CACCT 2011

IEEE

Proceedings of 5th International Conference on

ADVANCED COMPUTING & COMMUNICATION TECHNOLOGIES

Technically Sponsored by IEEE

November 05, 2011

Editors

R.K. Choudhary

Manoj Verma

Sanjeev Saini

Organised by

ASIA PACIFIC INSTITUTE OF INFORMATION TECHNOLOGY SD INDIA

Approved by AICTE, Ministry of HRD, Govt. of India and Dept. of Technical Education, Govt. of Haryana

Faridpur Road, G.T. Road Karnal Side, Panipat-132 103, Haryana (India) Ph.: +91-180-653 2444, 653 2555

Tele Fax: +91-180-257 7273 | e-mail: info@apiit.edu.in | Website: www.apiit.edu.in

5th IEEE International Conference

on

ADVANCED COMPUTING & COMMUNICATION TECHNOLOGIES [ICACCT-2011]

Technically Sponsored by IEEE

November 05, 2011

Editors

R.K. Choudhary
Manoj Verma
Sanjeev Saini

Organised by

Asia Pacific Institute of Information Technology SD India

1st International Engineering College
(Approved by AICTE, Ministry of HRD,
Government of India & Department of Technical Education, Govt. of Haryana)
Panipat-132 103, Haryana, India

ABC Group of Publication
9, Indira Colony, Vikram Marg, Karnal, Haryana, India

First Impression: 2011

5th IEEE International Conference on Advanced Computing & Communication Technologies

ISBN: 81-87885-03-3

DISCLAIMER

The authors are solely responsible for the contents of the papers compiled in this volume. The publishers or editors do not take any responsibility for the same in any manner. Errors, if any, are purely unintentional and readers are requested to communicate such errors to the editors or publishers to avoid discrepancies in future.

Published by
ABC Group of Publication
9, Indira Colony, Vikram Marg, Karnal
Tel: +91-184-6540168, +91-9215508638
Email: jobs@abccomputers.in
Website: www.abccomputers.in

Typeset & Printed by
Research India Publications
B-2/84, Ground Floor, Rohini Sector-16,
Delhi-110089, India
Email: ripublication@vsnl.net
Website: www.ripublication.com

Forward

It gives me immense pleasure in writing the message for the proceedings of 5th IEEE International conference on Advanced Computing & Communication Technology (ICACCT-2011), being organized on 5th November, 2011 by Asia Pacific Institute of Information Technology, Panipat. These proceedings comprises of technical research work of more than One hundred and twenty researchers from premier technical institutions and organizations.

Through the proceedings of ICACCT-2011, a forum is provided for sharing insights and experiences related to Advanced Computing & Communication Technologies. It will bring in new concepts, suggest approaches, and evaluate various tools for better understanding.

ICACCT-2011 provides the most appropriate opportunity to the faculties, researchers and the students of technical institutions to acquire practical corporate knowledge, as this conference provides a platform for interaction with the Honorable selves from academia and esteemed personalities from the corporate world.

I am pleased to see that in a short span of time we have attained a high level of maturity in organizing such research-oriented technical events. This is evident from the fact that the 5th IEEE International Conference is technically sponsored by IEEE & IETE – all leading professional bodies in the field of Engineering & Technology.

All the research papers compiled in the proceedings are well peer-reviewed and will be helpful to all the concerned Researchers, Industrialists and Technocrats of the same field. I wish the ICACCT-2011 a great success.

Prof. (Dr.) R. K. Choudhary
(Director, APIIT SD INDIA)

Preface

The world in which we are living today has become a global village, thanks to the developments that have taken place in the field of Advanced Computing and Communication Technologies.

By organizing this conference we have tried to bring Academicians, Researchers, Software professionals, Students, Corporate personnel and all such persons, who are related with this field, on a single platform so that they can share their ideas and experiences and may discuss what new developments are going on in this field.

People have responded to the conference with a great enthusiasm and academicians, researchers, students and corporate personnel have contributed a great deal. We hope all these participants will avail this opportunity to exchange their ideas with each other. We are sure that we will be able to make it a great success.

Editorial Board

Prof. (Dr.) R. K. Choudhary
Manoj Verma
Sanjeev Saini

Message

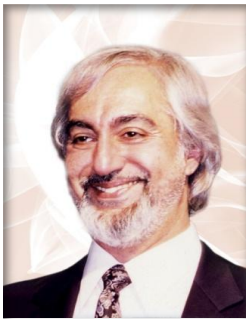
I am delighted that the 5th IEEE ICACCT-2011 conference is again being hosted by APIIT SD India. APIIT SD India is an organization with which Staffordshire University has had a long and successful partnership, particularly for the delivery of programmes in computing and engineering and for research activity. We are very proud of our collaboration with such an organization that is as successful as an engineering higher education college and research arena in India.

At the 5th IEEE ICACCT-2011 conference, some of my colleagues in the Faculty of Computing, Engineering and Technology will be presenting papers as part of the focus on Advanced Computing and Communication Technologies. This is a great opportunity for some of our researchers and all the other contributors to be sharing their work with academics, students and industry for the advancement of what information technology can deliver for the future to the benefit of economies and societies. It will clearly be a great event with around 400 delegates participating in various themes from all over the world. Making substantial advances in computing and communication is vital to our future as the world becomes smaller and the need to work together becomes greater. We all have so much to learn from each other and this conference will, no doubt, be a major contributor to that collaborative effort.

This event takes place on the same day as students graduate having achieved Staffordshire University awards delivered by APIIT SD India. We will be represented at both the graduation and the 5th IEEE ICACCT-2011 conference by the Dean of the Faculty of Computing, Engineering and Technology, Professor Michael Goodwin. I am only sorry that, on this occasion, I cannot partake in the celebrations of our graduates. I look forward to participating in graduation in the future.

Professor Michael Gunn
Vice Chancellor, Staffordshire University, UK

Message



Warmest greetings from Malaysia

On behalf of all the staff and students of the APIIT Education Group, I would like to first and foremost congratulate the management, staff and students of APIIT SD-India for their tremendous and collective efforts organizing this 5th IEEE International Conference on Advanced Computing and Information Technologies (IEEE ICACCT-2011).

This year's conference, which is organized by APIIT SD-India and technically sponsored by the IEEE Computer Society and the IETE, has received tremendous response from academics, researchers and industry practitioners not only from within India, but also from around the world. I understand that more than 300 technical contributions dealing with a wide variety of topics have been received, and this speaks volumes for the enhanced prestige and standing that the conference has continued to achieve since the first conference in 2005. The quality of papers submitted as well as the intensity of academic discourse during the conference is a reflection of the passion shown by conference participants in sharing and showcasing their impressive research findings

This year marks a major milestone in APIIT SD-India's development ad maturity with the launch of its new Degree Programmes in Engineering, as well as Masters Degrees in collaboration with APIIT in Malaysia, and with Staffordshire University, UK.

This augurs well for the future of APIIT SD-India. With the continued able stewardship of the SD Society, the dedicated work of its management and staff, the quality of its students, and with the continued support of our partners Staffordshire University, I am confident that APIIT-SD India will continue to excel in providing quality Higher Education so as to meet the ever-growing needs of employers in India and beyond.

Once again, I congratulate the staff and students of APIIT SD-India for organizing IEEE IACTT-2011, and I wish all participants a very fruitful and enjoyable conference.

Datuk Dr. Parmjit Singh
CHIEF EXECUTIVE OFFICE
Asia Pacific University College of Technology & Innovation (UCTI)
Asia Pacific Institute of Information Technology (APIIT)
PRO-CHANCELLOR
Staffordshire University, UK

भूपेन्द्र सिंह हुड्डा
BHUPINDER SINGH HOODA

सत्यमेव जयते

D.O. No. CMH-2011 /.....

मुख्य मन्त्री, हरियाणा,
चण्डीगढ़।
CHIEF MINISTER, HARYANA,
CHANDIGARH.

Dated 7.10.2011

Message

I am glad to know that Asia Pacific Institute of Information Technology SD India, Panipat is organizing 5th IEEE International Conference on the theme of 'Advanced Computing and Communication Technologies' on November 5, 2011.

I appreciate the decision to hold an international conference on such an important topic and that too at an opportune time as the advancement made in computer and communication technologies have, on one hand offered new opportunities and on the other these have also posed scientific and technical challenges. Therefore, it is high time to encourage innovations in high performance computing and communication technologies and stimulate the use of these technologies in key areas of computer application.

I am sure, the conference will promote the exchange of ideas and information among the delegates about new developments in high performance computing and communication technologies.

My best wishes for the success of conference.

(Bhupinder Singh Hooda)

Preamble

On behalf of the Organising Committee, it is my great pleasure to warmly welcome you in 5th IEEE International Conference on Advanced Computing and Communication Technologies (ICACCT-2011) at Asia Pacific Institute of Information Technology SD INDIA, Panipat. This conference is organized with the aim of disseminating knowledge amongst intellectuals, students and research scholars.

The objective of 5th IEEE ICACCT-2011 is to provide a highly interactive forum so as to bring together researchers of different disciplines, from academic and research institutions, industry and public organizations with the aim of collecting, exchanging and promoting the knowledge and new advances on Engineering Sciences specifically in the areas of Computer Science and Communication Technologies.

I am also honoured to have several plenary lectures by well-known leading experts from various engineering disciplines. I hope that the lectures and papers presented in this conference will stimulate and inspire future studies and advancement in engineering sciences.

I wish to welcome new and returning visitors to APIIT SD INDIA, Panipat and hope that you will find opportunities to tour this Historical-city and friendly people out here. I take this opportunity to express my sincere gratitude to the Conference patrons, IEEE Advisory Committee, International and National Advisory Committees, Program Committee, Session Chairs, Members of Review Committee, Organizing Committee and the Volunteers for their sheer efforts to make the ICACCT-2011 possible. I thank the management of APIIT SD INDIA, Sponsors and IEEE for their timely support and helping us in organizing ICACCT-2011.

Finally, I welcome you all to participate in ICACCT-2011.

Manoj Verma
(Convenor ICACCT-2011)

PATRON

Shri Roshan Lal Mittal
Shri Vijay Aggrawal
Shri Ishwar Garg
Shri Neeraj Aggrawal
Prof. (Dr.) R. K. Choudhary

IEEE Advisory Committee

Mr. Daman Dev Sood	(Chair – CS IEEE Delhi Section)
Prof. K. Subramanian	(Past/Vice Chair, CS Chapter, IEEE Delhi Section)
Dr. Deepak Garg	(Secretary, CS Chapter, IEEE Delhi Section)
Ms. Sharbani Bhattacharya	(Treasurer, CS Chapter, IEEE Delhi Section)
Prof. Ashok Bhattacharya	(Member, CS Chapter, IEEE Delhi Section)
Dr. Radhey G S Asthana	(Member, CS Chapter, IEEE Delhi Section)
Mr. Parkash V Eande	(Member, CS Chapter, IEEE Delhi Section)
Mr. Man Mohan S Puri	(Member, CS Chapter, IEEE Delhi Section)

International Advisory Committee

Prof. ADRIAN LOW	Director – RRE, Staffordshire University, UK.
Prof. ANTHONY ATKINS	Staffordshire University, UK.
Prof. HONG NAIN YU	Staffordshire University, UK.
Prof. ELAJSOLAN MOHAN	UCTI, Malaysia
Prof. ANDY SEDDON APIIT	Malaysia
Prof. S VENKATRAMAN	University of Ballarat, Australia
Prof. Y. K. MALAIYA	Colarado State University, U.S.A.
Mr. HIMANSHU GOEL	Country Manager, IBM
Mr. TARUN MALIK	Microsoft Corporation India.

National Advisory Committee

Prof. L. K. MAHESHWARI	Vice Chancellor, BITS Pilani
Prof. V. P. SAXENA	Ex-V.C., Jiwaji University, Gwalior
Mr. V. K. GUPTA	Sr. Technical Director, NIC
Prof. RAJENDRA SAHU	Director, RGUKT, Hyderabad
Sh. R. K. GUPTA	President, IETE, New Delhi.
Mr. P. K. ROY	Chief Scientist ONGC, Kolkata
Prof. TAPAN SENGUPTA	IIT Kanpur

Prof. R. P. MAHESHWARI	IIT, Roorkee
Prof. DEEPALI SINGH	AVB-IIITM, Gwalior
Prof. GOLDI MISHRA	HPCS, CDAC, Pune.
Prof. R. A KHAN	D.I.T, B.R.A. University, Lucknow
Prof. E.G. RAJAN	President, PRCentre. Hyderabad
Prof. VINOD KUMAR	Director General, VCE, Meerut.
Prof. K. MUSTAFA	Jamia Millia Islamia, New Delhi
Prof. BRAHMJEET SINGH	Gautam Buddha University, Noida
Prof. K. R. PARDASANI	MANIT, Bhopal
Dr. J. K. CHHABRA	NIT, Kurukshetra
Prof. B. PRASAD	Reader, Electronics Deptt., Kurukshetra University
Prof. PARVEEN KUMAR	MIET, Meerut
Prof. O. P. SAHU	NIT, Kurukshetra
Dr. AKHILESH UPADHYAY	SIRT, Bhopal

Review Committee

Prof. (Dr.) Dharmender Kumar	CSE Deptt. DCRUST Murthal
Prof. (Dr.) Dharminder Kumar	Dean of Faculty, Chairman CSE Deptt. GJU Hissar
Prof. (Dr.) Dinesh Chutani	Professor CSE Deptt. GJU Hissar
Prof. (Dr.) Nasib Singh Gill	CSA Deptt. MDU Rohtak
Prof. (Dr.) Pardeep Bhatia	CSE Deptt. GJU Hissar
Prof. (Dr.) Vikram Goyal	CSE Deptt. IIIT Delhi
Prof. (Dr.) Mukesh Yadav	Dean of Academics, SGTIET Gurgaon
Prof. (Dr.) Yudhvir Singh	CSE Deptt. UIET Rohtak
Dr. Akhilesh Upadhyay	SIRT, Bhopal
Dr. Virendra Srivastva	APIIT SD INDIA, Panipat

LOCAL ORGANIZING COMMITTEE

CONVENER
Manoj Verma

CO-CONVENER
Sanjeev Saini

SESSION MANAGEMENT
Dr. Virendra Srivastava, Ravi Sachdeva**,** Ankur Singla, Gaurav Gambhir, Radha krishana Rambola, Geetender Handa, Prateek Mishra, Geeta Nagpal, Vishal Kumar, Sohan, Dheeraj

WEBSITE MANAGEMENT
Umesh Verma, Praveen Saini, Manish Kumar

PUBLICATION
Manoj Verma, Dr. Virendra Shrivastava, Sanjeev Saini, Vikrant Sehgal, Sachin Jain

MEDIA & PUBLICITY
Sachin Jasuja, Gaurav Gambhir**,** Sachin Jain, Parveen Saini

SPONSORSHIP & FINANCE
Arun Choudhary, Suresh Hans, Sanjeev Jawa, Vinod Kumar

HOSPITALITY
Dr. Vikrant Sehgal, Sumit Pahwa**,** Suresh Rohila, Anil Jaukhani
Pramod Bhardwaj, Pawan Kumar, Narender, Maria

REGISTRATIONS & CERTIFICATES
Rajesh Tiwari, Ajeet Pathak, Kiran Malik, Sonia Arora, Himanshu Goel, Afzal Khan

INAUGURATION & VALEDICTORY
Sapna Bhandari, Kirti Jasuja, Kashish mathur, Umesh Verma

CATERING
Sunil Kumar, Kanhiya, Vikas Singhal, Jatin Kumar, Pramod Bharadwaj
&
Volunteers

Contents

Implementation Issues in Multi-View Rendering on Spatial Multiplex based 3D Display System

Dilip Kumar Dalei, Kuldeep Goyal and N. Venkataramanan

ANURAG, Defence R & D Organization (DRDO), Kanchanbagh, Hyderabad, 500058, India.
E-mail:dilip.dalei@gmail.com,goyalkuldeep@gmail.com,n_venkataramanan@rediffmail.com

Abstract
Auto-stereoscopic displays unlike stereoscopic systems provide a seamless 3D experience without any viewing aid like head gear or shutter glasses. These display systems can be broadly classified into three categories: Spatial Multiplex, Multi-Projector and Time Multiplexing. We have focused our research on spatial multiplex based 3D display technology. The paper explains two such display systems, namely Parallax barrier and Lenticular Lenlets. We are investigating the different approaches for multi-view auto-stereoscopic rendering and their implementation issues. It also presents a detailed study of performance issues in various approaches. The implementation is done in C++ using OpenGL graphics library.

Keywords: Stereovision, Auto-stereoscopic display, Multi-view Rendering, 3D Monitor, Spatial Multiplexing, OpenGL, Frame Buffer Object(FBO), Parallax Barrier, Lenticular Lenslets.

Introduction

Stereoscopy is a well known technique that enables to understand three dimensional visual information from two-dimensional planar images. It has got many areas of application like data visualization, virtual reality and entertainment, because it helps in understanding the information by reproducing our visual perception. Stereovision creates the depth perception by presenting a slightly different image to each eye. Anaglyph, polarization or shutter glasses are most popular methods to achieve stereovision, but these involve using external devices that impairs the feeling of immersion [2]. So, the trend is gradually shifting from stereoscopic to auto-stereoscopic displays.

Auto-stereoscopy is a next generation 3D technology that introduces the ability to watch 3D effects without any viewing aid. Moreover, current auto-stereoscopic methods can display much more than two images to provide an adequate rendering in several directions and to be adapted for multi-user purposes. Our work tries to understand the implementation and optimization issues of different methodologies available for auto-stereoscopic rendering. It also presents a comparative study of these methods.

The remainder of the paper is organized as follows. Section 2 explains the basic foundation of Auto-stereoscopic display systems and their guiding principles. This is followed by explanation of various approaches for multi-view rendering and their implementation in Section 3. The performance results are analyzed in Section 4. Finally, the paper is concluded in section 5.

3D Displays

Background

The extraction of three-dimensional information about the world from the two retinal images received by our eyes is a fundamental problem of human visual system. It provides us the information about scene objects and their relative position. Human stereo vision relies on two kind of visual hints to interpret three-dimensional information from a scene [4]. These are physiological (primary) and psychological (secondary, sometimes called pictorial) depth cues. The physiological depth cues are based on the physical structure of the eyes and include accommodation, convergence and retinal disparity. On the other hand psychological depth cues include relative size, linear perspective, height of objects above the line of sight, interposition, shading, shadow, relative brightness, color (chromostereopsis), and atmospheric attenuation. It helps to make sense of photos, paintings and images on 2D display systems. Even though these images are flat, we perceive depth or distances. The combination of these cues with physiological cues enhances the three-dimensional effect [6].

Stereoscopic displays require users to wear a device, such as analyzing glasses, that ensures left and right views are seen by the correct eye. Many stereoscopic display designs have been proposed for professional markets. In spite of appealing 3D effects they suffer from the drawback that the viewer has to use some external aid to separate the left and right eye views. This has limited the widespread appeal of stereoscopic systems as personal displays for home and office. However, they are particularly suited to multiple observer applications such as cinema and group presentation where directing individual images to each observer becomes difficult compared to providing each observer with a pair of analyzing glasses.

Auto-stereoscopic Technology

This technology has existed for many years, but the rapid growth in creation of 3D contents and their usage has popularized the technology worldwide. Auto-stereoscopic displays are based on the system creating two or more views of the scene and displaying each view into a discrete set of angles. There are mainly three categories of technologies that

are used for designing auto-stereoscopic displays. These are Spatial Multiplex, Multi-projector and Time Multiplexing [1]. The paper focuses the study on spatial multiplex based auto-stereoscopic display designs.

Spatial-multiplexing is an auto-stereo approach which involves image interleaving. Image interleaving splits the source images into strips and merge these strips into a single image. A significant issue in spatial multiplex is the loss of resolution of source images by half. The spatially-multiplexed category of displays is further classified by the mechanism employed to re-direct interleaved images toward the eyes [4]. These are *Parallax Barrier* and *Lenticular* display systems.

Parallax Barrier System

The parallax barrier method of 3D display is probably the oldest known auto-stereoscopic technique that preserves horizontal parallax. In this approach, a barrier is created by adding a second sheet of alternate opaque/transparent columns mounted on top of the display. These vertical columns reveal different parts of the underlying image depending upon the viewing direction.

The left and right eye would see different sets of columns as shown in figure 1. Thus, by encoding the left eye view on one set of columns and the right eye view on the other columns, a two view auto-stereoscopic image is created at a certain distance from the display. The main disadvantage of the parallax barrier is that it blocks part of the light, resulting in a lowered brightness of the display.

Examples of Parallax Barrier displays are 4D-Vision [10], the Varrier [11], SuperD HDB 24 etc.

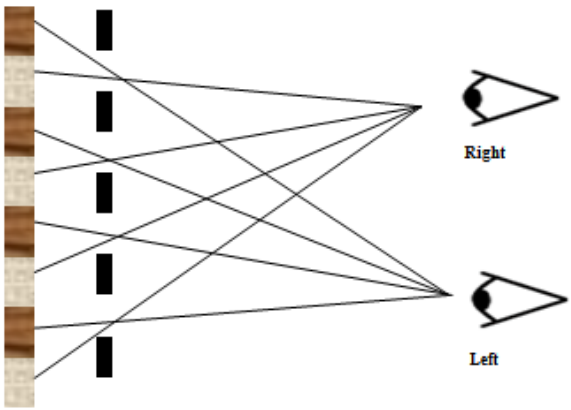


Figure 1: Parallax Barrier Display.

Lenticular Lenslets System

The other popular method for auto-stereoscopic displays is to use a lenticular array in place of the parallax barrier. It consists a sheet of long cylindrical lenses (lenticulars) placed over a flat display in such a way that the display's image plane coincides with the the focal plane of the lenses as shown in the figure 2. This helps the lenses to focus different portion of each image toward the user's eye.

As the lenses collect all the light from pixels there is no loss of brightness, which is an important advantage over barrier technology. But it still suffers from the same resolution and viewing zone problem as parallax barrier systems. A relatively low number of views have the advantage of a low resolution loss, but the disadvantage of limited look-around ability. Systems like SynthaGram [7], Phillips 3D-LCD [10], SuperD HDL 24 are examples of Lenticular displays. We use both Barrier and Lenticular 3D monitors from SuperD in our work.

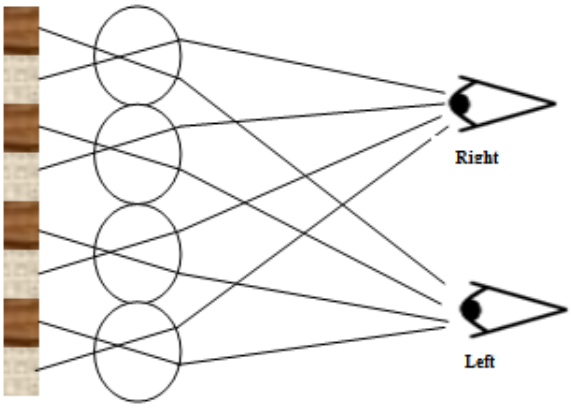


Figure 2: Lenticular Display.

Multi-View Rendering

Approach

In a scene user's eye see an infinite number of different images. The whole viewing space is theoretically divided into finite viewing zone. Each zone displays one image or view of the scene. But the user still sees different images for each eye and the image change upon movement of head. This preserves both stereo and horizontal movement parallax cues necessary for 3D perception. In multi-view rendering the scene is rendered to intermediate images in multi passes. All images are fused into a single 3D spatial image conforming to the auto-stereoscopic display. The final image accordingly projected into different viewing zones using optical filters (Parallax Barrier/Lenticular sheet).

Implementation

Many algorithms are proposed to create auto-stereoscopic images on a parallax display. We mainly investigate the Multi-pass/deferred approach to manipulate the image pixels to display 3D view. The parallax type of displays generally requires a single multiplexed 3D image constructed from multiple perspective views of the scene. This can be done either in single pass or multi pass. The paper limits the analysis to multi pass (deferred) approach.

In this approach the scene is rendered N times from N different viewpoints and the corresponding N images are collected and processed to form a composited image. As per technical requirement of 3D monitor nine views (N = 9) of the scene are captured in our experiment. The images are arranged in a special nine-tiled matrix format as shown in figure 3. This multi-view algorithm, rendering of N full-resolution images followed by N-fold image masking and compositing, leads to an increase in cost by N times.

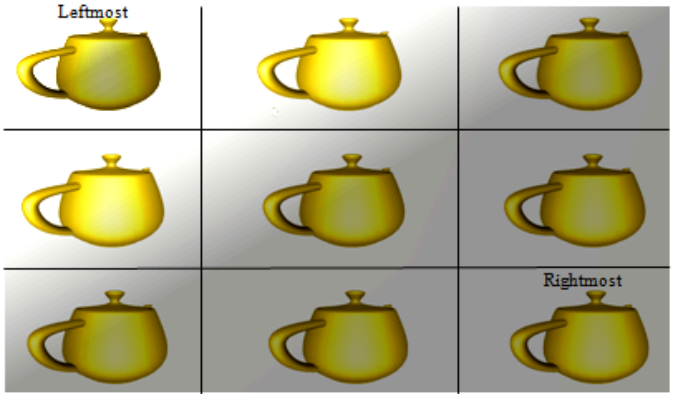


Figure 3: Nine-tile Format: Nine views of teapot arranged in a 3 X 3 Matrix.

The whole implementation is carried out in C++ using OpenGL API on Linux Platform. OpenGL is widely used as the industry standard 3D graphics library. It provides a pipeline approach to render a 3D scene. In OpenGL, multi pass algorithm is achieved in two phase called *view collection* and *composition*. The basic steps for implementation of the algorithm are outlined below.

View Collection Phase
Step 1: Prepare the scene model.
Step 2: Compute a view matrix for a particular viewpoint and multiply it to the current projection matrix.
Step 3: Render the scene N times and accordingly grab the final images.

View Composition Phase
Step 4: Prepare a single composited image by tiling the images according to a pre-specified order.
Step 5: Pass composited image to a shader program to generate a spatial multiplex 3D image.

In *view collection* phase, the scene is rendered from eight viewpoints and the corresponding images are captured from the frame buffer using *glCopyTexImage2D* function. This function copies the frame buffer image to texture memory. We have not used *glReadPixel* function which is slower than *glCopyTexImage2D* method. Later these are processed using spatial multiplex technique to create a single 3D image. The whole rendering pipeline is executed once for each viewpoint. The explicit copy of intermediate images from frame buffer to texture memory degrades the performance substantially. To improve the performance we engage the Off-screen approach provided by OpenGL Frame Buffer Object (FBO). By using FBO we can render a scene directly onto a texture. So, there is no need of framebuffer. Furthermore it avoids the additional data copy (from framebuffer to texture). This gives a performance gain in both time and frame rate.

In *view composition* phase the images are explicitly arranged in nine-tiled matrix format and the tiled image is passed to a pixel shader for generating the final 3D image. As the pixel shader runs directly on current graphics hardware it hardly poses any performance bottleneck. So it is observed that this phase has limited scope for optimization.

Results and Discussions
The hardware specification of our Testbed system is Intel Core2 duo with two Nvidia Quadro Fx 5800 graphics card under Linux System. Two HDL/B 24'' 3D monitors from SuperD Corporation are used in the experiment. A brief technical specification of the monitors is given in table 1.

For the experiment we have taken classical graphics models like *Teapot* and *Bunny* models which are available in public domain. The teapot model consists up 3241 vertices and 6350 polygons while the bunny model has 35947 vertices and 69451 polygons. We have tested the performance of these models for auto-stereoscopic rendering. This is illustrated in the figure 4 and figure 5.

Table 1: Technical Specification of 3D Monitors.

	SuperD HDL 24 (Lenticular Lenslets)	SuperD HDB 24 (Parallax Barrier)
Best Resolution	1920X1200	
Pixel Size	0.27 mm	
Viewing Angle	$160/160^0$	
Viewing Range	3m	

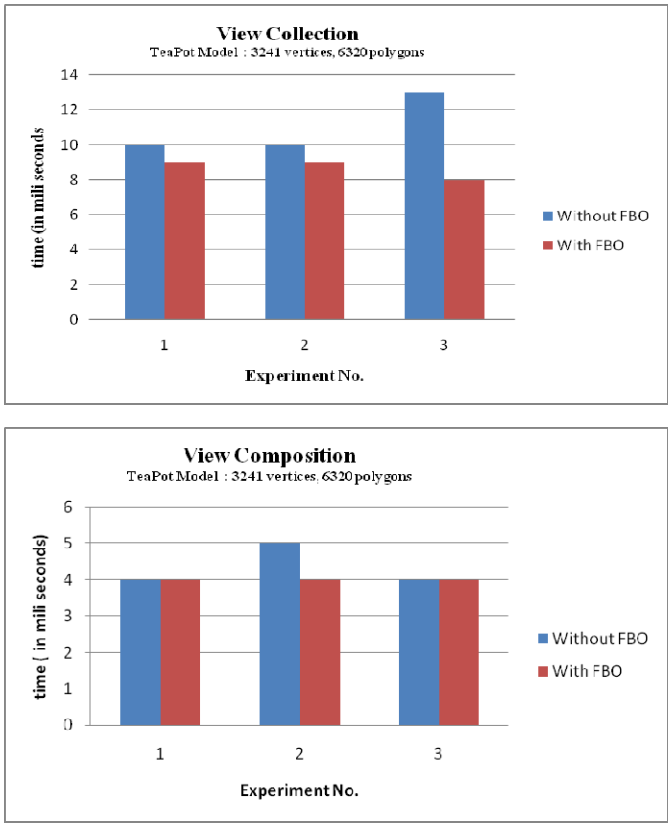
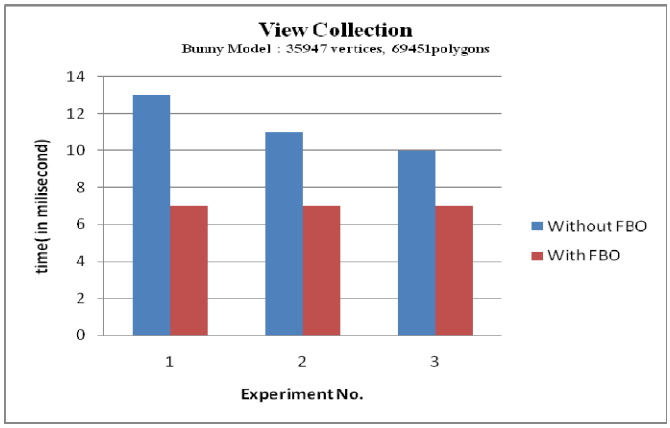


Figure 4: Performance Analysis of Teapot Model for view collection and composition time with/without FBO.



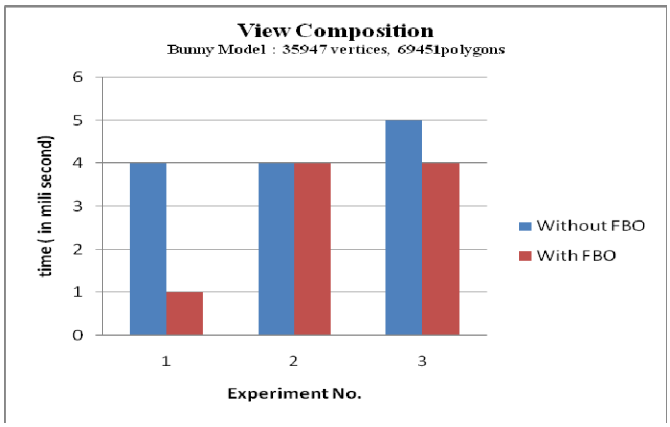


Figure 5: Performance Analysis of Bunny Model for view collection and composition time with/without FBO.

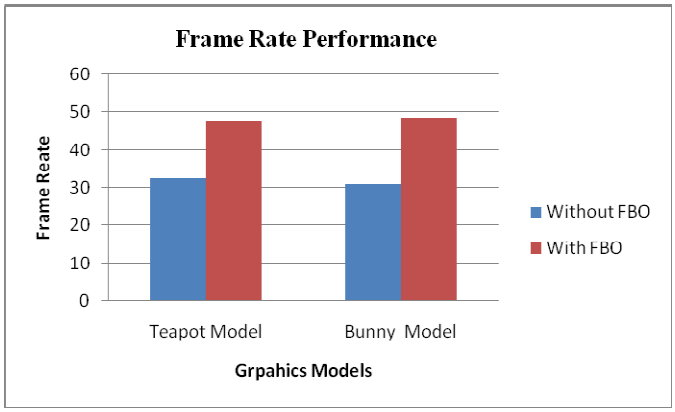


Figure 6: Frame Rate Analysis of Teapot and Bunny Model with/without FBO

Conclusion

In our experiment we mainly focus our analysis on the implementation methods of 3D rendering on parallax based auto-Stereoscopic displays. The paper clearly discusses the area of improvement in implementation methodology. We observe that the total performance of 3D image rendering depends mainly on view collection than view composition. This bottleneck can be improved using any faster image collection techniques.

This work opens many future directions for research. First Image warping can be used to generate multiple views from a reference view. Second many researchers have proposed the use of Geometry Shader Model to accelerate 3D Rendering.

Acknowledgement

We thank our team members for constant support and encouragement. We would to extend our courtesy to Stanford Computer Graphics Laboratory for distributing the graphics models for research purpose.

References

[1] N. A. Dodgson, "Auto-stereoscopic 3d displays", *Computer*, vol. 38, no. 8, pp. 31-36, 2005.

[2] John R. Moore, Neil A. Dodgson, Adrian R. L. Travis, Stewart R. Lang, "Time-multiplexed color auto-stereoscopic display", SPIE Symposium on Stereoscopic Displays and Applications VII, Jan 28–Feb 2, 1996

[3] F. de Sorbier, V. Nozick and V. Biri, "GPU rendering for auto-stereoscopic displays", *4th International Symposium on 3D Data Processing, Visualization and Transmission(3DPVT'08)*, June 2008.

[4] Robert L. Kooima, Tom Peterka, Javier I. Girado, Jinghua Ge, Daniel J. Sandin, Thomas A. DeFanti, "A GPU Sub-pixel Algorithm for Auto-stereoscopic Virtual Reality".

[5] Nick Holliman, "3D Display Systems", February 2, 2005

[6] E. Lynn Usery, "Auto-stereoscopy – Three-Dimensional Visualization Solution or Myth?"

[7] L. Lipton, M. Feldman. A New Autostereoscopic Display Technology: The SynthaGram. In Proceedings of SPIE Photonics West 2002: Electronic Imaging, San Jose, California, 2002.

[8] Ian Sexton, "PARALLAX BARRIER DISPLAY SYSTEMS".

[9] Michael Halle, "Auto-stereoscopic displays and computer graphics" ACM SIGGRAPH, 31(2), pp 58–62, May 1997.

[10] A. Schmidt, A. Grasnick. Multi-viewpoint Autostereoscopic Displays from 4D-Vision. In Proceedings of SPIE Vol. 4660, 2002.

[11] D. Sandin, T. Margolis, J. Ge, J. Girado, T. Peterka, T. Defanti. The Varrier™ Autostereoscopic Virtual Reality Display. In ACM Transactions on Graphics, Proceedings of ACM SIGGRAPH, 24, no.3, 2005, pp. 894-903

[12] C. van Berkel, "Image Preparation for the 3D-LCD". In Proceedings of SPIE, Stereoscopic Displays and Virtual Reality Systems. 1999.0020

Privacy Providing Authentication Scheme for Vehicular Networks

Upasana Singh and Pardeep Singh

Computer Science and Engineering, NIT Hamirpur, Hamirpur, India
E-mail: upasananith@gmail.com, pardeep@nitham.ac.in

Abstract
Frequent handovers are realized in vehicular communication because of the high speed of vehicles and hence there is always a requirement of secure and fast authentication for a seamless handover to take place. Apart from security of communication, privacy conservation of user specifics also an important requirement of vehicular communication. We proposed an authentication scheme that will not only provide security and but privacy and also will reduce the storage and communication overhead while increasing the efficiency.

Keywords: Vehicular Networks, Security, Privacy, Authentication, VANET.

Introduction

In present scenario vehicles are not only envisioned to communicate between each other, but also to get information from and send data to infrastructural units. Many R & D groups have shown enormous amount of interest in development of this technology. The CAR 2 CAR communication Consortium, SAFESPOT, eSAFETY, PReVENT, EASIS, SEVECOM are some of the European Initiatives. The vehicles and the Road Side Units (RSU) are equipped with On-Board processing and wireless communication modules. The vehicular communication could be Intra-Vehicular communication or Inter-Vehicle Communication. In Intra-Vehicular communication the On-Board Unit (OBU) communicates with several Electronic Control units (ECU). The Inter-Vehicular Communication could be Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication. Since a rich set of tools are offered to drivers and authorities, but a formidable set of exploits and attacks becomes possible. Hence, the security of vehicular networks is indispensable, because these systems can make anti-social and criminal behavior easier, in ways that will actually jeopardize the benefits from their deployment [24]. Vehicular communication is vulnerable to several kind of attacks like Jamming in which the attacker purposely generates interfering transmissions that prevent communication; an attacker might forge and transmit false hazard which are taken up by all vehicles in both traffic streams; attackers can replay messages, impersonation where an attacker masquerade of an emergency vehicle to mislead other vehicles to slow down and yield or impersonate a roadside units, spoofing service advertisements or safety messages; the attacker may select to alter the data at their source, tampering with the on-board sensing and other hardware will be relatively simple.

The paper is organized in the following way: Section II describes the related work. In Section III system model and problem statement are described. Our proposed solution is given in Section IV and security and performance analysis in V. Finally section VI concludes the paper.

Related Work

In case of vehicular networks a balance has to be made between the security and privacy for its proper functioning. Security and privacy related problems have been discussed by many researchers. While most of them addressed both security and privacy some of them fail to do so. Although communication in vehicular networks has to be real-time constrained, most of the proposals incur communicational overhead. Various techniques have been used to secure vehicular communication like symmetric key cryptography, asymmetric key cryptography, ECC, Id-based cryptography, and sometimes in combination with some hardware.

A secure communication architecture is proposed in [1] based on a public key infrastructure (PKI) and a virtual network controlled by cluster-heads but their approach produces a remarkable overhead and the use of cluster-heads can create bottleneck. The importance of privacy and secure positioning was considered in [4] and proposed the use of Electronic License Plates (ELP) to identify vehicles. Although they recognize the importance of conditional privacy, they do not provide any specific solution to the problem. To the best of our knowledge, there are few articles that consider both security and conditional privacy preservation in VANETs. In this line, [3] gave a foundational proposal of using pseudonym based approach using anonymous certificates and the public key infrastructure (PKI).The anonymous certificates are used to hide the real identities of users. This scheme required extra communication and had storage overhead. Also privacy could be invaded by logging the messages containing a given key and tracking the sender until her identity is disclosed. GSIS presented in [6], is a conditional privacy-preserving scheme using group signatures and ID-based signatures. In it a single membership manager is used to issues secret member keys to the vehicles. The conditional anonymity claimed applies only to the vehicles amongst the peer, with an assumption that the infrastructure points are trusted. An alternative way was proposed in [7] to overcome the limitation of pre-storing a large number of anonymous certificates while preserving conditional privacy. They proposed a group signature based scheme, making an assumption that vehicles and RSUs are able to collaborate actively. Every vehicle gets a short-time

anonymous certificate from a RSU after running a Two-round protocol when passing by the RSU. In order to prevent link ability of the messages, the vehicle should change the anonymous certificate regularly by interacting with RSUs. These frequent interactions may affect the network's efficiency. Group signature based schemes have the problem of identity escrow, i.e. the group manager could reveal the identity of any group member. The group based schemes could not be applied properly due to certain limitation as the difficulty in election of group leader due to the non-availability of a trusted entity among the peer vehicles; also there may be too few cars in the vicinity to create a group.

So it shows that asymmetric cryptography based solutions using certificates and signatures are secure but generate computational and storage overhead. Also group based schemes cannot be employed efficiently because of the reasons already discussed.

System Model and problem statement

System Model

Vehicular networks consist of several entities. A Trusted Authority which could be a law enforcement authority (or a group of authorities) could trace and disclose the identity in case of accident or crime. AAA server (authentication, authorization and accounting server) which authenticates the vehicle when it first enters the network and establishes the keys to be used. Road side units which act as the access points or access routers. The OBUs are installed on vehicles, RSUs and AAA server. In order to have seamless mobility and support the infotainment applications the network is FMIPv6 based.

Problem Statement

1. The OBUS have less storage and computational power than the RSU.
2. Even though tamperproof OBUS could be used to secure the data stored and prevent the attacker from reading it but while communication the energy could be intercepted.
3. In asymmetric cryptography like ECC (Elliptic Curve Cryptography), the most compact public key cryptosystem so far, the estimated security overhead of the signature and certificate is around 140 bytes.
4. The vehicles are highly mobile and hence have very less time to connect to a new RSU. The time to complete the handover is dependent on the number of messages exchanged during the handover.

The vehicular networks provide safety and commercial applications to the users and with the advent of infotainment applications the vehicular networks need support multimedia and real-time services. The handover in vehicular networks have to be secured while maintaining the computational and storage overhead. For real-time services such as infotainment application in vehicular networks the latency problems are not desirable. Therefore, a security framework has to be developed that will reduce the computational complexity.

Our Solution

In our solution we will be using terms vehicle and mobile node interchangeably similarly Access router (AR) and Road Side Unit (RSU) interchangeably. We have divided the Solution in three phases starting with mutual authentication and Key agreement phase, next verification phase and the handover phase.

Each vehicle is given a unique identity UID and password PSW by the TA. When the vehicle enters a network it enters UID and PSW in the OBU which generates a pseudoidentity ID_A as;

ID_A= (UID $\oplus$ PSW).

Mutual Authentication and Key agreement:

The AAA server chooses two large primes p and q and keeps them secret, it than computes n = (p.q). When the vehicle first enters the network it sends ID_A to the AAA server. AAA server than computes $J_A = f(ID_A)$ and sends J_A to the vehicle. AAA chooses a secret s such that $1 \le s \le n-1$. Than Computes $v = (J_A.s)^2 \bmod n$ which is the Vehicles public key. AAA selects and sends a shared secret 'g' to the vehicle. Both AAA server and vehicle choose respective secret numbers a and b such that $1 \le a$ and $b \le g-2$ each co-prime to g-1. They respectively compute $a^{-1} \bmod g-1$ and $b^{-1} \bmod g-1$. AAA server chooses a secret K such that $1 \le k \le g-1$, and computes (k.a) mod g and sends to the vehicle. Vehicle than multiplies the received value by b and sends it to AAA. AAA than multiplies the received value by $a^{-1} \bmod g-1$ which undoes its previous multiplication and sends it back to the vehicle. Vehicle multiplies the received value by $b^{-1} \bmod g-1$ which results in K mod g. This K mod g is the shared secret key between the AAA and the vehicle which is used as the Master key (MK).

This key will not be used for any kind of encryption it will only be used for deriving handover encryption key. The AAA computes handover encryption key (HEK) using the MK as $HEK = (MK||ID_{MN}||ID_{AR})$ and sends the HEK to the corresponding AR.

Verification Phase

Before the vehicle attaches to the new RSU and disconnects from the previous one, previous RSU is responsible to send V's related authentication information to the new one. Whenever a vehicle enter the vicinity of an RSU and have to communicate its identity must be verified. Verification steps are as follows:

Vehicle sends ID_A and $x = (J_A.r)^2 \bmod n$ *to the NAR(RSU).* NAR than randomly select a challenge bit e=0 or 1 and send to vehicle. The vehicle than compute $y=(ID_A.r) \bmod n$ if e=0 and if e=1 than $y = (ID_A.r.s) \bmod n$ and sends it to NAR. *NAR than computes J_A from ID_A using f and* $y^2 = (x.v^e) \bmod n$.

If both the values of y received and calculated are same than the verification is successful.

Handover Phase

The MN sends RtSolPr request to the PAR to which it is already connected for the information of the available ARs. The PAR than sends the information of the ARs to which the MN could attach via PrRtAdv. When the MN selects NAR to which it wants to connect it sends;

Msg1= HEK(ID_{MN}, ID_{PAR}, ID_{NAR}, $Nonce_{MN}$),
H(HEK, ID_{MN} || ID_{PAR} || ID_{NAR} || $Nonce_{MN}$)

along with the verification request. NAR on receiving Msg1 from MN firstly verifies the MN and than responds with

Msg2= HEK(ID_{MN}, ID_{PAR}, ID_{NAR}, $Nonce_{MN}$, $Nonce_{NAR}$), H(HEK, ID_{MN} ||ID_{PAR} || ID_{NAR} || $Nonce_{MN}$ || $Nonce_{NAR}$) .

After exchanging message Msg1 and Msg2 the Handover key HK it be computed by the MN using a one way hash function;

HK = (HEK, ID_{MN} || ID_{NAR} || $Nonce_{MN}$ || $Nonce_{NAR}$)

which will be used further during Handover. The MN will send FBU (Fast Binding update) message to the PAR along with some MAC (Message Authenticated Code) i.e. FBU, H(HK, NCoA, NonceNAR). PAR than sends handover initiation HI message along with received MAC i.e.

HI, H (HK, NCoA, NonceNAR).

On receiving the message NAR generates HK and verifies if what is received is same and if the verification is successful it sends Hack (Handover acknowledgement) to the PAR. PAR responds with an FBack (Fast Binding Acknowledgement). MN on attaching to the NAR transmits FNA (Fast Neighbor Advertisement) to the NAR to inform its presence.

Security and performance analysis

Our solution provides Identity privacy and anonymity via the use of pseudoidentity. Mutual authentication of MN and NAR is guaranteed via HK. Secrecy is achieved as Msg1 and Msg2 exchanged between MN and NAR are kept secret from an adversary. HK is only shared between MN and NAR. Neighbor ARs cannot derive HK and the key is kept secret from the attackers.Msg1 and Msg2 exchanged between MN and NAR cannot be altered by the attacker and hence integrity is achieved.

Denial of Service attack: Our proposal suggests a secure binding update authentication scheme using a security association between AR and MN. The scheme provides not only mutual authentication between MN and ARs, but also guarantees secrecy between ARs.

Know-key Security: If the attacker have intercepted the previous session key, still he can't use them to derive new session keys as both vehicle and RSU both generates new nonce for every new session, and in addition protected by the secure hash function. Hence our solution is secure against any adversary known key attacks.

Passive attack: A passive attack is possible if the attacker tries to guess the session key based on the information available publically. Even if the attacker performs a passive attack, he can't succeed as after verification both vehicle and the RSU will compute their session keys based on their secret shared information and the attacker could not compute. Therefore the proposed protocol resists against the passive attack.

Man in middle attack: It is a kind of active attack. Since no information about the secret key is revealed so the solution is safe against the man in middle attack.

Performance analysis

First we defined some computational parameters as follows:

- *thash* denotes the time for the hash operation.
- *tsym* denotes the time for the symmetric encryption/decryption operation.
- *tasym* denotes the time for the asymmetric encryption/decryption operation.
- *trandom* denotes the time for generating a random number.
- *tmul* denotes the time for the multiplication operation
- *tinv* denotes the time for the inverse operation
- *tsym* is at least 100 times faster than *tasym* in software. Moreover, *thash* ≈ 0.5ms, *trandom* ≈ 0.5*ms*, *tsym* ≈ 8.7*ms*, *tmul* ≈ 0.5 ms, *tinv* ≈ *19tmul* [27] [28].

Computational overhead during authentication and key agreement phase:

At V: *thash* + *tinv* + *tmul* + *tmul* + *tasym* ≈ 889.4 ms

At AAA: *tmul* + *trandom* + *tmul* + *tinv* + *tmul* + *tmul* + *thash* + *tasym* ≈ 882.2 ms

At V: Computational overhead during handover phase;

tmul + *tmul* + *tsym* + *thash* + *thash* + *trandom* + *thash* ≈ 11.7 ms

Total computational overhead at V is:

thash + *tinv* + *tmul* + *tmul* + *tasym* + *tmul* + *tmul* + *tsym* + *thash* + *thash* + *trandom* + *thash* ≈ *901.1 ms*

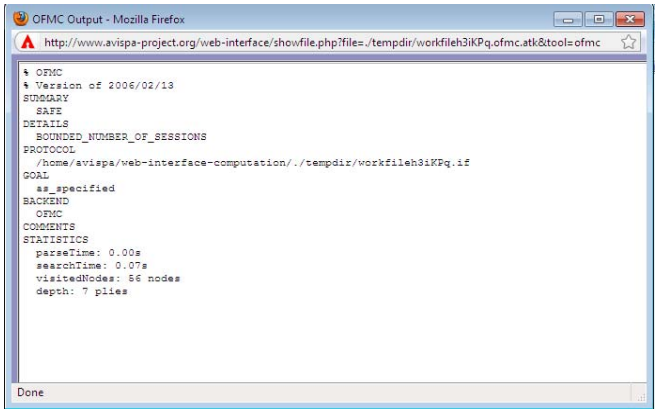
At RSU: Computational overhead during handover phase;

trandom + *trandom* + *tmul* + *tsym* + *thash* + *trandom* + *thash* ≈ 11.7 ms

Total computational overhead at RSU (with the AAA);

tmul + *trandom* + *tmul* + *tinv* + *tmul* + *tmul* + *thash* + *tasym* + *trandom* + *trandom* + *tmul* + *tsym* + *thash* + *trandom* + *thash* ≈ 893.9 ms

We simulated our protocol by using the Automated Validation of Internet Security Protocols and Applications (AVISPA) web interface. We assume that vehicular communications are insecure. That is, an adversary has got all transmitted messages *{Msg1, Msg2,* Msg3, *Msg4}* and has obtained the secrets *{MN, AR, f, h ( ), g}*. The proposed protocol is safe.



Conclusion

Vehicular networks are the vital solution to secure and efficient transportation system proving different types of applications to the vehicles. In order to take full advantage of the vehicular networks the communication must be secured meeting all the security requirements. Our proposed solution provides security and privacy both using symmetric key cryptography reducing the computation and storage required. Also, it enables to reduce the handover latency by reducing the number of messages exchanged with AAA server to zero.

References

[1] Blum, J., Eskandarian,A.: The threat of intelligent collisions

[2] J. Sun, C. Zhang, Y. Zhang, "An Identity-Based Security System for User Privacy in Vehicular Ad Hoc Networks", *IEEE Trans. Parallel and Distributed System 2010*, Vol. 21, No. 9, p 1227-1239.

[3] M. Raya, J. Hubaux, 'Securing vehicular ad hoc networks', Journal of Computer Security, Special Issue on Security of Ad Hoc and Sensor Networks 2007, Vol. 15, No. 1, p 39–68.

[4] X. Lin, X. Sun, P. Ho, "GSIS: A secure and privacy preserving protocol for vehicular communications", *IEEE Trans. Vehicular Technology 2007*, Vol. 56, No. 6, p 3442-3456.

[5] L. Gollan, C. Meinel, "Digital Signatures for Automobiles", *Technical Report, Institute for Telematike 2004*, Vol. 6, No. 1,p 24-29.

[6] J. Hubaux, J. Luo, "The security and privacy of smart Vehicles", *IEEE Journal on Security and Privacy 2004*, Vol. 2, No. 3, p 49-55.

[7] R. Lu, X. Lin, X. Zhu, "ECPP: Efficient conditional privacy preservation protocol for secure vehicular communications", *IEEE INFOCOM 2008*, pp. 1229-1237.

[8] J. Blum, A. Eskandarian, "The threat of intelligent collisions".

[9] D. Boneh, M. Franklin,"Identity-based encryption from the Weil pairing", *Journal on Advances in Cryptology-CRYPTO 2001, Lecture Notes in Computer Science*, Vol. 2139, p 213-229.

[10] D. Chaum, E. Van Heijst, "Group signatures", Advances in Cryptology 1991, *Lecture Notes in Computer Science*, Vol. 576, p 257-265.

[11] T. W. Chim et al, 'SPEC: Secure and Privacy enhancing communications schemes for VANETs", *Journal on Ad Hoc Networks (2010)*, doi: 10.1016/j.adhoc.2010.05.005.

[12] A. Shamir, "Identity based cryptosystems and signature schemes", *Lecture Notes in Computer Science 1984*, Vol. 196, p 47-53.

[13] J. Freudiger, M. Raya, M. Feleghhazi, "Mix zones for location privacy in vehicular networks", *ACM Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS), 2007*.

[14] C. Zhang, R. Lu, X. Lin, An efficient identity based batch verification scheme for vehicular sensor networks, in *Journal IEEE INFOCOM 2008*, p 246-250.

[15] X. Lin, X. Sun, P. Ho, "GSIS: A secure and privacy-preserving protocol for vehicular communications", *IEEE Trans. Vehicular Tech. 2007*, Vol. 56, No. 6, p 3442–3456.

[16] A. Studer, E. Shi, F. Bai, "TACKing together efficient authentication, revocation, and privacy in VANETs", Proc. 6th Annual IEEE SECON Conference (SECON'09), 2009.

[17] P. Kamat, A. Baliga, W. Trappe, "An identity-based security framework for VANETs", *Proc. 3rd ACM Int'l Workshop on Vehicular Ad Hoc Networks, VANET'06*, p 94–95.

[18] P. Kamat, A. Baliga, W. Trappe, "Secure, pseudonymous, and auditable communication in Vehicular Ad Hoc Networks", *Journal on Security and Communication Networks* 2008, Vol. 1, No. 3, p 233–244.

[19] J. Sun, C. Zhang, Y. Fang, "An id-based framework achieving privacy and non-repudiation", vehicular ad hoc networks in *Proc IEEE Military Communications Conf. 2007*, p 1–7.

[20] J. Sun, Y. Fang, "Defense against misbehaviour in anonymous vehicular ad hoc networks", *Journal on Ad Hoc Networks 2009*, Vol. 7, No. 8, p 1515–1525.

[21] G. Calandriello, P. Papadimitratos, J. Hubaux, "Efficient and robust pseudonymous authentication in VANET", *Proc. 4th ACM Int'l Workshop on Vehicular Ad Hoc Networks, VANET'07*, p 19–28.

[22] P. Kocher, "Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems", in *Lecture Notes in Computer Science 1996*, Vol. 1109, p 104-113.

[23] F. Standaert, T. Malkin, M. Yung, "A unified framework for the analysis of side-channel key recovery attacks", in Lecture Notes in Computer Science 2009,Vol. 5479, p 443-461.

[24] M. Raya, P. Papadimitratos, J. Hubaux, "Securing Vehicular Networks".

[25] P. Papadimitratos, V. Gligor, J. Hubaux, "Securing Vehicular Communications - Assumptions, Requirements, and Principles".

[26] E. Schoch, F. Kargl, M. Weber, "Communication Patterns in VANETs", *IEEE Communications*

Magazine 2008, p 119-125.

[27] K. Xue, P. Hong, X. Tie, “Using Security Context Pre-Transfer to Provide Security Handover Optimization for Vehicular Ad Hoc Networks”, Vehicular Technology Conference Fall (VTC 2010-Fall), 2010 IEEE , p 1-5.

[28] S. Atay, A. Koltuksuz2, H. Hışıl, “Computational Cost Analysis of Elliptic Curve Arithmetic”, Hybrid Information Technology, 2006. ICHIT '06, p 578-582.

An Advance ATM Machine Service: Making Demand Draft through ATM Machine

[1*]Mohd. Arif Siddique and [2]Dr. Amit Kumar Awasthi

[1]Department of Computer Science, Radha Govind Engineering College,Meerut (U. P.), India
E-mail: marifs2009@gmail.com
[2]Associate Professor, Department of Applied Science & Humanities
Pranveer Singh Institute of Technology, Kanpur (U.P.), India

Abstract
This paper suggests an advanced ATM machine service by which customer make Demand Draft by ATM (Automatic Teller Machine) machine without the need for a cashier, human clerk or bank teller. This technique can gives strength to the anywhere banking.

Keywords: Banking, Advanced ATM Services, Anytime and anywhere Banking, DD Exit Slot, DD through ATM, Paper Selector.

Introduction
The existing problem with Demand Draft (i.e. DD) which cannot be prepared on the time such as in night or off-days as well as customer cannot get Demand Draft at the spot. This paper suggests an Advanced ATM Service by a which customer can make Demand Draft by him self through ATM (Automatic Teller Machine) machine with proper authorization and without the need of a cashier, human clerk or bank teller. This is done by the E-Banking. Finally the objective of this paper is to create an add-on service for the On-Premise ATM Machine so that ATM machine can make Demand Draft. The basic idea is taken from [7] "ATM as video conferencing station".

1. The Problem with current DD-Making Process

- Customer has to manually visit the bank in working hours and fills the required form, and he has to wait for some authorization.
- DD making can be possible in bank working hours only.
- DD making can not be possible on the time such as in night or off days.
- Customer cannot get Demand Draft at the spot.
- Instead of this, many banks are offering Online Demand Drafts, but in this process customer have to wait 8 days after filling online DD request because it deliver by the currier.

2. Benefits of proposed technique

- Costumer becomes free from the rush of the bank.
- Anytime like in night or off days DD making is possible.
- From anywhere costumer can make DD.
- Customer can get DD at the spot.
- It is the better service than online DD.

3. Introduction of ATM Machine (in short ATM)

Now An Automated Teller Machine (ATM) [1] is well known machine, it is a computerized telecommunication device that provides the clients of a financial institution with access to financial transactions in a public space without the need for a cashier, human clerk or bank teller. The customer is identifies by inserting an ATM card. Authentication is provided by the customer entering a personal identification number (PIN). Using an ATM, customers can access their bank accounts in order to make cash withdrawals, credit card cash advances, and check their account balances, purchasing, booking tickets, etc.

ATMs are placed not only near or inside the premises of banks, but also in locations such as shopping centers/malls, airports, grocery stores, petrol/gas stations, restaurants, or any place large numbers of people may gather.

According to location there are two types of ATM machine: *On-Premise and Off-Premise* (also known as On-Site and Off-Site). *On-premise* ATM machines are typically more advanced, multi-function machines thus it is more expensive. Off premise machines do work of cashier, so it is typically the cheaper mono-function machine. Figure 1 is showing component of ATM machine. ATM machine has *LCD display* to show output messages, *Card reader* to read ATM card, *Keypad* to enter PIN and amount, *Cash Dispenser* to dispense case from ATM, printer to print transaction detail over the paper, Camera for security purpose.

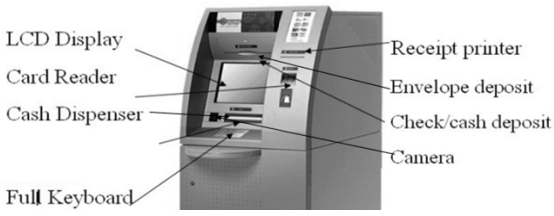


Figure 1: Components of ATM machine.

Demand draft
Demand Draft in the Indian context is essentially a written

order, where a buyer pre-deposits an amount of money in a bank and hands over the 'written order' to the seller, so that the seller can 'demand' the deposited money from the bank where it way deposited.

Methodology

This paper suggests the technique that can implement with minor the change in existing model of ATM Machine.

Brief Description of the proposed process

Costumer first login to the system by swapping his ATM card then system will ask for PIN as costumer enters valid PIN system shows various options on screen now user selects *DD Making* then system ask for some detail, as user enters all required details system adds some details with the detail given by the costumer and displays on screen and ask for verification, user when verifies all detail then printing of DD and transaction slip starts and dispense from exit slot. User now has DD and transaction slip

Figure 2: DD Making process.

Addition/Modification in existing component of ATM

Key Board

Currently most of the ATM has numeric keypad[1] (as shown in figure 3) but this paper suggesting new keypad like in mobile phone with two key 'Backspace' and 'Blank Space' (as shown in figure 4) by which customer can enter text data like *Beneficiary Name and Account Number, Branch* etc. just like writing SMS in mobile phone.

Figure 3: Current numeric keypad

Figure 4: Proposed alpha-numeric keypad

Printer and Paper Selector

The aim of this paper is to achieve goal with minor change in the design of existing ATM Machine so that hardware and software cost could not increase.

Here this paper does not suggest to change existing ATM Transaction Slip Printer (as shown in figure 5) and add new printer that can print DD as well as ATM transaction slip.

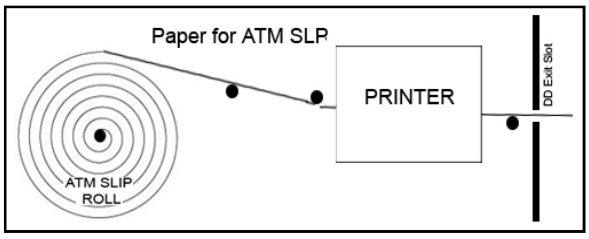


Figure 5: Current printer of ATM Machine with ATM Slip Roll

This paper proposed a new device '*Paper Selector*' (shown in figure 6), by the addition this new device *ATM slip printer* becomes able to print *Transaction Slip* as well as *DD*. Block diagram of printing device is shown in figure 4. In this figure we have a roll of paper for transaction slip and a stack of paper for DD, both papers are running over the roller and inserted into *Paper Selector*.

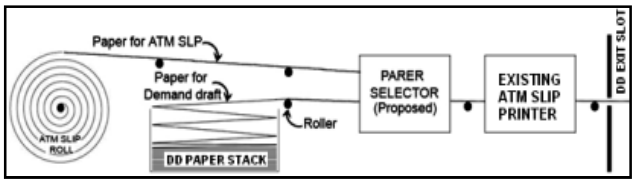


Figure 6: Suggested printer with Paper Selector.

Paper Selector has two roller sets say *roller set 1* and *roller set 2* and an *instruction translator*. *Instruction translator* receives two types of instructions that are "forward paper of transaction slip" and "forward paper of DD" form system and find what is to be done. Paper selector can be in one of the three stages

1. Idle position, see figure 7(a)
2. DD paper forwarding, see figure 7(b)
3. Transaction Slip forwarding, see figure 7(c)

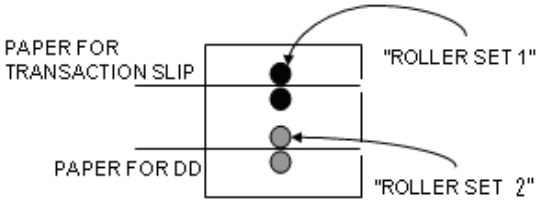


Figure 7(a): idle position of paper selector.

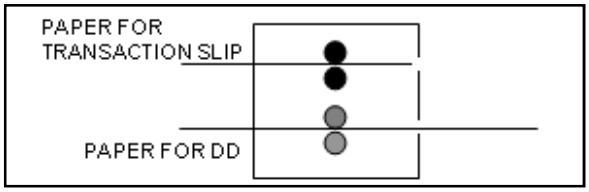


Figure 7(b): DD paper forwarding.

Figure 7(c): Transaction Slip forwarding.

In *idle position* both roller sets remains stop.

In the case of ***DD paper forwarding,*** when printer get instruction to ***print Demand Draft*** at the same time *instruction translator* receive "forward paper of DD" and then the roller set 2 runs and send paper from *DD paper stack* to the printer and at the same time roller set 1 remains stop. Now printer simply prints details of DD over the received paper.

In the case of ***Transaction slip forwarding***, when printer get instruction to ***print Transaction slip*** at the same time *instruction translator* receive "forward paper of Transaction Slip" and then the roller set 1 runs and send paper from ATM Slip roll to the printer. Roller set 2 remains stop. Now printer simply prints details of transaction over the received paper.

Demand Draft

Information needed to make DD

By the simple GUI user enters DD amount, Beneficiary Name, Payable Branch, Code of Branch through alphanumeric keypad just like typing an SMS on mobile phone.

Paper for DD

This paper suggest new layout for DD. Why we change the layout of DD? The answer is, first, Paper target is to make minor change in existing system. To use existing printer we have to change layout of DD from Landscape to portrait. Now DD can print with the existing ATM Transaction-Slip-Printer. Second, width of paper of DD must be same as the ATM Transaction Slip so that it can print by the printer and can exit from ATM Transaction Slip *exit slot* new slot is not required. All the security issues [2] like DD number, MICR number and Transaction code etc. is still available on the DD, suggested sample layout of DD is shown in figure 6.

Only two differences are between *off-line DD* and ATM generated DD, first, ATM generated DD has ATM Identification Number [3], Associated Branch of Bank [3], ATM Location [3] to recognize ATM that has printed this DD. Second, it has a digital signature for authentication.

DD Layout

In figure 8, Proposed DD divided into 4 areas First area has: First area has Date, Time, Issuing *ATM ID, Branch Name, Code Number, Phone Number* to which ATM is Associated (Cancellation of DD is only done in bank to which ATM machine is associate). Second area has: *Beneficiary Name, DD Amount (in Number and Word*). Third aria has: *Drawee Branch Name and Drawee Branch Code Number*. Fourth area has: All security issues [2] (as per the requirement of bank) like *DD Number, MICR Number and Transaction Code, Authority Digital Signature*, etc.

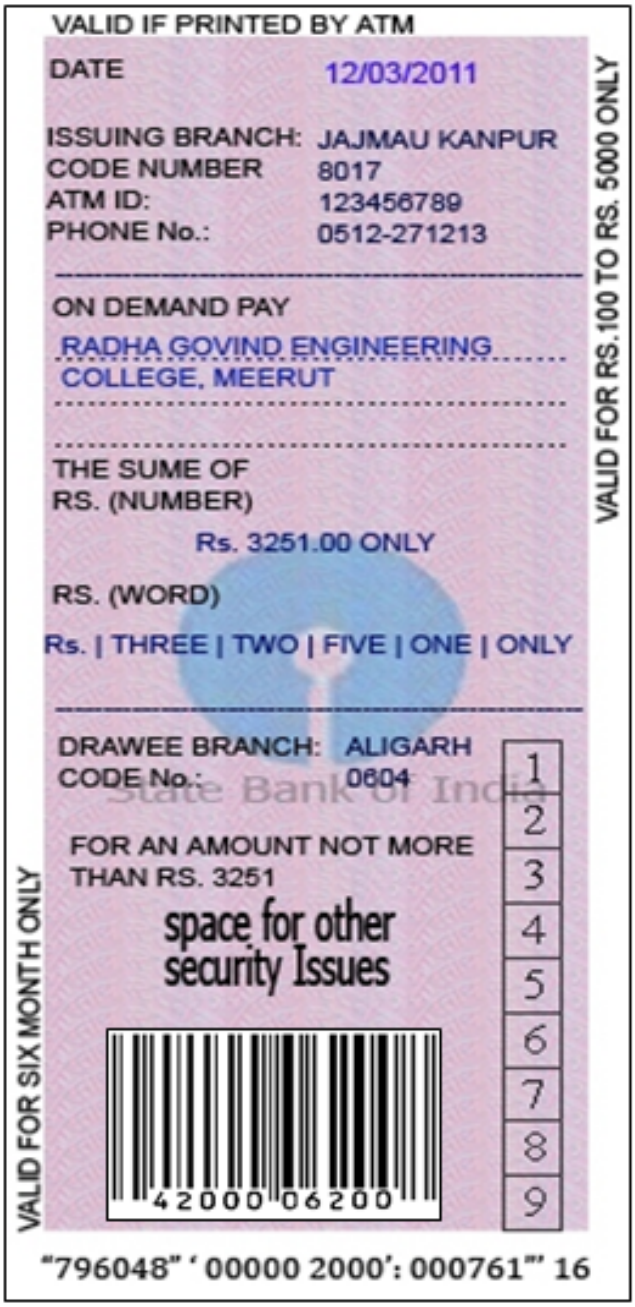


Figure 8: Sample Demand draft

Transaction Slip

Proposed sample transaction slip is shown in figure 9. It has: *Date, Time, Issuing ATM ID, Branch Name, Branch Code Number, Debit Card Number, transaction Number, Beneficiary Name(in very short), DD Number, DD Amount, Deducted DD making Charge, security code, Available A/C Balance.* And A massage "*Note: In case of cancellation of DD card holder must present personally with any ID card, this slip and DD in original".* With all these suggested information some space is still for the other information that is as per requirement of bank.

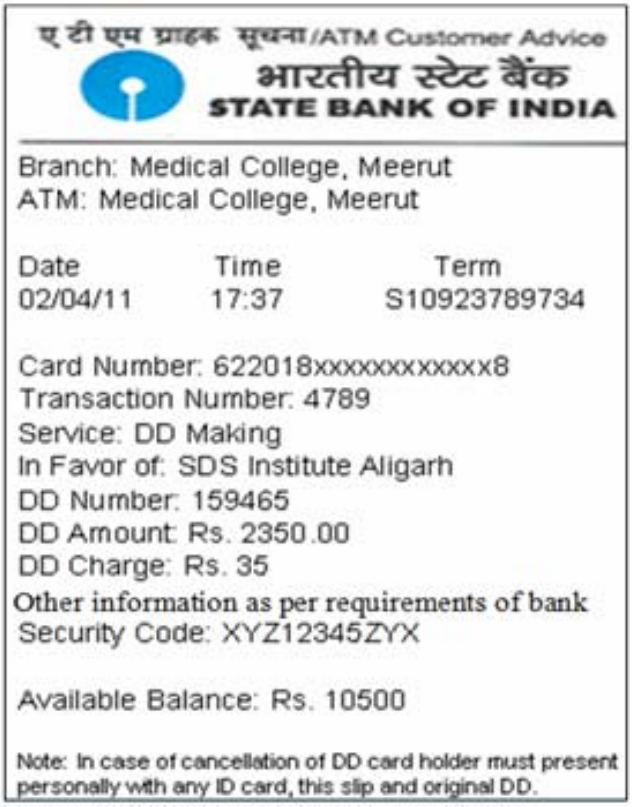


Figure 9: Sample ATM transactions slip.

DD Exiting Slot

As per discussion ATM Transaction-slip printer will print DD so the DD can exit from the slip-exit slot of ATM. There is no need of new exit slot.

DD Making Process

The DD making process is illustrated in following algorithm.

Algorithm DD-Making

Input: ATM Card and PIN, Beneficiary Name, DD Amount, Payable Branch Name and code.

Output: Printed DD and transaction Slip

Step 1: User login by swapping his ATM card and entering PIN

Step 2: System display option window.

Step 3: User selects DD Making option from various options.

Step 4: screen display DD Making window and ask for entering details

Step 5: User enters Beneficiary Name, DD Amount, Payable Branch Name, and Code of Branch.

Step 6: Screen displays all entered information with Date, Time, Issuing ATM ID, *Branch Name, Branch Code, DD Number, DD Amount, and Deducted Charge.* Ask for confirmation.

Step 5: If user presses 'yes' // either by pressing side button or touching screen

{

If AC-Balance >= DD Amount + DD Making Charge

{

Screen display 'Pay?'

If user select 'yes'

{

System display "DD Making is started, Please wait!" and system sends a signal to the *paper selector*,

If papers for DD and slip are available

{

AC Balance = AC Balance – (DD Amount + DD Charges)

Display "DD is printing, collect it from exit slot".

Paper selector sends paper to the printer.

As printer received paper and detail of DD it just prints all following details on received paper:

Current Date and Time

Issuing ATM ID, Branch Name, Code, Address where ATM is associated

Authority Digital Signature

Amount in digit and words for example if DD amount is Rs. 2563/- Then it is *Rs.|Two|Five|SIX|THREE|ONLY* in word

Other security issues

}

Else Display "Paper for DD is not available" and exit.

}

Else Display "Transaction cancels by the user" and exit.

}

Else Display "Account balance is not enough" and exit.

Display "Collect Transaction Slip"

}

Else goto step 4 // user wants some editing

Step 6: Exit

DD CANCELLATION PROCESS

Algorithm DD-Cancellation

Input: Card holder must be present personally in bank with any ID card, ATM Card, ATM transaction slip for DD and original Demand Draft.

Output: Refund DD amount.

Step 1: Authenticate Customer by any ID proof and ATM card.

Step 2: Authenticate DD by the security issues

Then goto step 5 else exit.

Step 4: Original DD and ATM receipt collected by the bank correspondent**.**

Step 5: Make a credit equal to DD Amount - DD Cancellation charge.

Step 6: Exit

DD Modification Process

Algorithm DD-Modification

Input: Card holder must be present personally in bank with any ID card, ATM Card, ATM transaction slip for DD and original Demand Draft.

Output: Refund DD amount.

Step 1: Customer have to cancel current DD as discussed in Algorithm DD-Cancellation

Step 2: Remake DD from ATM.

Signing Authority

In manual case Branch Manager is the Signing Authority that can sign over the DD. But in case of DD making form ATM Machine, Two questions are arising first, who is the Signing Authority? And second, how he/she can sign over the DD?. The answer of the first question is still Signing Authority is Branch Manager of branch where ATM machine is associated. Answer of second question is Proxy Blind Signature Schemes [5].

D. Chaum [5] introduced the concept of Blind Signature scheme in 1982. Using this schemes one user can obtained the signature of another on any given message, without revealing any in formation about the message or its signature. This scheme also ensures untractability and unlinkability. In 1996 Mambo et al [6] introduced the concept of proxy signature. In this scheme an original signer delegates his signing authority to another (proxy) signer in such a way that the proxy signer can sign any message on behalf of the original signer and the verifier can verify and distinguish between normal (original) signature and proxy signature.

In proposed method, we use the concept of "Proxy Blind Signature Schemes [5]". See figure 10, Branch Manager **M** of any Branch **B** has its own signature given/verified by the *Central Security System* of concern *financial institution.* '**n**' ATM machines $\mathbf{A_1, A_2, \ldots, A_n}$ are associated with this branch **B**. Each machine has its own proxy blind signature generated by the *signature* of Branch Manager **M**. When a customer makes DD form any ATM Machine $\mathbf{A_i}$ then the proxy signature of that ATM Machine printed over the DD in the form of Barcode.

During authentication, this barcode (proxy signature on DD) is used to authenticate proxy signature of ATM Machine $\mathbf{A_i}$ i.e signature of **M**.

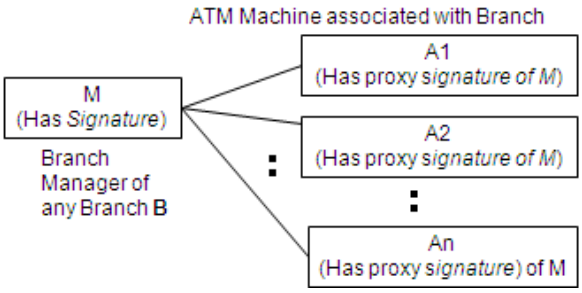


Figure 10: Signature and Proxy signature.

Transaction

The e-Commerce modules as e-Transactions and e-Banking are used to make transaction for DD making, DD cancellation. Transaction for DD making is just like a like withdrawal an amount equals to DD amount plus DD Making charge. it means DD Amount and DD Charge will deducted from account balance, with a particular 'DD Making'. Similarly transaction for DD cancellation is just like a like deposit an amount equals to DD amount minus DD cancellation charge. Security related issues remain same during above discussed transaction.

Application

- DD is the popular [4] Payment Instruments in India over the cheque, because A cheque is not a cash, as it does not assume the finality of payment. The funds may not be available with the drawer or the drawer may have withdrawn funds from his bank account in the interim leading to the possibility of the cheque being dishonoured on presentation.
- Cheque is not always acceptable in several business transactions particularly where the drawer and the payee are not known to each other then there is use of DD.
- DD is a popular medium to pay examination fee.
- Because of these advantages, implementation of proposed technique is beneficial.

Conclusions and Future Scope

In this paper, we present our approach of making demand draft from ATM machine for improving anytime/anywhere banking. Proposed approach involves using of paper selector so that existing printer can print DD as well. New layout of DD seems difficult in first view but you know this is the printed material and it can print portrait beside landscape on same cost. These minor changes are not expensive. For authentication of DD we are using digital proxy signature. In cancellation of DD we suggested remaking of DD because once DD has printed no editing is allowed. Transaction of DD making is similar to withdraw amount equals to DD amount plus DD making charges and transaction of DD cancellation is similar to deposit amount equals to DD amount minus DD cancellation charges. Refund of the cancellation can not be given to the customer in cash; it is automatically credited in the account of customer as in the railway ticket cancellation. This prevents the fraud. All security related to transaction is unchanged. Digital signature of signing authority is the area for future work.

Acknowledgments

We thank Dr. Mohd. Sahiq Khan, Dr. K. Prasad Pamulapati, Dr. Ajay Singh for their wonderful comments and suggestions.

References

[1] http://en.wikipedia.org/wiki/Automated_teller_machine
[2] http://www.sriraj.org/misc/demand-draft-micr-code-sbi-other-numbers-charges/
[3] http://www.sbhyd.com/services_atmlocations.asp
[4] http://rbidocs.rbi.org.in/rdocs/Publications/DOCs/4453.doc
[5] D. Chaum, Blind signatures for untraceable payments, Advances in Cryptology Crypto 82 Plenum Press (1982), 199-203.
[6] M. Mambo, K. Usuda, and E. Okamoto, Proxy signatures: Delegation of the power to sign messages, IEICE Trans. Fundamentals E79-A, no. 9.
[7] Mark M Grossi , Grant C Paton, George E Schneider ATM as video conferencing station.

Analysis of Downlink Scheduling for Network Coverage for Wireless Systems with Multiple Antenna

[1]Harish Kumar, [1]Manish Kumar and [2]Pushpneel Verma

[1]Department of ECE, [2]Department of CSE,
Bhagwant Institute of Technology, Muzaffarnagar, India
E-mail: hc78@rediffmail.com, pushpneelverma@gmail.com

Abstract
In this paper, the focus on the downlink scheduling design optimized for network coverage for wireless systems with multiple antennas. & proposed a wireless system consisting of a base station (with n_T transmit antennas) and K client mobiles (each with single antenna). with multiple antennas, we proposed a systematic frame-work based on information the critical approach and formulate the scheduling design as a mixed convex and combinational optimization problem. network coverage is based on the maximum cell-radius such that the outage probability of a single user at the cell edge (at a target bit rate) is below a specified target P_{out}, The scheduling algorithm in the MAC layer is responsible for the allocation of channel resource at every fading block. The system resource is partitioned into short frames. We assume time division duplexing (TDD) systems so that channel reciprocal holds. At the beginning of every frame, the base station estimates the channel matrix from the participating mobile users. The uplink channel estimation is used as the downlink channel information. was found that network coverage could also benefit from the *multi-user selection diversity* through wireless scheduling.

Index Terms: *MIMO,* Coverage-capacity tradeoff. Cross layer, Wireless

Introduction

It is well-known that cross-layer scheduling could achieve significant performance gain in *network capacity due to multi-user selection diversity.* Basically, system resource is allocated adaptively to user (s) with the best channel condition. There have been a lot of works that try to take advantage of the cross-layer optimization by cooperative scheduling which takes the link-level metrics into scheduling decisions which takes the ink-level metrics into scheduling decisions [1]. In [2]. [3]. it is shown that maximizing the link diversity order (which maximizes the link capacity) does not always result in the optimal system capacity. Therefore. joint optimization to the system level performance of wireless systems and the scheduling design optimized for network capacity has been relatively well-studied.

While system capacity is an important measure to optimize. System coverage is also an important dimension to consider. especially during the initial deployment stage where network coverage is usually the bottleneck. The advantage of cross-layer scheduling on the *network coverage* is a relatively unexplored subject. Conventional concept of network coverage is based on the maximum cell-radius such that the outage probability of a single user at the cell edge (at a target bit rate) is below a specified target P_{out}, In [4] the coverage performance of an uplink scheduling algorithm has been studied. It is found that network coverage could also benefit from the *multi-user selection diversity* through wireless scheduling. However. the scheduling algorithm considered is restricted to selecting one user at a time. Hence, the analysis does not generalize to the general case with multiple antennas where *spatial multiplexing* allows selecting multiple transmission at any scheduling slot. Moreover, the design of the scheduling algorithm is heuristic and it is not clear what is the optimal scheduling design with respect to network coverage. In this paper, we shall focus on the downlink scheduling design optimized for network coverage for wireless systems with multiple antennas.

In this paper we consider a wireless system consisting of a base station (with n_T transmit antennas) and K client mobiles (each with single antenna). With multiple antennas, we have additional degrees of freedom for *spatial multiplexing* and *spatial diversity* To include the spatial multiplexing into the framework, we first extends the conventional concept of network coverage to a more general *utility-based network coverage* to deal with the possibility of allocating resource to multiple users at the same time. We consider the *network centric utility* and the *user centric utility* as two examples of the utility based coverage concept. Raised on the generalized concept of network coverage, we propose a systematic frame-work based on information the critical approach and formulate the scheduling design as a mixed convex and combinational optimization problem. Due to the huge search space. The complexity of the optimal algorithm is enormous. We consider a *genetic-based* scheduling [5]. which offers a reasonable complexity-performance tradeoff.

This paper is organized as follows. In section II. we shall outline the multi-user forward link physical layer model as well as the MAC layer model. In section III. we shall define the *utility-based network coverage* and formulate the downlink scheduling design optimized for network coverage. Optimal solution is outlined. In section IV. we shall introduce the *genetic-based* algorithm. In section V. we present numerical results to evaluate the performance of the optimal and genetic based schedulers. Finally. we conclude with a brief summary of results in section VI

Overview of System Model with Multiple Transmit Antennas

We first elaborate the multi-user for ward link channel model with multiple antennas, the physical layer model and the MAC layer model below. To decouple the data source statistics from the system performance, we shall assume that source buffers are large in size so that they always contain source packets waiting to be transmitted. In other words, there will be no empty scheduling slots due to insufficient source packets at the buffer.

Downlink Channel Model

We consider a communication system with *K* mobile users having single receive antenna and a base station with n_T transmit antennas. The microscopic channel fading between different users and different antennas is modeled as i.i.d. complex Gaussian distribution with unit variance. Furthermore, it is assumed that the encoding and decoding frame are short bursts which are much shorter than the coherence time of the fading channel.

Let Y_k be the received signal of the *k*-th mobile. The *K* x *1* dimension vector of received signal Y at the *K* mobile stations is given by

$$Y = \begin{bmatrix} Y_1 \\ . \\ . \\ . \\ Y_K \end{bmatrix} = \begin{bmatrix} \sqrt{L_1} H_1 \\ . \\ . \\ . \\ \sqrt{L_K} H_K \end{bmatrix} X + \begin{bmatrix} Z_1 \\ . \\ . \\ . \\ Z_K \end{bmatrix} \quad (1)$$

where X is the n_T x 1 transmit symbol from the base station to the *K* mobiles, Z_k is the complex Gaussian noise with variance σ_z^2, H_k is the 1 x n_T dimension channel matrix between the n_T transmit antennas (at the base station) and the *k*-th mobile and L_k is the path loss between the base station and the *k*-th mobile. The entries of H_k are i.i.d. zero mean complex Gaussian with unit variance. The path gain [6] is given by (2) at the top of the next page, where d_k denotes the distance between the base station an the *k*-th mobile, G_t, G_r are transmit and receive antenna gains and λ is the wavelength of the carrier.

Multi-user Physical Layer Model

Before we could discuss the scheduling optimization problem, it is very important to define the physical layer model because different physical layer implementations will definitely affect the system level performance. To isolate the physical layer performance from specific implementation details (such as channel coding and modulation, multiple access schemes like TDMA, FDMA, CDMA), and information theoretical approach is adopted Specifically. the maximum achievable rate at the physical layer (with arbitrarily low error probability) is given by the Shannon's capacity and is realized by random codebook and Gaussian modulation.

For simple processing, we shall adopt a simple downlink ero-forcing (ZF) scheme[2] at the base station [5], [7] as illustrated in Figure 1.

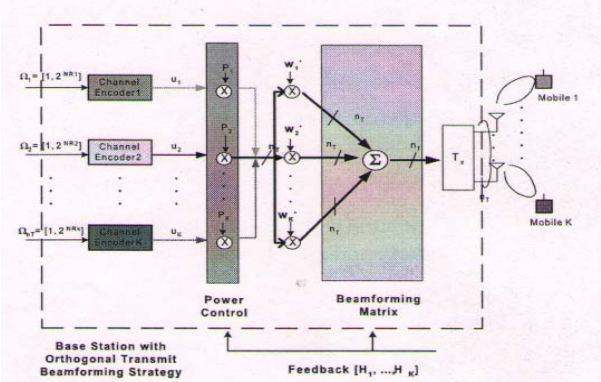


Fig. 1. Orthogonal Transmit Beam-Forming strategy with multiple antennas.

There are *K* streams of information data to *K* individual users at the base station transmitter. They are channel encoded independently. The vector $(K \times 1)$ of encoded symbols, $U = [U_1, ..., U_K]$, are processed by the power control diagonal matrix $(K \times K) P = diag(p_1, ... p_k)$ followed by the ZF matrix $(n_T \times K), W = [w_1, ... w_K]$ where w_k Is the $n_T \times 1$ complex ZF weight of user *k*. Hence, the transmitted vector of symbol, X. is given by :

$$X = w\sqrt{PU} = \sum_{k=1}^{K} \sqrt{p_k} U_k w_k \quad (3)$$

Where $p_k \geq 0$ is the average transmit power during the current scheduling instance for user *k* and $\varepsilon[|U_k|^2] = 1.$ Since encoding frame is short burst with quasi-static fading, no power adaptation within an encoding frame is needed. At any scheduling slot. Individual user(s) could be turned off by assigning $p_k = 0$. an *admissible set* is defined as a set of user indices with non-zero transmit power $A = \{k \in [1, K] : pk > 0\}$. The total transmit power out of the base station at any scheduling slot is constrained by P_0 That is:

$$\sum_k P_k \leq P_0 \quad (4)$$

Calculation of the ZF Weights : Given an *admissible set*, the transmit power $\{p_1, \ldots p_k\}$, and a realization of the channel fading $\{h_1 \ldots . h_k\}$. the received signal of user *k* is given by :

$$Y_k = \underbrace{\sqrt{p_k L_k} h_k w_k u_k}_{Information} + \underbrace{\sum_{J \in A, J \neq k} \sqrt{p_j L_j} h_k w_j U_j}_{Multi-beamInterference} + Z_k \quad (5)$$

Where the first term contains the desired signal and the middle term represents the *multi-beam interference* due to simultaneous transmission of independent information streams. The OTBF weight, w_k is selected to satisfy:

$$w_k^* w_k = 1 \forall k \quad (6)$$

And the *orthogonal conditions*

$$h_j w_k = 0 \forall j \in A, j \neq k \quad (7)$$

where A denotes *admissible set* of users with non-zer allocated power. The operator of * means complex conjugate transpose. Note that when p_k= 0, the information stream for user *k* is turned off. In other words, the number of simultaneous transmission is given by the cardinality of the admissible set A. Observe that there are $2n_T$ degree of freedom in w_k and there are 2|A|-1 equation from the constraints (6) and (7). Hence, we have.

$$|A| \leq n_T \quad (8)$$

This means that with n_T transmit antennas, the base station could support at most n_T spatial channels. The remaining degrees of freedom is utilized to maximize $w_k^*(h_k^* h_k) w_k$. Please refer to [5] for the solution of the ZF weight.

Date rates supported by the Orthogonal Spatial Channels : With the ZF weights $\{w_k\}$, the *multi-beam interference* becomes zero and there are |A| independent spatial channels. The received signal for mobile user *k* is given by.

$$Y_k = \sqrt{p_k L_k} H_k w_k U_k + Z_k \quad (9)$$

Hence, the maximum achievable date rate of the *k*-th spatial channel during the fading block is given by the maximum mutual information between U_k and Y_k and is given by :

$$T_k = \log_2 \left(1 + \frac{p_k L_k | h_k w_k |^2}{\sigma_z^2} \right) \quad (10)$$

MAC Layer Model

The scheduling algorithm in the MAC layer is responsible for the allocation of channel resource at every fading block. The system resource is partitioned into short frames. We assume time division duplexing (TDD) systems so that channel reciprocal holds. At the beginning of every frame, the base station estimates the channel matrix from the participating mobile users. The uplink channel estimation is used as the downlink channel information. Due to short burst transmissions, the channel estimation is used as the downlink channel information. Due to short burst transmissions, the channel fading remains the same across the entire burst duration. The estimated CSI is passed to the scheduling algorithms in the MAC layer. The output of the scheduler consists of an admissible set. $A = \{k \in [1, K]; p_k > 0\}$ (the set of user indices with non-zero power allocated at the current fading block). the corresponding power allocation $\{p_k\}$ and the *instantaneous rate* allocation $\{r_k\}$of the selected users. The downlink payload is transmitted at the scheduled rate and the rate is also broadcast on the downlink common channels to mobile users.

General Formulation of the Downlink Scheduler Design

Before we proceed with the scheduler design, it is important to quantify what is meant by *system performance*. In multi user systems. the system performance is can be defined by an *instantaneous utility* $G(r_1, ... r_K)$ where r_k is the instantaneous data rate of the *k*-th user. For meaningful optimization. we have:

$$\frac{\partial G}{\partial r_k} > 0 \forall r_k \geq 0 \quad (11)$$

Utility-Based Network Coverage

Consider a sequence of *N* realization of fading blocks, the instantaneous achievable data rate (s) of a scheduled mobile user (s) are random variables (functions of the specific fading realization) in that fading block. In conventional wireless systems where the scheduling is constrained to select one active user at any scheduling instance (fading block), *outage* is defined as the event that the instantaneous data rate of the scheduled user is below a target data rate R_0. *Outage probability* is defined as the likelihood of the outage event, $P_{out}(R_0) = \lim_{N \to \infty} N_{outage}(R_0) / N$. Traditional concept of network coverage is there fore defined as the maximum distance between a mobile and the base station such that the outage probability of the scheduled user (at a target bit rate, R_0) is below a specified target $P_{out}(R_0)$. That is, coverage is given by the maximum distance, $d(P_{out}, R_0)$, such that

$$\Pr[r(d) \leq R_0] \leq P_{out} \quad (12)$$

where *r(d)* denotes the instantaneous data rate of any user at a distance *d* from the base station. This is a commonly employed definition in the cellular systems [6] as well as wireless systems with scheduler constrained to select one user at a time [4].

However, when the base station has multiple transmit antennas, there are multiple spatial channels as a result of the additional degrees of freedom. This implies that a general multiple antenna scheduler must allow more than one active transmission at any scheduling slot (fading block), Hence, before we could discuss the design of schedulers optimized for network coverage, we must extend the definition of coverage to include multiple active user transmissions at a time.

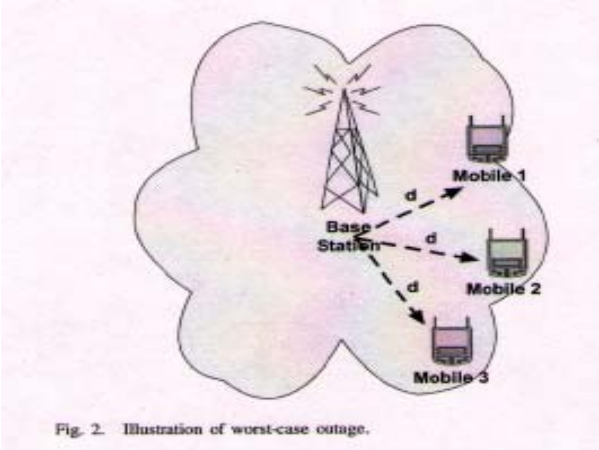


Fig. 2. Illustration of worst-case outage.

Coverage-Optimized Scheduling Problem Formulation

Observe that given a fixed target utility value G_0, the *utility-based coverage is maximized if the worst case out-age probability* is minimized. Hence, the coverage-optimized scheduler design is formulated below and is a mixed concave optimization and combinatorial search problem.

Examples of Utility Functions

The utility-based coverage is a general concept and depending on the specific forms of the utility function, the *coverage* can have very different physical interpretation. In this section we shall illustrate the concept of utility-based coverage using two examples, namely the *network-centric coverage* and the *user-centric coverage.*

Network Centric Coverage :The generalized coverage is called *network centric* if the utility function is given by :

$$G_{network}(r_1...r_K) = \sum_{k\in A} r_k \tag{15}$$

The corresponding outage event will be interpreted as the situation when the total instantaneous sum throughput $\sum_k r_k$ corresponding coverage (defined based on the outage event) is not measuring what a single user gets but is measuring how the network deliver as a whole but on how much the *worst user* gets. Hence, the physical meaning of the outage and coverage is *Network centric.*

User Centric Coverage : The generalized coverage is called *user centric* if the utility function is given by :

$$G_{user}(r_1,...., r_k) = \min_{k\in A} r_k \tag{16}$$

In this case, the outage event is determined by the *worst case* user instead of the contribution from all selected users. The outage is measured not based on how much the network delivers as a whole but on how much the *worst user* gets. Hence, the physical meaning of the outage and coverage is *user centric.*

Optimal Solution-Mixed Combinatorial and Convex Programming

Since the optimization problem involves minimizing $\Pr[G(r_1,...,r_K) \le G_0]$, which is not very easy to deal with, we have the following lemma which establishes the equivalence between minimizing the worst-case outage and maximizing the utility function $G(r_1...,T_K)$.

Step 1. For every possible admissible set A, we compute the optimal power allocation $(p_1,....p_K)$ for those selected users.

Since r_k is a function of weights $w_1,....w_K$ and $(p_{1,....}\ p_K)$ and since the weights are also functions of the admissible set A, we could express the utility function $G(r_1,....r_K) = G(A; p_1..., p_K)$ *as* :

$$G(r_1,..., r_K) = G(A; p_1,....p_K) \tag{18}$$

The optimal power allocation with respect to the *network centric utility,* $Gnetwork(r_1, r_2,...., r_K)$ *. as well as the user centric utility,* $G_{user}(r_1...r_K)$, *are both given by:*

$$p_k^* = \left(\frac{1}{\lambda} - \frac{1}{L_k \mid h_k w_k \mid^2}\right)^+ \tag{19}$$

for all $k \in A$ and λ is the Lagrange multiplier given by the solution of the equation $\sum_{k\in A} P_k^* = P_0$.

Step 2 : Repeat step 1 to compute the utility functions for all other possible admissible set A.

For the *network-centric* utility. we have the following lemma about the admissible set A.

$G_{network}(r_1,...r_K)$ *satisfies* $|A| = n_T$.

On the other hand, we have the following lemma about the admissible set A with respect to the *user-centric utility.*

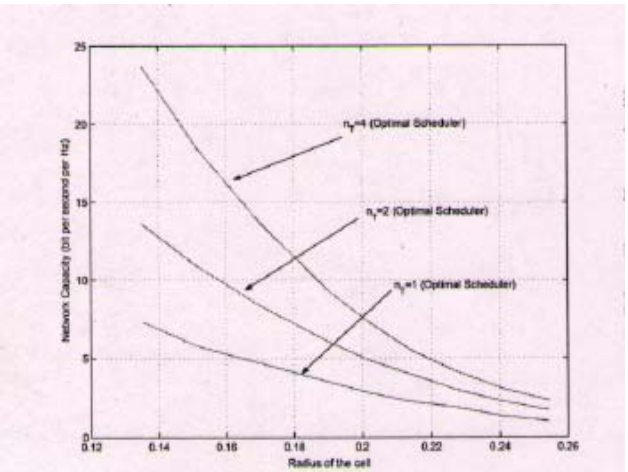


Fig. 3. Network coverage (km) versus network capacity (b/s/Hz) of coverage-optimized scheduler at $n_T = 1, 2, 4$. Total transmit power of the base station is 30dBm. Path loss exponent is 4. Total number of users is 10. The network coverage refers to the *network centric coverage* at a target outage of 1%.

Genetic-Based Coverage-Optimized Scheduler Design

The computational complexity of the optimal algorithm in general exceeds the implementation limitation in most designs for moderate *K* and n_T. In this section. we shall introduce a real-time *genetic algorithm* [5]. The main template of genetic algorithm is illustrated below.

Algorithm 1. Step 1-Initialization : *Initialize a population with* N_P *chromosomes (A chromosome is a sample of the optimizing variable* $(\alpha_1,...\alpha_K)$ *where* $\alpha_k \in \{0,1\}$*). These chromosomes are randomly picked, satisfying the constraint:* $\sum_{k=1}^{K} \alpha_k \le n_T$.

$$P_m = \frac{1}{\beta_1 + \beta_2 \sigma_G / G} \tag{20}$$

Where σ_G is the standard derivation of the fitness of the current population (before selection). β_1 and β_2 are two constants.

These two processes introduce randomness into the intermediate generation so that the new population will be a combination of the best chromosome in the current population as well as some new random elements.

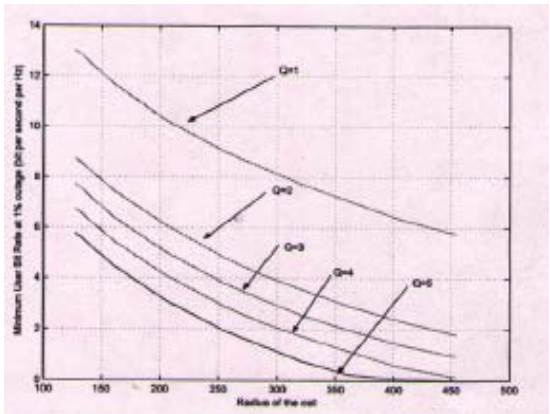


Fig. 4. Minimum scheduled data rate (b/s/Hz) versus cell coverage (m) at various $Q = 1, .., 5$ where $Q = |A|$. Total transmit power of the base station is 30dBm. Path loss exponent is 4. Total number of users is 10. The cell coverage refers to the *user centric coverage* at a target outage of 1%.

Step 4: Termination :*Replace the original population with the new population and repeat Step 2 and Step 3 until the number of interactions reaches N_g. When forming new population, it is ensured that the fitness chromosome in the current population is saved and inserted into the next population. And all members of the next population is checked against the constraint* $\sum_k \alpha \, \mathrm{k} \le n_T$. *If any chromosome violates this constraint, 'O' is inserted into a randomly selected bit position in the violating chromosome until the constraint is satisfied.*

The computation complexity of the genetic algorithm is bounded by $N_g \times N_p$ function evolutions. As will be illustrated in the next section, this represents enormous computational saving compared with the optimal algorithm.

$^4\sigma_G$ is an indication of the population convergence because a population converging onto a local optimal solution will have small σ_G while a population before convergence will have large σ_G. Hence, we would like to reduce the randomness introduced through mutation p_m to speed up convergence when σ_G is large but increase the mutation probability p_m to avoid getting stuck at local optimal points when σ_G is large but increase the mutation probability P_m to speed up convergence when σ_G is large but increase the mutation probability P_m to avoid getting stuck at local optimal points when σ_G is small. In this paper. We use β_1 =1.2 β_2 = 10.

Numerical Results and Discussions

In this section, we shall compare the performance of the downlink schedulers optimized for *utility-based coverage*, we shall investigate the contributions of *multi-user selection diversity and spatial multiplexing* to the network coverage. To highlight the contribution of *multi-user selection diversity*, we compare the coverage with respect to the *random scheduler*, where n_T users are randomly selected irrespective of their channel matrices at every fading block. To highlight the contribution of *spatial multiplexing*. We compare the coverage with respect to various n_T. Furthermore, we shall compare the effectiveness and performance-complexity tradeoff of the *genetic algorithm*

In the simulation, each data point consists of 5000 realizations of channel fading. Channel fading of the *K* users are generated based on independent complex Gaussian distribution (with unit variance). We assume 0dB antenna gain in the transmit and receive antenna. The carrier frequency is assumed to be 2GHz. Data rate is expressed in terms of bits per second per Hz.

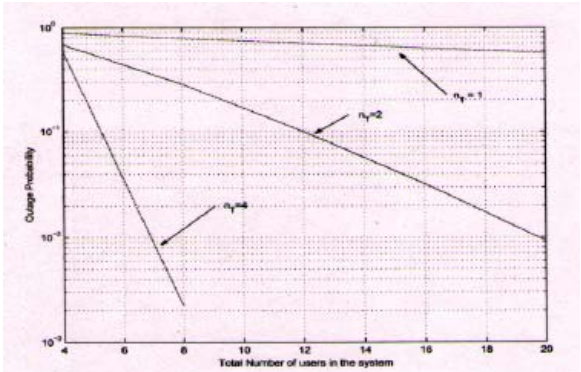


Fig. 5. Outage Probability versus Number of users *K*. Target network capacity is 7.0 b/s/Hz. Cell radius is 0.19 km. The transmit power at the base station is 30 dBm. The outage event refers to the *network centric utility* at a target outage of 1%.

Contribution of Spatial Multiplexing/Spatial Diversity to Network Coverage

Figure 3 illustrates the *network centric* coverage versus the network capacity (network loading) of the coverage-optimized scheduler at N_T = 1.2, 4. We observe that a significant gain in network coverage is achieved by increasing the number of transmit antenna n_T at high SNR and this illustrates the contribution of spatial multiplexing to network coverage. For example, in Figure 3, there are 50% and 90% area coverage gain at target network capacity of 5b/s/Hz comparing relative to N_T=*1* for n_T=2, 4 respectively.

On the other hand, Figure 4 illustrates the user-centric utility $G_{user}(r_1,...r_K) = \min_{k \in A} r_k$ threshold versus cell radius at various Q=1, 2…., 5 and n_T=5 where Q=|A|. The user centric utility threshold G_0 we considered refers to the threshold at 1% outage probability where outage is defined as the event $G_{user}(r_1,...r_K) < G_0$. Hence a point (*x.y*) in the graph means that over 99% of the time, scheduled user at xm from the base station will be able to transmit at least *y* b/s/Hz. We observe that to achieve high coverage gain (w.r.t. user centric utility), the space time scheduler should exploit the spatial diversity (Q=1) instead of spatial multiplexing (Q=5). For example, there is a 13 times are a coverage gain at target network capacity of 5b/s/Hz relative to Q=1 and Q=5 respectively.

Contribution of Multi-user Selection Diversity to Network Coverage

Figure 6(a-b) illustrates the capacity versus coverage between optimal scheduling and random scheduling at n_T = 1.4. For example, there is 3.2 times are coverage gain between the optimal scheduler and the random scheduler at n_T=4 and target network capacity of 2 b/s/Hz. This illustrates that multi-user selection diversity contributes significantly to coverage area gains.

Figure 5 illustrates the worst-case outage probability vs number of users *K* at various n_T Observe that as *K* increases. The efficiency of multi-user selection diversity increases because at any scheduling instance, it is more likely to select user with good channel conditions. Hence. The outage probability is reduced at the same target network loading (network capacity). Yet, supporting a large *K* would induce a large signaling overhead for channel estimation at the base station. In practice. *K*=20. could deliver a majority of the multi-user selection diversity gain for n_T = 4.

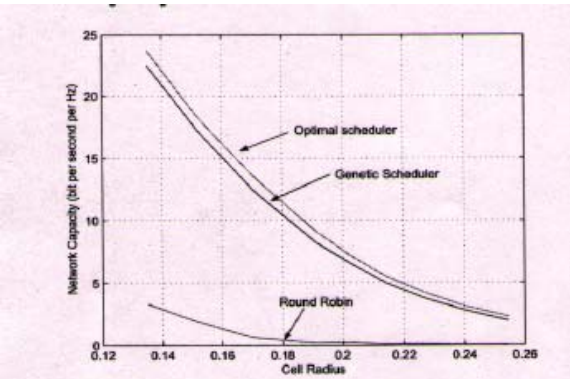


Fig. 6. Coverage (km) Performance of Genetic Downlink Schedulers at n_T = 4. Base station transmit power is set at 30dBm. Path loss exponent is 4. The network coverage refers to the *network centric coverage* at a target outage of 1%.

Performance Comparisons of the Genetic Algorithms

Figures 6 illustrates the performance of the genetic algorithm. The genetic algorithm has relatively small performance loss compared with the optimal scheduler.

Complexity Comparisons

Table I compares the number of function evaluations of the optimal algorithm and the genetic algorithm at various n_T and *K*. Observe that there is a 8 times and 36 times saving in computation of genetic algorithm (compared with the optimal algorithm) when ($K.n_T$) = (10.4) and (20.4) respectively, Furthermore, the MFPS (Million Function evaluations Per Second) requirement of the genetic algorithm is still within implementation limit. Hence the genetic algorithm may be used as real time scheduling algorithm.

Conclusion

In this paper, we focus on the design of *coverage-optimized* downlink scheduler for systems with multiple antennas. A systematic framework is proposed for the scheduler design

Problem based on information theoretical approach. There are *K* mobiles with single receive antenna and one base station with n_T transmit antennas. We proposed a generalized definition of network coverage, namely the *utility-based coverage*. The contributions of *multi-user selection diversity* and *spatial multiplexing* to the network coverage are analyzed. While the optimal scheduler delivers the best coverage, the computation complexity is huge. A real-time genetic algorithm is proposed, which offers enormous computational savings compared with the optimal algorithm and is an attractive candidate given it's performance complexity tradeoff.

Table 1 : Comparison Of Computational Complexity (In Terms Of Number Of Function Evaluations) Of Genetic And Optimal Algorithms. Is The Genetic Column. The Format Of The Results Is $N_P \times N_g$. The Brackets Indicates The Meps Assuming 2ms Packet Duration. [Meps= Million Function Evaluation Per Second].

(k, n_T)	Genetic Algorithm	Optimal Algorithm
(10.2)	10x2=20 [0.01]	55[0.027]
(10.4)	10x5=50 [0.025]	385[0.194]
(20.2)	10x5=50[0.025]	210[0.150]
(20.4)	20x5=100 [0.05]	3645[1.32]

References

[1] A. Jalali. R. Padovani, and R. Pankaj. "Data throughput of CDMA. HDR a high efficiency-high data. rate personal communication wireless. System." In IEEE Trans. Veh. Technol. Vol. 28 no. 1 pp. –Jan. 2000.

[2] K.N. Lau, Y.J. Liu, and T.A. Chen, "Optimal space-time scheduling for wireless communications with partial power feedback, " *Bell Labs Technical J.*, Nov. 2002.

[3] W. Yu, W. Rhee, and J.M. Cioffi, "Optimal power control in multiple acess fading channels with multiple antennas." In *Proc ICC 2001*.

[4] C.J. Chen and L.C. Wang, "Coverage and capacity enhancement in multiuser MIMO systems with scheduling, " in *Proc.* IEEE Clobecom 2004. *pp* - 1024-1038

[5] K.N. Lau, "Optimal down space time scheduling for wireless systems with umultiple antennas." IEEE Trans. Veh. Technol. Vol 54, pp. 1322-1333 July 2005.

[6] T.S. Rappaport *Wireless Communications : Principles and Practice.* Prentice Hall, 1996.

[7] H. Viswanathan. S. Venkatesan. And H.Huang. "Downlink throughput evaluation using multiple antennas in packet data cellular systems." *IEEE J. Sel. Areas Commun.* Special issue on MIMO. June 2003

[8] Harish Kumar, VK Sharma, Pupshaneel Verma, G. Venakat " Analysis of A New base station Receiver Increasing Diversity in a CDMA Cellular System" published International journals of Springer, March 2011, CCIS 142, pp 104-110, 2011

GigNet for Papua New Guinea

N. Gehlot and Simo Kaupa

Department of Electrical & Communications Engineering, The PNG University of Technology, Lae, Morobe, Papua New Guinea
E-mail: 1gehlotn@gmail.com, skaupa@ee.unitech.ac.pg

Abstract
Papua New Guinea (PNG) is the second largest island nation in the world with more than six million people of which bulk of the population (75 percent) are scattered across the rough and rugged terrain of the country. Currently it is experiencing a boom in the economy as a result of Liquefied Natural Gas (LNG) projects and other mining industries. Considering the economy boom demand for bandwidth requirement for converged traffic is up by several magnitudes. Among several competing technologies, Optical Fiber Communications (OFC) system is the most promising media of choice for long-haul, regional, metro and access networks. This paper proposes solutions to deliver GigNet (1000 Mb/s) in PNG by innovative methods to overcome challenges posed by unique socio-economic, cultural, and land ownership by its inhabitants, rather than the PNG government; and the implementation of network is further compounded by PNG's rugged terrains, lack of electric power and related infrastructure.

Keywords: Rugged terrain, GigNet, remote pumping, DWDM and OFC.

Introduction

As expected underdeveloped and upcoming country such as PNG is in dire need of optical fiber communication (OFC) GigNet infrastructure for the economy backbone and improve people's quality of life in all walks of sphere. The aggregate several Gigabits application bandwidth demands can only be fulfilled by OFC compare to wireless and coaxial communication medium. With the optical GigNet infrastructure last mile services such as voice, cable television, radio, e-banking, e-commerce, e-learning, telemedicine and Internet can be made accessible to everyone all the time anywhere in the country. The premier university of South Pacific Region - *The PNG University of Technology (UNITECH), Lae, Morobe, Papua New Guinea attempts to find solutions to the unique hurdles of building OFC based GigNet, first of its kind in PNG. This solution is expected to create a gold rush era in terms of GigNet for both the suppliers and users in South Pacific Region in southern hemisphere.*

One of the most recent examples is of India and China meeting bandwidth demands to sustain and assure growth in all sectors of industries is by using OFC based system. In India, Reliance, Airtel, government owned BSNL, MTNL, Railways, Department of Defence and many other private and public OFC network co-exist and demand is still growing. Countries such as those in Africa and Ireland which also competes with India in business process outsource (BPO) / knowledge process outsource (KPO) for services industries mandate OFC systems; and either continuous deployment or upgrading of the same.

There are several differences between India, Africa, China and similar countries compared to PNG in deploying OFC system. This paper vividly lays out the differences, challenges and proposes ubiquitous solutions in details of Mr. Kaupa's MS EE Thesis at The PNG University of Technology (UNITECH, Lae, and PNG) in Electrical Engineering department in its pioneering efforts of graduate program that in line with the larger PNG development issues of the government's national plans - the vision 2030 & vision 2050.

In most of the world the challenges described below are not of significance but are a major obstacle in PNG. This has been proven in the Exxon-Mobile LNG project worth more than US$50 Billion. Large amount of cash flow has provoked both inter-tribal and intra-tribal acrimonies leading to all out wars, occasionally resulting in villages being swamped and wiped out over night by competing tribes. The land ownership rights are further ignited by people of vested interest in public and private organizations.

Hence, building an OFC network is quite challenging even with the off-the-shelf components are available at throw away price from around the globe, such as Huawei-China, Tell Labs, US and Alcatel-Lucent, US/France, etc. The lack of due diligence required at decision making and trained manpower only makes OFC network deployment difficult if not impossible.

We briefly, summarize the following challenges for the clarifications and are discussed elsewhere in detail.

Availability of Electric Power Supply

Most of the rural areas and district towns do not have electricity supply; and wherever electric power is available it is sporadic due to innumerable blackouts on daily basis as PNG Power Limited (PPL-the energy supplier) generators are hard pressed to keep up with the growing demand.

Landowner issue

PNG land law is unique when compared to most of the countries elsewhere. In PNG, 85% of the total land is owned by several tribes, family or individuals and not by the

government, commonly called customary land[1]. The land is either leased or outright purchased (rarely) from the tribe, family or individual to set up any infrastructure for development. Before the deal of lease or purchase is completed several other tribes will come up for claiming the booty and lead to inter-tribal warfare and pose a major hindrances to speed at which OFC may be deployed. There may be recent occupiers to land rather than traditional tribes on a certain territory and this leads to inter-tribal disputes as well. Compounding factor is the participation by civil servants and political leaders joining the foray to promote self and/or one's own tribe.

When a land lease/sale deal has been done and sealed, the delay in processing payments as prevailing habit kickbacks in most of the commonwealth countries results in gross dissatisfaction among land owners that guarantees delays up to several months which ultimately results in two scenarios:

Loss of revenue in terms of bank interest to investors due to delay in completion of network and cash flow returns.

Land owners may get frustrated; tempers flare as additional deal brokers joins the foray and may result in calling army, police or mercenaries to protect the partially completed network.

Securing OFC network in PNG

There is an additional challenge of security of personal, network, equipment, OFC cables, joints etc., during and after deployment as anti-social elements (RASKALs as commonly referred in PNG) may cut or dig up the network when payment is refused. PPL has lost several grounding copper cables dug out of ground fast buck by anti-social elements resulting in fires and destruction of protective elements all together.

Limited working hours due to security concern is an additional cost factor that must be taken into consideration while building an OFC network in PNG.

There are in excess of 800 languages in land of unexpected – as PNG is often referred to where Australian army has been recently deployed to safeguard Exxon Mobile LNG project. Seeking local help for security can be treacherous even for the US marines as locals are guaranteed to be faithful, committed and sincere only to own tribal brethrens and this is a Molotov grenade in a volatile situation from the point view of security.

Related infrastructure

Poor road conditions do exist connecting a few of major cities only in patches and there is no national surface road network grid to avail at this time. There is absolutely no train track system to move around and piggy back OFC conduits on railway land as often done in India/China etc.

The poor drainage system can result in soaking of the OFC system landing points and nodal grids.

There is some intra-city and inter-city public transport system in a few cities operating only during bright day light. Most of the travel is by expensive air flights or water conduits.

[1]NiuMedia Pacific PNGInLaw (FAX +61 3 6224 0814) [March 2011 update, Chapter 10, Page 77, 2004.

Additional Concerns

Cultural and social problems are also contributing factors to development progress or OFC network deployment. In PNG cultures, values, languages and dispute settlement methodologies varies every 100 km. Law of tribes override civil laws. Often, tribal laws and civil laws must be fulfilled to settle issues related to both living and non-living commodities that includes land for OFC network.

Keeping in mind all of the above the challenges, a creative, out of norms and novel solutions are required to develop OFC based GigNet for PNG. A careful planning and design is required at the physical layer of the OFC network even though the plug-and-play, off-the-shelf OFC system components for bandwidths of several Terabits Network are available.

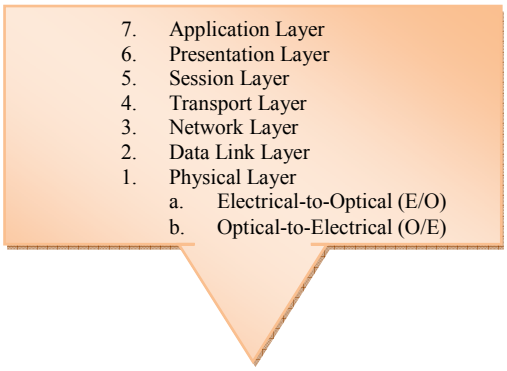


Figure 1: PNG OFC backbone and GigNet design challenge is concentrated at Layer-1 (Physical Layer) as high layers are off-the-shelf with plug and play.

Emulated Scenarios

To comprehend the rugged terrain of the country four scenarios are generated for broader classification and solutions in the emulated model for the long-haul and regional optical fiber networks.

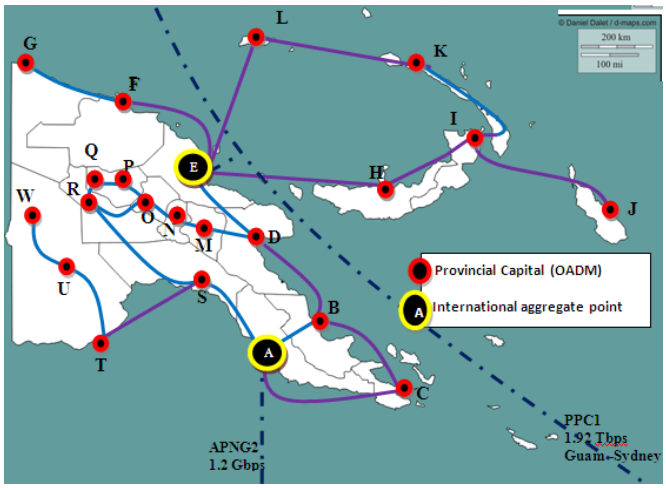


Figure 2: PNG map showing prospect long-haul and regional optical network for GigNet. Repeater span for 32 channels with 10 Gb/s per channel is assumed to be 80 km.

Table 1: Rugged terrain challenges along with availability of power, highway and utility network for the deployment of OFC network for GigNet in PNG summarized.

Case	Availability YES NO	Stations PAIR	Inter-Station Distance (km)
I	Electric power - NO Utility network - NO Inter Station highway - NO	A-B A-C B-C B-D S-T E-F E-H H-I I-J K-L L-E T-U	160 440 350 340 360 320 550 320 523 397 393 488
II	Electric power - NO Utility network - YES Inter Station highway - NO	A-S-R Q-R	750 80
III	Electric power - NO Utility network - NO Inter Station highway - YES	F-G P-Q U-W	340 71 178
IV	Electric power - YES Utility network - YES Inter Station highway - YES	D-E D-M-N-O-P-R	59-260

Laying OFC Fiber Cable

Scenario 1: Case 1 (Mainland)

A consortium of public and private entities in PNG may help to reduce deployment cost and result in a win-win partnership to build towers and run aerial optical ground wire (OPGW) for connectivity. But it may take a long time to hammer out formalities, capital investment among PNG government entities and unclear policies given the unstable government could result in substantial delays to deploy GigNet within 12 months.

For example, recent adventure of PPL in to deploying OFC cable resulted in contentious issues because telecommunications being a global cash cow and everyone wants to milk it. Thus a consortium is a decent proposal but hard to progress due to clarity in PNG national policies.

Free space optical communication (FSOC) is an alternate solution but not quite possible for GigNet since repeater span and bit rate are limited by factors such as beam dispersion, atmospheric absorption, rain, fog, line of sight obstruction, background light. FSOC system could be quite expensive for rough and mountainous terrain covered with dense tropical forest and vegetation. PNG is pristine and consists of lush green forest responsible for giving the world hundreds of new species flora, fauna and creatures but FSOC is not practical and cost effective.

One promising and innovative solution is to use the **existing river network** in PNG to deploy end-to-end fiber connectivity for the GigNet delivery. The most remote parts of PNG can be reached via the existing river network which flow into the main ocean (hub) from the higher lands.

River networks offers attractive alternative to conquer the rugged mountainous geography of PNG with the least cost and fast deployment. The existing and well matured technology for shallow water deploying of submarine cable can be readily availed. This is analogous to military global positioning system (GPS) success in global commercial market with wide range of applications. Hence, the expertise of submarine system should be adopted for using river beds for OFC network deployment. Deep river gorges challenge are surmountable as they have been well understood in submarine cables laying at depths of greater than few kilometers both in Atlantic and Pacific oceans.

Rivers are not owned by individuals or tribes anywhere in the world and that includes PNG. Therefore this will also help overcome land ownership issues. To reduce the overhead cost the regulated river and stream act be made known to the people and use local help to infuse cash in to communities. PNG water act stipulates that the right to the use, flow and control of water is vested in the State[2].

Scenario 2: Case I (Coastlines and between islands)

Run optical submarine cables customized for both shallow and deep water. Hybrid erbium doped fiber (EDF) and Raman amplification with bidirectional remotely pumped lasers could be used for long optical distance without requirement of electric power. Power for the underwater optical amplifiers could be provided by the DC current as required delivered from power feed equipment at the terminal stations [1].

Scenario 3: Case II (Mainland)

Share the existing utility network (LNG gas pipe line), infrastructure to deploy OFC network. Authors recommend the use of optical amplifiers that are remotely co-pumped or counter-pumped at 980 nm or 1480 nm due non-availability electric power at several points.

Scenario 4: Case III (Mainland)

It is prudent to use the existing rivers and brooks networks because townships are located in close proximity to water ways. In some cases, existing PPL towers and poles for aerial suspension of fiber may be used.

Scenario 5: IV (Mainland)

Use the existing high voltage transmission towers to run OPGW.

Fiber cable and the amplifiers

Polarization mode dispersion (PMD) and non-linear effects in the dense wavelength division multiplexing (DWDM) system can be minimized by deploying non-zero dispersion shifted fiber (NZDSF) which are optimized to operate in the C band

[2]PNG Constitutional Laws of Water, Part VII, Section 79, Subsection 1.

and emerging L band. G.655 fibers such as Corning LEAF, Lucent TrueWave and Alcatel TeraLight can be deployed in the metro, regional and long-haul networks which enable system to evolve from today's 10 Gb/s to 40 Gb/s and beyond.

Inter-station separation distance calls for optical amplification with remote pumping (where there is no power available) and minimum number of traditional 3R receivers where necessary to minimize costs.

Remote Pumping

For Case I, II, III where power is not available for distances up to 600 km a usage of remote pumping technique for either EDFA or Raman amplifier is recommended to flatten the optical gain bandwidth along the passive network while operating within the performance index [2]. The hybrid erbium doped fiber (EDF) and Raman distributed amplifications are deployed to obtain necessary optical signal to noise ratio (OSNR) and increase repeater span. For the emulated scenarios where electric power is not available at the transmitting end the counter-optical pump is recommended at the receiving end as illustrated in Figure 4. Deploying of EDFA requires two separate channels to carry signal and the secondary is used for control and monitoring (traditionally referred to as line monitoring system).

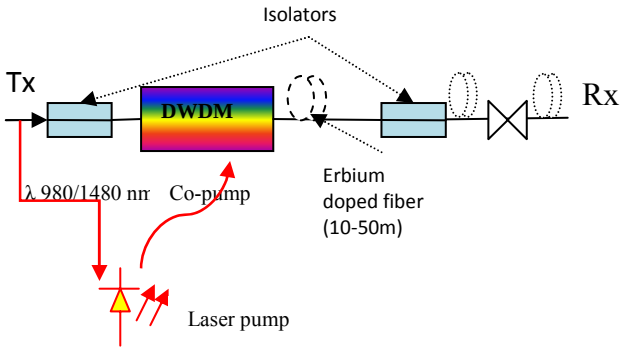


Figure 3: Co-pump configuration for remote amplification

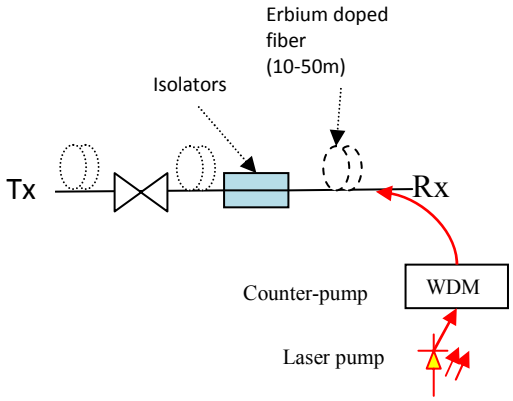


Figure 4: Configuration showing counter-pumping for remote amplification

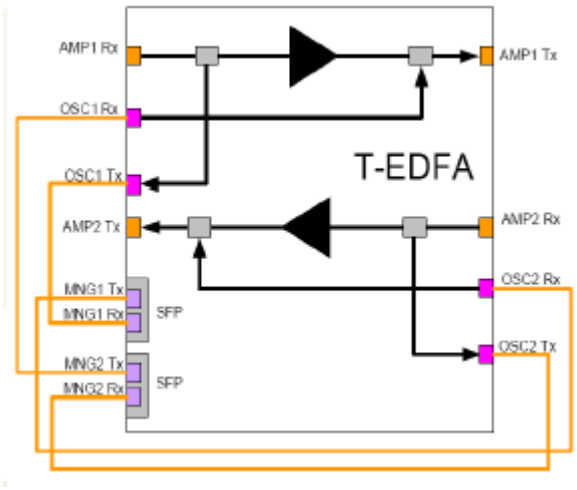


Figure 5. Off-the-shelf Optical Amplifier solutions from PacketLight manufacture for networks that can be deployed as Booster, PreAmp or In-line amplifier.

Optical Fiber Amplifier (OFA) are cost effective, housed in a single unit or module consisting of a laser pump, wavelength division multiplexer (WDM), gain fiber erbium doped) and isolators. Off-the-shelf OFA solutions from PacketLight Networks that support 4/8/16/32/40 wavelengths and amplification function of Inline, Power-booster or PreAmp with Optical Supervisory Channel for Remote Management illustrated by Figure 5 are best suited for PNG GigNet deployment [3].

Land and submarine crossing distance up to 600 km urge the use of laser pump (980 or 1480 nm) whose wavelength is pumped into the erbium doped fiber ultimately to achieve population inversion in the gain fiber medium and thus obtain optical amplification via stimulated emission for the optical spectrum. For emulated cases I, II and III with very large crossing distance use hybrid EDF and Raman distributed fiber amplification technique which has the advantage of flattening the optical gain of the bandwidth [4].

OFC System design parameter limit

DWDM technology increases carrying capacity of a single pair with each input signal independent of the others [5]. For the GigNet a single fiber pair will simultaneously operate at 32 wavelengths. Every station is emulated for 32 channels reconfigurable optical add/drop multiplexers (ROADM) each operating at 10 Gb/s with a channel spacing of 100 GHz[3].

[3]ITU Grid G.692

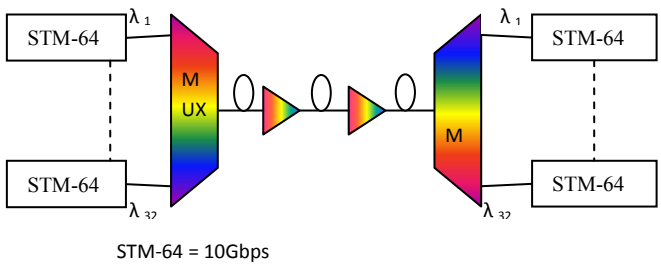


Figure 6: DWDM and SDH interface for 10 Gb/s Channels.

Synchronous Digital Hierarchy (SDH) is the family of optical fiber transmission with a line rate range from 155.520 Mb/s to 39.813 Gb/s that is used outside the US and Japan. Every channel is simulated for synchronous transport module (STM-64) indicated by Figure 6. All these equipment are off the self products.

Table 2: WDM transmission using the maximum distance and 32 channels per station.

Case	Channel N	Bit Rate B (Gb/s)	Capacity NB	Distance L (km)	NBL Product [(Tb/s)-km]
I (E-H)	32	10	320	550	176
II (A-S-R)	32	10	320	750	240
III (F-G)	32	10	320	340	108.8
IV (D-E)	32	10	320	260	83.2

Capacity = Channel x Bit Rate (10 Gb/s)
Performance index = Bit Rate x Distance
NBL product = Capacity x Distance (km)

Two major cities (A and E) as illustrated by Figure 2 are chosen for the international aggregation where the submarine optical branching units (BUs) for PIPE Pacific Cable (PPC1) and direct Australia to PNG link (APNG-2) are terminated.

Last Mile CONNECTIVITY

For Access network within town centres lease or rent existing utility networks such as sewage system, drinking water pipes, electricity poles, cable TV poles and existing telecommunications ducts to run G.652 and G.653 fiber to the last mile customers. To meet the exponential bandwidth demand fiber-to-the house (FTTH) is the way forward. It can be configured as point-to-point or point-to-multipoint using xPON technologies. Other retail telecommunications providers choosing to use wireless or copper cables to redistribute services to the last mile customers can use GigNet as their backbone for layer 2 and 3 traffic as illustrated by Figure 1.

Conclusion

River network provides optimum solution to all of the above challenges of landownership (as river beds are owned by the government); rugged and difficult terrain as explained in *Scenario 1*; security (it will be difficult for the rascals to swim in rivers and cut the OFC cable and seek ransom); and the cost and speed of the deployment is enhanced too.

From the above discussions, it is clear that all the challenges posed in deploying GigNet in PNG are realistically overcome with cost reduction [6] and faster deployment. Table 3 summarizes the proposed solutions for difficult terrain, socio-economic conditions along with peculiar land ownership issues of PNG.

Table 3: Summary of the Challenges and innovative solutions.

Challenges	Workable Solutions
Need to services such as voice, television, radio, telemedicine, e-commerce, e-banking, e-learning, internet and internet applications such as VoIP, etc.	GigNet
Deploying of fiber in the rugged terrain where no power, no utility network exist	Use river network
Deploying of fiber between islands and along the coastline where there is no road or utility network	Use the submarine cables
To send signal over distance up to 600 km with no electric power available	Use remote pumping techniques (co, counter , bidirectional or hybrid)

The above innovative solutions are customized for the PNG GigNet but may use elsewhere under similar demanding conditions.

Acknowledgement

The Authors would like to acknowledge the Vice Chancellor, Prof. (Dr.) Misty Baloiloi and the Registrar, Mr. Allan Sako of UNITECH for the vision, need and championing the MS-PhD program in Electrical Engineering in the best interest of PNG as a nation. The Authors are grateful to The Chancellor, Mr. Philip Stagg and the council members of UNITECH for encouraging research and industrial alliance. The authors sincerely thank Mr. Tony Koiri, CEO, PNG Power and Mr. John Yanis, General Manager, PNG Power for fostering and nurturing research in PNG.

References

[1] Southern Cross Cable Network
[2] Govind P. Agrawal, Fiber-Optic Communication Systems, 2nd Edition.
[3] PacketLight Networks
[4] Electro-Optics Handbook, Chapter 28: Optical Amplifiers by Beth A. Koelbl
[5] Fiber Optics Infro.com, Optical Amplifiers, http://www.fiber-optics.info/
[6] e-optolink, 181 DWDM passive systems Catalog

Bibliography

Dr. Narayan Gehlot, Senior IEEE Member.
Dr. Gehlot is a renowned leader in communications with over 15 years (post doctorate) of research and development experience in systems, board and chip design. He has worked in some of world's leading laboratories in telecommunications and computer networking such as Indian Institute of Technology, Madras, India; Bellcore, Morristown, NJ, US; AT&T Bell Laboratories, Holmdel, NJ, US and Lucent Technologies Bell Labs Holmdel, NJ, US. A genuine innovator who has contributed to more than 51 (33 issued) patents globally in a wide range of technologies such as wired and wireless communications, fiber optics for FTTx, long haul, metro, intercontinental submarine systems, network management system, line monitoring systems, computer, internet, security, database, networks, vehicular technologies, Raman amplifier etc. Dr. Gehlot's patents have been cited in more than 333 issued patents.

Dr. Gehlot strongly believes that *ideas*[4] *are the key assets* to success whereas solution to a problem is only a matter of time. He is an outstanding researcher, innovator known for successful collaborations with Universities and intellectual property creation. He is bestowed with the unique ability of foresightedness to look ahead and plan. At Bell Labs Dr. Gehlot was honored as an "Outstanding Asian American for Lucent's Success" along with world-renowned researchers and 1998 Nobel Prize laureate Dr. Daniel Tsui. Dr. Gehlot is an alumnus of BITS, Pilani, India; NJIT, Newark, NJ, US and University of Pittsburgh, Pittsburgh, PA, US. He is a Senior Member of IEEE and has keen interest on fundamentals of nature beyond science. Dr. Gehlot is currently a faculty in EE department at UNITECH, Lae, PNG.

[4]Albert Einstein "The formulation of a problem is often more essential than its solutions, which may be merely a matter of mathematical expressions or experimental verifications."

Mr. Simo Kaupa, Student Member IEEE.
Mr. Kaupa is a faculty at UNITECH since 2005 in electrical and communications department. His area of research is communications and computer networking discipline with emphasis on fiber optic network specifically to PNG to improve connectivity through research. He is UNITECH alumnus with a Bachelor in Electrical Engineering. He is currently a Graduate (MS EE) Student in Electrical Engineering department at UNITECH, Lae, PNG and pioneering student member of IEEE at UNITECH PNG.

Dynamic Certification Authority in Mobile Adhoc Network

[*1]Vijender Hooda, [2]Yashvardhan Soni and [3]Amninder Kaur

[1]Associate Prof., Department of Information & Technology
Dronacharya College of Engineering, Gurgaon, India
E-mail:viju_hooda@rediffmail.com
[2]Asstt. Prof., Department of Information & Technology
Dronacharya College of Engineering, Gurgaon, India
E-mail:soni.yashvardhan@gmail.com
[3]Asstt. Prof., Department of Electronics & Communication Engineering
Dronacharya College of Engineering, Gurgaon, India
E-mail:amninder.kaur@rediffmail.com

Abstract
Mobile Ad Hoc Network (MANET) is an infrastructureless network of wireless mobile nodes that cooperate among each other to maintain connectivity of the network. In comparison to wired networks, securing a MANET is more difficult and challenging Ad hoc networks have lots of applications; however, a vital problem concerning their security aspects must be solved in order to realize these applications. The dynamic and cooperative nature of ad hoc networks present challenges in securing these networks. In this paper we propose the design of a dynamic Certification Authority which follows dynamic approach to solve the above problem using a suite of network monitoring protocols.

Keywords: Mobile Ad hoc networks (MANET), authentication, certification Authority (CA), reputation, threshold, security.

Introduction
Mobile ad hoc networks are complex wireless networks, which have little or no existing network infrastructure. These networks can be established in a spontaneous manner allowing organizations and network members to work together and communicate, without a fixed communication structure. The term ad hoc is defined as: “Meaning "to this" in Latin, it refers to dealing with special situations as they occur rather than functions that are repeated on a regular basis.” (The American Heritage Dictionary of the English Language, Fourth Edition. Houghton Mifflin Company, 2004).[1] Due to recent wireless technology advances, mobile devices are equipped with sufficient resources to realize implementation of these dynamic communication networks. However, for ad hoc networks to find a wide spread within both the military and commercial world, they must be secured against malicious attackers. Mobile ad hoc networks have distinct characteristics, which make them very difficult to secure. Such characteristics include: the lack of network infrastructure; no pre-existing relationships; Unreliable multi-hop communication channels; resource limitation; and node mobility. Users cannot rely on an outside central authority, like a trusted third party (TTP) or certificate authority (CA), to perform security and network tasks. The responsibility of networking and security is distributed among the network participants. Users have no prior relationship with each other and do not share a common encryption key. Therefore, only after the network has been formed, the users establish trust and networking links. Figure 1 [Zhou & Hass, 1999] demonstrates these autonomous, multi-hop characteristics. Connection between nodes is made by means of other nodes within the network.

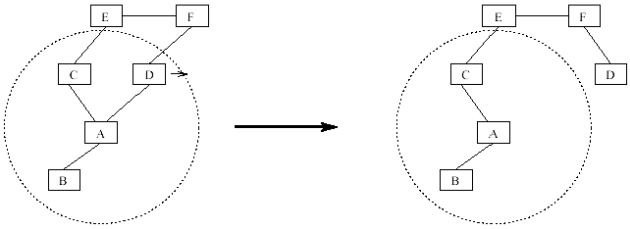


Figure 1: Ad Hoc Network Topology.

In Figure 1, the circle represents wireless range of node A. In Figure 1, when node D appears within the range of node A, the topology changes to maintain the connection. Note that all network functions are performed by the nodes and no host or outside authority exists. An ad hoc network is a dynamic type of network which is both similar and very different to its parent fixed communication network [2].

Dynamic Network Architecture
Ad hoc networks have no fixed or existing network infrastructure. The network architecture is continuously changing as the network evolves. There is no pre-existing or fixed architecture which handles all network tasks such as: routing security and network management. Instead, the network infrastructure is spontaneously set up in a distributive manner. Each participating node shares the network’s responsibilities. Distribution of network functionality avoids

single point attacks and allows for the network to survive under harsh network circumstances. A fixed entity structure, such as a base station or central administration, is crucial for security mechanisms. A trusted third party member [William, 1999], which is expected in traditional networks, is similar to a fixed entity as both define security services; manage and distribute secret keying information (which allows secure communication of data through encryption and decryption techniques). Therefore the absence of such a control entity introduces new opportunities for security attacks on the network.

Security objectives and services

Securing mobile ad hoc networks requires certain services to be met. A security service is a made available by a protocol which ensures sufficient security for the system or the data transferred. The security objectives for mobile ad hoc networks are similar to that of fixed wired networks [3]. The security objects are described in six categories, adapted from discussions in [Stalling, 2003]:

- Authentication
- Access Control
- Data Confidentiality
- Data Integrity
- Non-repudiation
- Availability Services

Security model

A security model for mobile ad hoc networks is illustrated, in general terms, in Figure 2. A message *M* is to be transmitted from the source *A*, across a network of nodes, to a destination node *B*. The two entities who are primary participants must collaborate for the transaction to occur. The multi-hop route will involve secondary participating nodes. Security is provided by two accompanying techniques: a security related transformation applied to the message (resulting in an encrypted message *C)* and secret keying information shared by the principal participants. The general mobile ad hoc security model shows four basic tasks for a security mechanism:

- The design of a security algorithm.
- Generation of secret keying material used in conjunction with this security algorithm.
- Distribution of secret keying material.
- Protocol for the participants to follow which will achieve the required security services.

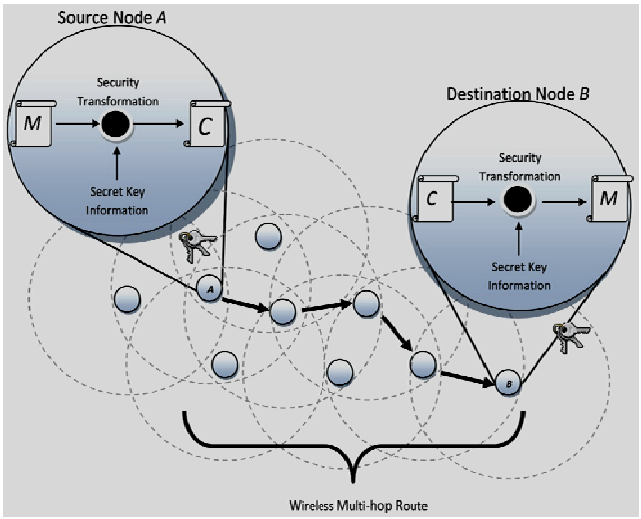


Figure 2: General Security Model.

Certification Authorities

This section focuses on CAs. In ad hoc networks, trust is managed locally at the individual nodes. A node is not trusted by a given node until it presents a certificate, and the node in question verifies that the certificate was issued by a trusted CA, and it has not expired nor been revoked [4]. The CAs have the following trust management tasks:

- Issuing of certificates
- Storage of certificates
- Certificate validation
- Revocation of certificates.

Beyond managing certificates, it is also the CA's responsibility to disseminate the public keys of principals to inquiring clients. Every response from the CA is signed with the CA's private key, and so can be validated with the CA's public key. The success of this approach lies in maintaining the secrecy of the private key of the CA [5]. It is also necessary for the CA to remain on-line (i.e. available) to provide these services. There are three major parameters to a distributed key management framework: fault tolerance, vulnerability and availability. The first parameter is associated with the number of node failures the system can handle; the second is associated with the number of compromised nodes the system can withstand, whereas the third is associated with the ability of the client to contact the required number of CAs. The optimization of any one of these parameters may adversely affect other parameters and so adversely affect the success of the system. In addition, mobile networks present hostile environments where nodes may easily die or be compromised and no guarantees can be made about the ability to access the necessary nodes for authentication. An ideal key management service for ad hoc networks should provide the best of both worlds: it must be light-weight and simple to mobile nodes, and it must be available in highly dynamic networks.

Certification Authorities Selection

A single centralized authentication server is unsuitable for ad hoc networks, from the security point of view, as it may be subject to a single point attack. To provide better fault tolerance, it is possible to deploy many copies of the CA in the network. With many such replicas, the system can withstand a number of replicated CAs - 1 failure because the CA service is available as long as there is at least one operational CA. Availability has also been improved since a client node will have a better chance of reaching one of the multiple CAs to get service. Unfortunately, the system has become more vulnerable. An adversary need only compromise one of the many CA nodes to acquire the secret key and so compromise the whole system. The problem of using replicated CAs stems from the fact that each replica has full knowledge of the system secret. The approach is vulnerable against any attacks that compromise a single replica, which should not be considered too difficult considering the inherent physical vulnerability of mobile nodes. The Threshold Digital Signature scheme was proposed to address this problem. With threshold digital signatures, again the key is divided into n pieces and distributed. But now if a client needs a signature on its data, each secret holder will use its piece of the key to generate a partial signature over the data. When client collects k of these partial signatures, the client can reconstruct the full signature [6].

Certification Schemes in Adhoc Networks

Different certification schemes have been presented in the literature. We classify a these schemes into cluster-based schemes and non cluster-based schemes and present them in subsections 7.1 and 7.2 respectively.

Cluster-Based Certification Schemes

In A cluster-based architecture for a distributed public key infrastructure that is highly adapted to the characteristics of ad hoc networks was introduced in. In order to adapt to the highly dynamic topology and varying link qualities in ad hoc networks, central instances that would form single points of attack and failure were avoided. Instead, the ad hoc network was divided into clusters, and the cluster heads jointly perform the tasks of a certification authority. A proactive secret sharing scheme distributes the private network key to the cluster heads in the ad hoc network. Instead of a registration authority, arbitrary nodes with respective warranty certificates may warrant for a new node's identity. Based upon this authentication infrastructure, a multi level security model ensuring authentication, integrity, and confidentiality is provided. Authentication itself is realized in two stages. First, a node gets the status of a guest node. After sufficient authentication, the node will become a full member. An additional important feature is the possibility to delegate the cluster head functionality to another node.

Non Cluster-Based Certification Schemes

A Certification protocol called MP (MOCA Certification Protocol) was proposed. Given the threshold value, *k*, the total number of nodes, *M*, and the number of MOCAs, *n*, the communication pattern between a client and *k* or more MOCA servers is one to (k or more) then back, which means that a client needs to contact at least k MOCAs and receive replies from each of them. To provide an efficient way of achieving this goal, a certification protocol called MP (MOCA certification Protocol) was proposed. In MP, a client that requires certification services sends Certification Request (CREQ) packets. Any MOCA that receives a CREQ responds with a Certification Reply (CREP) packet containing its partial signature. The client waits a fixed period of time for *k* such CREPs. When the client collects *k* valid CREPs, the client can reconstruct the full signature and the certification request succeeds. If too few CREPs are received, the client's CREQ timer expires and the certification request fails. The client is left with the option to initiate another round of certification requests [7].

Certificate Revocation Schemes

In this section we focus on the certificate revocation in ad hoc networks. The importance of the certificate issue is discussed and the certificate revocation schemes that have been used for ad hoc networks are presented. Certificates can be revoked if nodes are found to be corrupt or compromised. This revocation service assumes that all nodes monitor their one-hop neighbour nodes and are capable of retaining their own certificate revocation list (CRL) [Luo & Lu, 2000]. When a user node identifies a neighbouring node is corrupt, it adds the node in question to its CRL and announces this to all neighbouring nodes. The neighbouring nodes in turn check if this announcement is from a reliable source, i.e. the source is not on the receivers CRL. If the source is reliable, the announced node is marked as suspect. If a threshold of *k's* reliable accusation is made against a single node then the node's certificate is revoked. This procedure allows for compromised nodes to be identified and explicitly quarantined from CA involvement, until such a time as they have become secure again. Implicit revocation is implemented by setting lifetimes for certificates t_{cert}. When the time has expired and the certificate has not been renewed it is implicitly revoked [8].

Threshold Cryptography

Threshold cryptography is used to share the CA service between nodes. A threshold cryptography scheme allows the sharing of cryptographic functionality. A (*t-out-of-n)* threshold scheme allows *n* nodes to share the cryptographic capability. However, it requires *t* nodes, from the *n* node set, to successfully perform the CA's functionality jointly. Potential attackers need to corrupt *t* authority nodes, before being able to exploit the CA's functionality and analyze secret keying information. Therefore, a (*t-out-of-n)* threshold scheme tolerates *t-1* compromised nodes, from the *n* node set [Aram et al, 2003]. When applying threshold cryptography to the shared CA problem, the CA service is shared by *n* nodes across the network called authority nodes. The private key *k*, crucial for digital signatures, is split into *n* parts (*k1, k2, k3, …, kn*) assigning each part to an authority node (*an*). Each authority node has its own public key, *Kn*, and private key, *kn*, (as seen in Figure 3).[9]

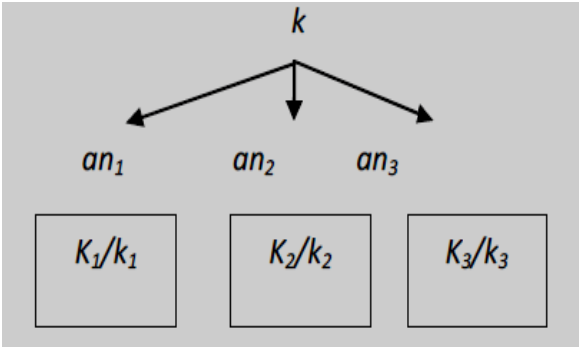


Figure 3 (2-out-of-3): Threshold Key Management.

It stores the public keys of all the network nodes (including other authority nodes). Nodes wanting to set-up secure communication with node *i* need only request the public key of node *i* (K_i) from the closest authority node - therefore increasing the CA's availability. For the CA service to sign and verify a certificate, each authority node produces a partial digital signature using its respective private key, k_p, and then submit the partial digital signature to a combining node. Any node may act as a combiner in the ad hoc network. The partial digital signatures are combined at a combiner (*c*) to create the signature for the certificate, *t* correct partial digital signatures are required to create a successful signature. Therefore, protecting the network against corrupt authority nodes, up to *t-1* corrupt authority nodes may be tolerated [Lidong & Zygmunt, 1999]. For example, Figure 4 shows a (*2-out-of-3*) threshold scheme where the message *m* is signed by the CA, two partial signatures (*PS*) are accepted, while the third (*an2*) was corrupted. The partial signatures meet the threshold requirements and the partial signatures are combined at *c* and applied to the message [10].

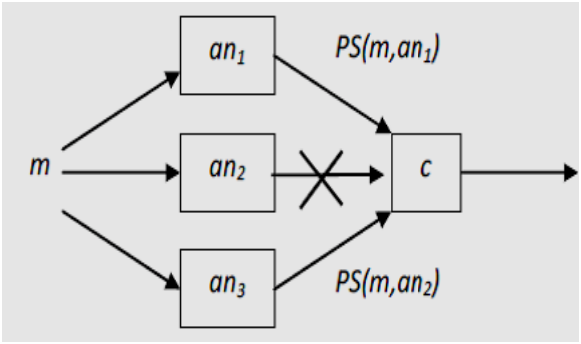


Figure 4 (2-out-of-3): Threshold Signature.

Conclusion

In this paper we discussed the design of a dynamic CA for MANETs based on threshold cryptography. We found that the delay experienced by nodes for certificate renewal increases when the number of nodes in the network is reduced. We proposed a set of monitoring protocols which extends the design of a distributed CA by providing dynamic behavior. The protocols enable the distributed CA to dynamically update the threshold value by monitoring the Average Node Degree of the network and thereby prevent an increase in the certificate renewal delay.

References

[1] C. Siva Ram Murthy, B.S. Manoj, "Ad Hoc Wireless Networks: Architectures and Protocols", Prentice Hall PTR, New Jersey (May 2004).

[2] L. Zhou and Z. J. Haas. Securing ad hoc networks. IEEE Network Magazine, 13(6), (1999)

[3] J. Kong, P. Zerfos, H. Luo, S. Lu, and L. Zhang. Providing Robust and Ubiquitous Security Support for Mobile Ad-Hoc Networks. In Proceedings of ICNP '01.

[4] R. Gennaro, S. Jarecki, H. Krawczyk, and T. Rabin, "Robust threshold DSS signatures, ". In U. Maurer, editor, Advances in Cryptology – Proceedings of Eurocrypt '96, number 1070 in Lecture Notes in Computer Science, pages 354–371, Zaragoza, Spain, May (1996).

[5] Y. Desmedt, "Threshold cryptography, ". European Transactions on Telecommunications, 5(4):449–457, July (1994).

[6] Broch and D. B. Johnson, "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks", IETF Internet Draft, October (1999).

[7] S. Vassilaras, D. Vogiatzis and G. Yovanof, "Security and Cooperation in Clustered Mobile Ad Hoc Networks With Centralized Supervision, " IEEE Journal on Selected Areas in Communications, vol. 24, no. 2, February (2006).

[8] T. Clausen et. Al., " Generalized MANET Packet/Message Format", draft-ietf-manet-packetbb-June (2008).

[9] Chakeres et. Al., "Dynamic MANET On-demand (DYMO)Routing", draft- ietf- manet- dymo June (2008).

[10] M. Gasser et al., "The Digital Distributed Systems Security Architecture", Proc. 12th Natl. Comp. Security Conf., NIST, (1989).

Performance Analysis of DSR, AODV Routing Protocols based on Wormhole Attack in Mobile Ad-hoc Network

Gunjesh Kant Singh, Amrit Kaur and A.L. Sangal Email:

E-mail: Gunjesh31@gmail.com, amrit.tiet@gmail.com, sangal62@yahoo.com

Abstract
Mobile ad-hoc network are able work without any existing infrastructure. MANET is a self configure network connected by wireless links. Mobile ad-hoc network uses temporary network which is able to work without any centralize administration or stand alone infrastructure. In mobile ad-hoc network each device move in any direction without any restriction so it changes it links to often with other devices present in same network. Mobility of mobile device anywhere in the network without any centralize administration makes it difficult to manage routing. In mobile ad-hoc network each device need to forward traffic that is not related to its own use and therefore each device work as a router. MANET's protocol has different security flaws and using these flaws many kind of attack possible on mobile ad-hoc-network. Wormhole is one of these attacks. Wormhole attack causes serious affect on performance of the MANET protocol and preventing the attack has proven to be very difficult. In wormhole attack attacker place some malicious node in the network. A malicious node captures data packets from one location in the network and tunnels them to another malicious node at distinct location, which replays them locally. These tunnels works like shorter link in the network and so act as benefit to unsuspecting network nodes which by default seek shorter routes. This paper illustrates how wormhole attack affects performance of routing protocol in mobile ad-hoc network using random waypoint mobility model with varying node mobility.

Index Terms: AODV, CBR, DSR, MANET

Introduction

Mobile networks can be classified into infrastructure networks and mobile ad hoc networks (MANET) according to their dependence on fixed infrastructures [2]. In infrastructure based mobile network wired access point is used and within the transmission range of access point all mobile device are free to move in any direction. In mobile ad-doc network each device is free to move any direction so the routes use to reach from one device to another change frequently. In mobile ad-hoc networks each device need to forward traffic that is not related to its own. Routing paths in MANETs potentially contain multiple hops, and every node in MANET has the responsibility to act as a router [4]. There are various mobility models such as random way point, reference point group

Mobility model (RPGM), Manhattan mobility model, freeway mobility model, Gauss Markov mobility model etc that have been proposed for evaluation [6, 13]. Several parameters such as mode mobility, traffic load and node density and pause time has been used to evaluate performance of MANET routing protocols.. Biradar, S. R. et al.[11] have analyzed the AODV and DSR protocol using Group Mobility Model and CBR traffic sources. Biradar, S. R. et. al.[11] investigated that DSR performs better in high mobility and average delay is better in case of AODV for increased number of groups. Also Rathy, R.K. et. al [8] investigated AODV and DSR routing protocols under Random Way Point Mobility Model with TCP and CBR traffic sources. They concluded that AODV outperforms DSR in high load and/or high mobility situations.

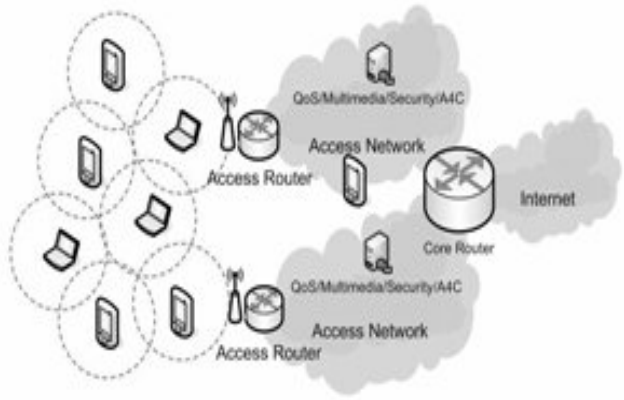


Figure 1: Mobile Ad-hoc Network.

Random Waypoint Mobility Model

MANET's protocol performance frequently observes and studied by simulation and their performance depends heavily on the mobility model that governs the movement of the nodes [5]. Random way point is a mobility model that use random based mobility to manage mobility of mobile devices in a wireless communication system. This mobile model describes various property of mobility like movement patter of the mobile users and their location velocity and acceleration change over time. Mainly this type of mobility model is use for simulation when network protocol performance is evaluated. The Random waypoint model, first proposed by Johnson and Maltz[17], soon became a "benchmark" mobility model[20] to evaluate the Mobile ad hoc network (MANET) routing protocols, because of its simplicity and wide availability.

Description of Routing Protocol

In this section we will provide review on couple of typical ad hoc network protocol namely DSR and AODV.

Ad-Hoc on Demand Distance Vector (AODV)

The Ad-hoc On-demand Distance Vector routing protocol [1, 3, 12] enables multi hop routing between the participating mobile nodes wishing to establish and maintain an ad-hoc network. AODV is a reactive protocol based upon the distance vector algorithm. ADOV uses many type of message in order to find route from one mobile device to another mobile device. Route discovery process starts when a source node needs to send a packet to destination node but it does not have a valid route to destination node. AODV initiate a path discovery process to locate the other node. Source node broadcast route request (RREQ) packet to all it neighbors. Then their entire neighbors forward this request to their neighbors and so on. This process is continuing until either the destination node is found or an intermediate node with "fresh enough" route to destination is located. Sequence number is use by AODV to ensure all route are loop-free and contain most recent route information. In AODV to avoid looping each node maintains it own sequence number as well as a broadcast ID. The broadcast ID is incremented for every RREQ the node initiates, and together with the node's IP address, uniquely identifies an RREQ. Along with its own sequence number and the broadcast ID, the source node includes in the RREQ the most recent sequence number it has for the destination. Intermediate nodes can reply to the RREQ only if they have a route to the destination whose corresponding destination sequence number is greater than or equal to that contained in the RREQ. AODV uses periodic local broadcast hello message. Hello message help a node to inform its neighbor that it active and working. However, the use of hello messages is not required for Nodes listen for retransmissions of data packets to ensure that the next hop is still within reach. If such a retransmission is not heard, the node may use any one of a number of techniques, including the reception of hello messages. Hello messages may list the other nodes from which a mobile has heard, thereby yielding a greater knowledge of network connectivity.

Dynamic Source Routing (DSR)

This is an on-demand routing protocol based on source routing concept. In DSR mobile nodes stores source routes in it caches for which mobile device are aware. When new routes are learned by nodes entries of cache is updated for these new routes. Working of this protocol can be divided in two parts. (a) Route discovery (b) Route maintenance. When a mobile node need to send any packet it first consults with its route cache that whether it already have a route for destination. It an unexpired route is present it send the packet using this route. But if node does not have such route it initiates broadcasting of route request packet. This route request message contains the address of the destination, along with the source node's address and a unique identification number. Each node that receive that packet check it cache to know whether a route for this destination exists or not. If route does not exists it adds it own information to the packet and send it to outgoing link. To limit the number of route requests propagated on the outgoing links of a node, a mobile only forwards the route request if the request has not yet been seen by the mobile and if the mobile's address has not already appeared in the route record. A reply packet is generated when request packet either reach to destination node or it reach to a intermediate node who have unexpired route for destination in its cache. By the time the packet reaches either the destination or such an intermediate node, it contains a route record yielding the sequence of hops taken.

The paper is organized as follows. In the section II, we give brief description of Random waypoint Mobility Model. In section III, we give brief introduction of AODV and DSR routing protocol. Section IV, describes the wormhole attack. In section V, cover the simulation setup and result of simulation and at the end in section VI, we draw the conclusion of simulation scenarios.

Wormhole attack

In this attack, an adversary receives packets at one point in the network, tunnels them to another point in the network, and then replays them into the network from that point [20]. Malicious nodes are connected via a link called "wormhole link" using private high speed network. Wormhole attack is simple to deploy but it may cause significant damage to network. Wormhole attack can be carry out by using different techniques. Here we discuss two methods to generate wormhole attacks in mobile ad-hoc network. In the first type of wormhole, all packets which are received by a malicious node are duly modified, encapsulated in a higher layer protocol and dispatched to the colluding node using the services of the network nodes. These modified packets reach to colluding node just like normal node traverse form one node to another node. Once packets reach to intended malicious node, its extract the packet make the requisite modifications and send them to intended destination. In second type of attack after packets are modified and encapsulated they are sending using a point to point specialized link between the malicious nodes.

Simulation Setup and Result

We have used Network Simulator Qualnet 5.0.2 in our evaluation. In Scenario we have place 50 nodes uniformly distributed in area of 500m x 500m. For this study, we have used random waypoint mobility model for the node movement with 0 sec pause time and 5, 10, 15, 20, 25, 30, 35, 40 meter/sec node mobility speed. The parameters used for carrying out simulation are summarized in the table 1.

Performance Metrics

We have used the following metrics for evaluating the Performance of two on-demand Reactive routing protocols (AODV & DSR):

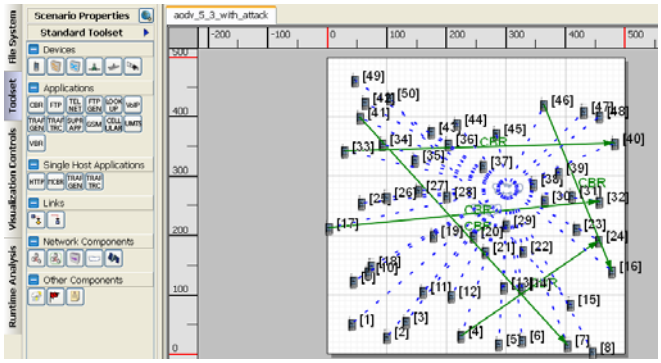


Figure 2: Simulation scenario in qualnet simulator.

Packet delivery ratio

It is the ratio of data packets delivered to the destination to those generated by the sources. It is calculated by dividing the number of packet received by destination through the number packet originated from source.

PDF = (Pr/Ps)*100

Where Pr is total Packet received & Ps is the total Packet sent.

Table I: Simulation Parameters.

Parameters	Value
Routing Protocols	AODV, DSR
MAC Layer	802.11
Packet Size	512 bytes
Terrain Size	500m * 500m
Nodes	50
Mobility Model	Random waypoint
Data Traffic Type	CBR
No. of Source	5
Simulation Time	200 sec.
Node Mobility Speed	5, 10, 15, 20, 25, 30, 35, 40
CBR Traffic Rate	8 packet/sec
Maximum buffer size for packets	50 packets
No of Malicious Node	2, 3, 4

Average End-to-End Delay (second)

This includes all possible delay caused by buffering during route discovery latency, queuing at the interface queue, retransmission delay at the MAC, propagation and transfer time. It is defined as the time taken for a data packet to be transmitted across an MANET from source to destination. D = (Tr – Ts) Where Tr is receive Time and Ts is sent Time

Average jitter

Jitter is used as a measure of the variability over time of the packet latency across a network. A network with constant latency has no variation (or jitter). Packet jitter is expressed as an average of the deviation from the network mean latency. Jitter is cause by network congestion, timing drift, or route changes. At the sending side, packets are sent in a continuous stream with the packets spaced evenly apart. Due to network congestion, improper queuing, or configuration errors, this steady stream can become lumpy, or the delay between each packet can vary instead of remaining constant.

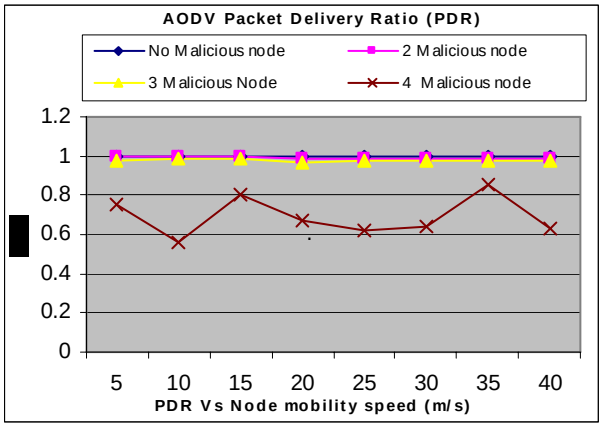


Figure 3: Packet Delivery Ratio vs. nodes mobility speed.

AODV Packet delivery ratio under wormhole attack

In AODV protocol if many nodes are sending and receiving data traffic simultaneously placing more malicious node uniformly in MANET network causes severe damage because it increases the probability of route affected malicious node. As show in figure 3 when no of malicious node are less (2 or 3) there is very less probability that any route involve malicious node and packet delivery ratio decreases only one percent as compare to the network that has no malicious node. But once the number of malicious node increases a particular level and it placed uniformly all over network effect of attack become severe as we can see in figure 3 when number of malicious node become 4, packet delivery ratio decreases significantly (Between 60% to 80%). One more important behavior is observed that packet delivery ratio under wormhole attack does not affected by node mobility speed.

AODV Average Jitter under wormhole attack

Jitter is another significant application layer parameter in mobile ad-hoc network especially in case where quality of

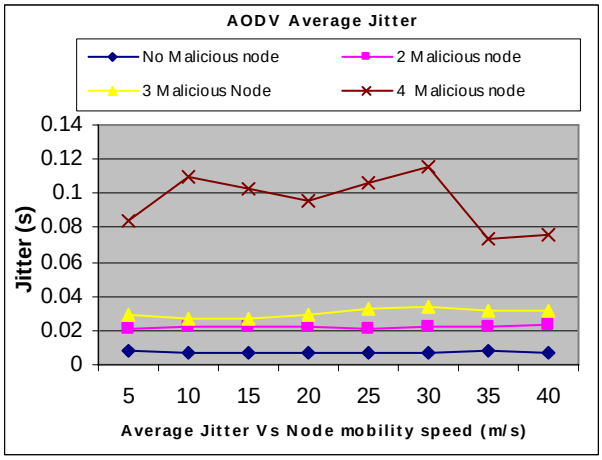


Figure 4: Average jitter vs. Nodes mobility speed.

Service is required. Study of wormhole attack effect on jitter in AODV protocol show that when the number of malicious node in mobile ad-hoc network are low (2 or 3) Jitter increase almost two times as compare to network without any malicious node. this is because when number of malicious nodes are less then number of route affected by these malicious node are also low which cause less delay. Another important characteristic can be seen from this figure 4 that in case of less malicious nodes in network (2 or 3) jitter increases as node mobility speed increases. When we increase number of malicious node from 3 to 4 there is a significant increase is jitter but at this case jitter decreases at very high node mobility speed (35 and 40 m/s).

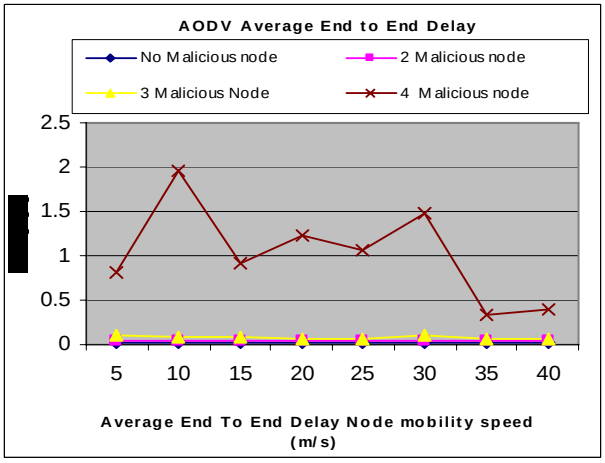


Figure 5: Average End to End-Delay vs Nodes mobolity speed.

AODV Average End to End delay under wormhole attack

Average End to End delay does not affected by the attack much when number of malicious nodes or less (2 or 3 nodes) also these is no change in End to End delay with respect to node mobility speed. However there is a significant increase in average end to end delay when number malicious node are high (4 nodes) and there is a negative relationship between End to End delay and node mobility speed. As we can see from figure 5 that in case when number of malicious node are high(4 nodes) with high node mobility (35 and 40 m/s) Average end to end delay become almost 3 time less as compare to other (5 to 30 m/s) node mobility speed.

DSR Packet delivery ratio under wormhole attack

In mobile ad-hoc network DSR protocol uses a complete list of node that contain by each packet that it has to traverse in order to reach destination node. This feature, although excludes intermediate nodes form making any routing decisions. Still From figure 3 and 6 we can see that DSR is more badly affected by wormhole attack as compare to AODV. And it shows that wormhole attack does not depend on working of intermediate nodes. When number of malicious nodes are less (2 or 3 nodes) packet Delivery ratio decreases as nodes mobility speed increase.

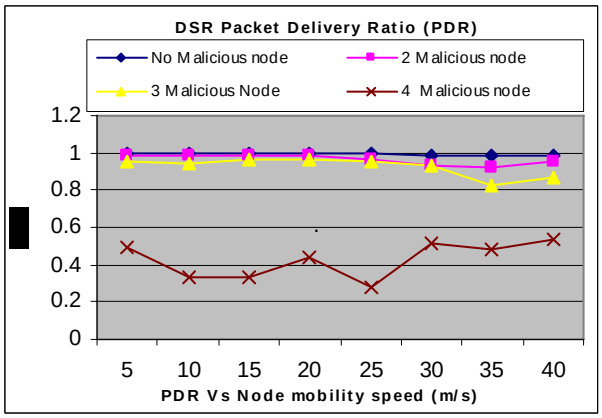


Figure 6: Packet Delivery Ratio vs. Nodes mobility speed

However result of packet delivery ratio with high malicious (4 nodes) node show that packet delivery ration increases as node mobility speed increases. As in case of DSR nodes maintain exiting or secondary route to it cache memory it increase the probability that a attack route is use by more than one source node to send traffic to destination node over a period of time which further magnify the impact of wormhole attack in DSR protocol.

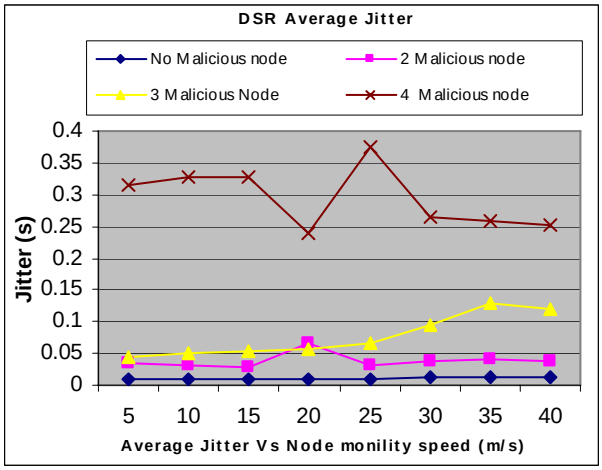


Figure 7: Average jitter vs. Nodes mobility speed.

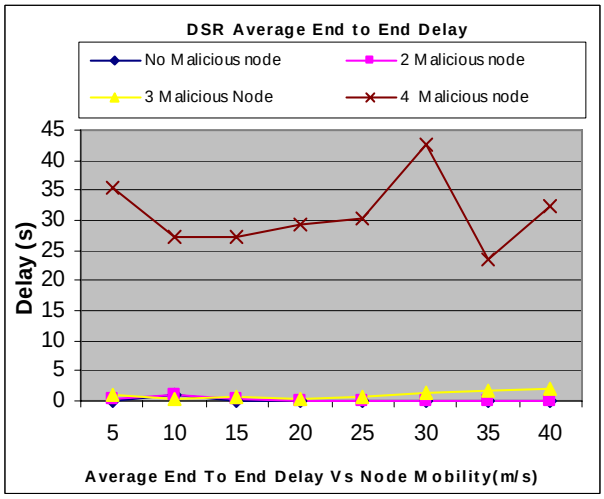


Figure 8: Average jitter vs. Nodes mobility speed.

DSR Average Jitter

AODV perform better then DSR under wormhole attack for jitter parameter. As DSR maintain existing route and secondary route so route discovery process is faster but this property DSR help wormhole attack to become more danger for DSR protocol. Once a root is attack by wormhole it used again and again by DSR since it maintain existing root from figure 5 we can see that average jitter become low when node mobility is high (35 to 40 m/s). In the case of AODV Average jitter is almost three times less than the DSR. When number of malicious node is less (2 or 3) and node mobility speed is also low average jitter is very low (Between.05 to.10 sec.). However average Jitter increases as node mobility speed increases. Performance of jitter with high number of malicious node (4 nodes) shows that average jitter is very high in this case (.30 to.35 sec) and as Jitter decreases as node mobility increases.

DSR Average End to End delay under wormhole attack

AODV outperform DSR when we compare Average End to End delay under wormhole attack. In the case of DSR there is no significant difference in average End to End delay when no malicious node present and less malicious nodes (2 or 3nodes) are place in network. But with high number of malicious nodes are high (4 nodes) average End to End delay increases significantly (between.25 to 43 sec.) as show in figure 8.

Conclusion

From the figure 3 to 8, we obtain some conclusion that under wormhole attack with CBR traffic sources, AODV perform better than DSR for packet delivery ratio, average jitter and End to End delay parameter on both low (2 or 3) and high (4 nodes) number of malicious nodes scenarios. In this paper, only two routing protocol are used and their performance have been analyzed under wormhole attack. This paper can be enhanced by analyzing the other MANET routing protocols under different mobility model and different type of attack.

References

[1] S. Das, C. E. Perkins, E. Royer, “Ad Hoc On Demand Distance Vector (AODV) Routing”, IETF Draft, June 2002

[2] C-K Toh “Ad Hoc Mobile Wireless Networks Protocols and Systems”, First Edition, Prentice Hall Inc, USA, 2002

[3] C.E. Perkins and E.M.Royer, “Ad-Hoc On Demand Distance Vector Routing”, Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, New Orleans, LA, USA, pages 90-100, February 1999.

[4] Elizabeth M. Royer and Chai-Keong Toh, “A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks”, IEEE Personal Communications, pages 46-55, April 1999.

[5] Fan Bai, Ahmed Helmy “A Framework to systematically analyze the Impact of Mobility on Performance of Routing Protocols for Adhoc Networks”, IEEE INFOCOM 2003

[6] Tracy Camp, Jeff Boleng, Vanessa Davies “A Survey of Mobility Models for Ad Hoc Network Research”, Wireless Communication & Mobile Computing (WCMC): vol. 2, no. 5, pp. 483-502, 2002

[7] D. Johnson, Dave Maltz, Y Hu, Jorjeta Jetcheva, “The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks”, Internet Draft, February 2002

[8] Suresh Kumar, R.K. Rathy and Diwakar Pandey, "Traffic Pattern Based Performance Comparison of Two Reactive Routing Protocols for Ad-hoc Networks using NS2", 2nd IEEE International Conference on Computer Science and Information Technology, 2009.
[9] D. Johnson, Y. Hu, and D. Maltz, "The Dynamic Source Routing Protocol (DSR) for Mobile", RFC 4728, Feb 2007
[10] S.Corson and J.Macker, "Routing Protocol Performance Issues and Evaluation considerations", RFC2501, IETF Network Working Group, January 1999.
[11] S. R. Biradar, Hiren H D Sharma, Kalpana Shrama and Subir Kumar Sarkar, "Performance Comparison of Reactive Routing Protocols of MANETs using Group Mobility Model", IEEE International Conference on Signal Processing Systems, pages 192-195 2009.
[12] C. Perkins, E. Belding-Royer, S. Das, quet, "Ad hoc On-Demand Distance Vector (AODV) Routing", RFC 3561, July 2003
[13] N.Aschenbruck, E.Gerhands-Padilla, P.Martini, "A Survey on mobility models for Performance analysis in Tactical Mobile networks, " Journal of Telecommunication and Information Technology, Vol.2 pp.54-61, 2008
[14] X. Hong, M. Gerla, G. Pei, and C.-C. Chiang, "A group mobility model for ad hoc wireless networks, " in ACM/IEEE MSWiM, August 1999.
[15] http://www-scf.usc.edu/~fbai/important/, referred on February 2010.
[16] http://nile.usc.edu/important/, referred on February 2010.
[17] Bai, Fan; Helmy, Ahmed (2006). A Survey of Mobility Models in Wireless Adhoc Networks. (Chapter 1 in Wireless Ad-Hoc Networks. Kluwer Academic. 2006.
[18] Broch, J; Maltz DA, Johnson DB, Hu Y-C, and Jetcheva J (1998). "A performance comparison of multi-hop wireless ad hoc network routing protocols". roceedings of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking(Mobicom98), ACM, October 1998.
[19] Broch, J; Maltz DA, Johnson DB, Hu Y-C, and Jetcheva J (1998). "A performance comparison of multi-hop wireless ad hoc network routing protocols". roceedings of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking(Mobicom98), ACM, October 1998.
[20] A. Perrig, Y. C. Hu, and D. B. Johnson, Wormhole Protection in Wireless Ad Hoc Networks, Technical Report TR01-384, Department of Computer Science, Rice University, 2001.
[21] Rashid Hafeez Khokhar, Md Asri Ngadi and Satira Mandala, "A Review of Current Routing Attacks in Mobile Ad Hoc Networks", International Journal of Computer Science and Security, pp. 18-29, Volume-2 Issue-3

Security Threats, Attacks and Countermeasure, Trust Model in Wireless Sensor Network: Research Challenges

Pranav Lapsiwala and Ravindra Kshirsagar

Electronics & Communication Department
Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, India
E-mail: plapsiwala@gmail.com, ravi_kshirsagar@yahoo.com

Abstract
A wireless Sensor network (WSN) consists of battery operated sensing and computing devices deploying for monitoring application. With the advent of new technologies WSNs are providing a new class of information to human beings, In most cases the network are stationary but as a revolutionary step the WSNs have to consider mobility. Recent research has provided means by which WSN has more secure communication or data aggregation. The unreliable network and deployment of nodes in open environments leads to several kind of attack. In this research paper analyzed existing attacks, counter measure model and methods are proposed.

Keywords: Wireless Sensor Network, Security issue, Trust based Model.

Introduction

The Wireless Sensor Network (WSN) consists of large number of deployed sensor nodes and base nodes. In the near future, due to immense development of MEMS technology the wireless sensor networks are anticipated to consist of thousand of inexpensive nodes, each having sensing capability with limited computational and communication power.

The node consists of one processing unit like microcontroller or microprocessor (generally 8 or 16 bit), memory unit and power unit. Base node or sink node received data from the deployed node which measure any physical quantity.

The sensor node can be used in variety of application such as military, environment, wildlife, home automation and health monitoring etc. As, node are deployed in environment, the network is more susceptible for attacks. The cryptographic solution is used for other network are not fusible due to resource constraints. This leads WSN in need of a new approach for providing security. It must be light weighted in terms of processing and communication cost in order to increase lifespan of the network.

- By using various counter measure algorithm process become delayed
- Overall system performance becomes very slow as number of nodes is increase.
- Multiple message transmission affected.

The section II describes various kind of attacks in wireless Sensor Network, Section III provides Trust and Trust based Management System, and Section IV discuss about Trust based model and Limitations followed by conclusion and references

Attacks in Wireless Sensor Networks

The wireless sensor networks are more prone to attacks as they are deployed in the environment. This section provides the list of existing attacks.

Sniffing attack

Sniffing attack is generally found in following kind of topology in which data can reach to the sink by hop to hop basis. A node can listen to all the packets sent by its all neighbors. However, if two nodes are exchanging information regarding routing, data or signaling information and the malicious node overhead it, then in future the malicious node can go for new types of attacks in the network.

Countermeasure

Cryptography of messages which needs point to point communication reduces the sniffing attack. But key management and heavy cryptography algorithm like RSA, DES, and AES make process slow and consume more power in processing rather in communication.

Hello Flood Attack

Hello packets are used to exchange routing information. Each node broadcast the information, if it comes to know that there is an update for shortest path to sink node. This creates flooding in entire network. An intruder will always try to broadcast the packet subsequently advertising low-cost routes.

The sensor node is generally deployed for application in the environment and nodes configured themselves and preparing the routing table for reaching sink via neighbor nodes with minimum hop count. In case of dynamic network every new node in the network broadcast the advertising of minimum hop count message and that will cross verify by other nodes. It causes the delay in entire process. The energy consume by the node for routing information is more rather data aggregation.

Countermeasure

The static and pre planned location of node avoid the flooding attacks in the network. If dynamic topology required then Hash key authentication algorithm is good solution with

lowest bit authenticate key.

Energy Drain Attack
The attacker initiates the large no of traffic through which the energy of the network gets drained. Eg. Denial-of-services attacks.

Countermeasure
As, WSN is energy constrained network and this kind of attack drain the network lifespan. The data aggregation and cluster head based approach overcome the energy drain attacks.

Droop Packet Attacks
In this type of attack node try to catch the packet and drop them. In multihop kind of network and data aggregation algorithm affect lot with this kind of attack

Countermeasure
Authentication key, secure routing algorithm with proper weight channel and monitor the behavior of the sleep and wake time of node with reduce amount of acknowledge and handshaking frame can over come Droop packet attacks.

Sink, black, grey and worm hole attack
In sink hole attack, false node misguide the other node of network by advertising shortest path to base station or sink node. The advertisement packet consist of black hole attack signifies the low latency channel for sink node. The attacker forward this packet to other node and try to find out other information by proper study of packet like node Id, packet size and authentication key and MAC address of the node. Where worm hole attack drop certain type of packet like routing and handshaking information packet. The combination of Sink, Black and worm hole packet Detroit the network performance.

Countermeasure
Secure routing algorithm with proper end to end communication. Static and fixed topology overcomes this problem.

Stealthy attack
In this type of attack, attacker tries to send the high or low value of data rather than aggregate data value. The stealthy attack is targeted for dynamic and distributed type of network where cluster head send the aggregate data to the sink node. Cluster head algorithm works on selection of minimum or aggregate data collected from the neighbor node. So, poor data aggregation algorithm easily broke by sending false value of aggregation function.

Countermeasure
To overcome stealthy attack the cluster head threshold design such a way that it can sense the drastic change of neighbor data. Cluster head has to compare the current data with the previous aggregate data and threshold data.

The attacks which discussed above are applicable for hop to hop architecture and every node depends on the neighbor node for data, routing and control information. For proper management of node functionality still it require more research work in synchronization, sleep and wake time MAC algorithm. To overcome intruder attacks trust based mode between the neighboring node proposed by many researchers and few popular model discuss in the section III but still more research work require for tiny device have low processing, power and memory capability.

Identify applicable sponsor/s here. ***(sponsors)***

Trust Management system
The trust is subjective, dynamic, asymmetric, non transitive and reflexive in characteristics. In the network, each node collects information about services provided by its neighbors. When a node needs the service of its neighbor nodes, it uses the collected information to calculate the trust, based on which it decides whether to get services with its neighbor or not.

The trust is generally for node trust, communication trust and path trust in the network. The following methods are for trust management system

Threshold of Trust value
The threshold value for particular value can be high or low depending on the application and trust based model and threshold value could be continuous or discrete.

Data Gathering
Node can be collecting the data implicitly based on observation of other entity or by explicitly sending request to other entity to send required information.

Based on trust value, the node performs necessary action with other entity. After action gets completed, based on result of action, the node updates the trust value. If entity behaves in the same manner as predicted, then update the value positively else negatively.

The Trust Model in Wireless Sensor Network
Agent based Trust Model
Chen et al., [7] proposed an agent based trust model. ATSN (Agent based trust on sensor network) runs at middle-ware of every agent node. As, agent based approach is applied for multi hop WSN communication topology. In that every node monitor the behavior of the neighbor node like forwarding data time and control frame time and processing time for algorithms. The ATSN and RFSN are almost similar with one another and differ with several uncertainties.

Limitation of Trust based Model
ATSN uses agent nodes with more power, long radio range and large storage space then normal sensor node to perform operations. ATSN work with fixed window and aging is specified by considering the positive outcome from current window. The agent node propagates the trust value to sensor node with encryption technique

Weight based Trust Model
Hur et al. [9] proposed a weight based trust model. The trust is used to eliminate data from malicious node, during data aggregation. The every node is capable enough to compare the receive data with sensed data duration and develop a weight trust model on it.

Limitation of Weight based trust model
The model highly based on synchronism phenomenon.

Beacon based Trust Model
Srinivasan et al. [6] propose a reputation based beacon system. The system was developed on the location information of the node. When node advertise beacon that time it also advertise its location to neighbor node. The node with location information consider as a malicious node and reputation develop with location information is quite slow process for huge node network.

A sensor node uses a neighbor reputation table to determine whether or not to use a given beacon's location information based on a simple majority scheme. This will help to find out malicious node.

Limitation of Beacon Trust Model
Beacon Trust model is depend on the neighbor reputation table that highly vulnerable for attacks.

Attack related to Trust Base Model in WSN
The good node sends the negative feedback for reputation.
There is little method by which malicious node try to establish trust in WSN.

- The node just spoofs their identity with different MAC address or with different authentication key id.
- The malicious node sends different trust or threshold value for all neighboring node to develop the trust.

Conclusion
Importance of Wireless Sensor Network can not be denied as the world of computing is getting portable and compact. Unlike wired networks, WSN pose a number of challenges to security solution due to their unpredictable topology, wireless shared medium, heterogeneous resources and stringent resources etc. Security is not a single layer issue but multilayered issue. It requires multi fence security solutions that provide complete security spanning over the entire protocol stack. The existing protocols are typically attack-oriented in that they first identify several security threats and then enhance the existing protocol or propose a new protocol. Therefore, more ambitious goal for WSN, MANET and ad hoc network security is to develop a multifence security solution that is embedded into possibly every component in network, resulting in depth protection that offers multiple lines of defense against many both known and unknown security threats.

References

[1] C. Karlof, D. Wagnor, “Secure routing in wireless sensor networks: attacks and countermeasures”, University of California at Berkley, USA.

[2] D.Martin, H.Guyennet,”Wireless Sensor Network Attack and Security Mechanisms: A short Survey”, in 13th IEEE International Conference on Network-Based Information Systems(NBIS),14th -16th Sept., 2010, France, pp. 313-320.

[3] F.Stajano, R.J.Anderson, The resurrecting duckling: Security issue for ad-hoc wireless networks, in seventh international security protocol workshop, 1999, pp. 172-194.

[4] L.Zhou, Z.Hass, Securing ad hoc networks, IEEE network magazine 13 (6) (1999) 24-30.

[5] A.Perrig, R.Szewczyk, J.D.Tygar,V.Wen and D.E. Culler, “ SPINS: Security protocol for Sensor Networks” in proceeding of 7th Annual International Conference on Mobile Computing & Networks (MOBICOM), July 2001, University of California, Berkley, USA., pp.189-199.

[6] D.Liu, P.Ning and W.Du, “ Detecting Malicious Beacon node for Secure Location Recovery in Wireless Sensor Networks” in the 25th IEEE International Conference on Distributed Computing Systems (ICDCS ’05), 2005.

[7] H.Chen, H.Wu, X.Zhou,”Reputation-based Trust in Wireless Sensor Network”, in IEEE International Conference on Multimedia and Ubiquitous Engineering, 26th -27th April, (MUE’07), 2007, Shanghai.,pp.603-607.

[8] A.Boukerche, Xu. Li,”An Agent-based Trust and Reputation Management Scheme for Wireless Sensor Networks” in IEEE International Conference on Global Telecommunication conference, 28th Nov. -2nd Dec., (GLOBECOM’05), CANADA.,pp 5

[9] P. Lapsiwala, R. Kshirsagar, “Authentication and Intrusion Detection Topology for Wireless Sensor Network” in International Conference on Electronics, Information and Communication System Engineering (ICEICE ’10), Jodhpur, India.

General Lightweight Scheduling in Game Artificial Intelligence

Mr. Trevor Adams and Dr. Clive Chandler

Faculty of Computing, Engineering and Technology, Staffordshire University, Staffordshire, England
E-mail: t.j.adams@staffs.ac.uk c.chandler@staffs.ac.uk

Abstract
Game Artificial Intelligence requires an interactive AI, which by its very nature presents many challenges to game developers. As AI tasks become more complex, the need to manage the execution of those tasks becomes more important. All but the most complex routines can be managed with some simple abstractions for execution management. These abstractions, through extension, could be used to map functionality onto hardware specific implementations; paving the way for hardware thread support and multi-core support. This paper discusses the need for scheduling in game artificial intelligence (AI); it presents a design for a lightweight scheduler that forms the basis of an extensible framework for basic control flow and execution. Whilst not part of a larger API the design can be compatible with and adaptable to the general tasks of AI control.

Index Terms: Games, AI, Software Engineering, Scheduling

Introduction
GAME AI presents many challenges. Computer games operate over many hardware architectures and are built using a myriad of software frameworks. Games are now presenting increasingly engaging interactions with players. Platforms are now able to give more hardware resources to AI routines as hardware is devoted to other areas, chiefly graphics. As AI becomes more complex, a mechanism is required to manage the load and facilitate smooth running. A scheduler is a computing control system designed to manage the execution of code. Chiefly, the purpose is to make good use of available resources by distributing tasks efficiently. Efficiency in this case is determined by the application requirements and the constraints of the host environment. Game developers have been using finite state machines (FSM) to manage game interactions for many years [1], but recently developers have also begun using behaviour trees in place of hierarchical FSM as they rely on simpler primitives and scale well to more complex scenarios [2].

Game AI is a good focus for the purpose of a general scheduler as many games present indeterminate problems at any given point in time. A game such as Madden NFL 2011 [3] may be calculating a path, traversing a path and planning strategy for one agent or multiple agents simultaneously and a version of the game may be produced for multiple platforms. A high level, abstract system for scheduling can help an AI system manage the execution of these tasks.

There are three key elements to a game AI scheduling system [4];

- Having algorithms that can work over multiple frames, and yield results in any frame (Anytime).
- Dividing execution amongst routines (Frame / Phase).
- A mechanism for giving preferential treatment to high priority operations (Prioritisation).

A scheduler is a core support structure to general game AI routines [2]. Each of these elements will be considered and a set of suggested classes that are used to implement a solution.

Anytime Algorithms
Anytime algorithms are designed, to a certain extent, to emulate the way a human may perform an action – we start acting before we have finished thinking [4]. The idea that a first guess will be roughly correct and begin acting upon it allows an AI to be responsive to input. The extent of visible thinking will be determined by the game that is hosting the AI. Anytime algorithms are used in commercial games and are often bespoke, you can see evidence of anytime algorithms when playing games such as StarCraft 2 [5], a real time strategy (RTS) game. When a unit is directed to travel to a currently unexplored region of the game world it responds immediately and is able to traverse the route without having to display any thinking time; the agent heads in the general direction and updates information as time goes on.

Anytime algorithms have been shown to perform well in regards to pre-calculated counterparts [6]. Anytime algorithms compliment scheduling as the nature of them involves being interruptible. It could be said that interruptible, anytime algorithms add to the illusion of intelligence within a game due to responsiveness and retracing that happens when better information is found [7]. As movement and path finding are often the most time consuming processes[4], they tend to be prime candidates for anytime algorithms in the constrained environments of game development.

The nature of anytime algorithms requires that an AI agent have access to a resource for the purpose of memory [8]. A blackboard system as used in real time team planning [9] and behaviour systems [8] would be ideal for the purpose of scheduling. A blackboard system can also be made thread safe and can be used by multi-core processor systems, easing the load placed on context switching. Shared context and centralised data can also assist in breaking down the predictable nature of game AI staple techniques E.g. FSMs.

Dividing Execution

Millington and Funge [4] present a simple notation for execution contexts and refer to them as behaviours. Behaviours are the core execution component for a game AI entity. Whether a game is to use a basic FSM for behaviours [1], select behaviours dynamically [10] or use a hybrid approach [11], having a behaviour be a fixed point of execution provides a context for a scheduler system.

```
interface IProcess {
   void Update()
}
class Behaviour {
   int Frequency
   int Phase
   IProcess Process
   void Execute()
}
```

Some schedulers will refer to behaviours as a composition of tasks. Tasks have their own context, referred to as a closure [2], and can be configured to return information such as completion status.

Most games code operates by way of a game update loop [12]. Games tend to operate using a basic loop structure. Input is gathered; game logic is updated and then rendered to the screen. AI tasks are required to run within the logic update section of the game loop. Many tasks are run during this time, including graphical operations. The lightweight scheduler will split tasks over a developer-supplied frequency and stagger them using a phase identifier. This is a default behaviour supplied as part of the framework, it need not be fixed and could be exchanged through use of the strategy pattern [13]. A record class will keep an instance of the behaviour object, the frequency of execution and a phase step. As it is likely a game will require multiple tasks to be scheduled, a schedule manager will be required to maintain a list E.g.

```
foreach(Behaviour b in listBehaviours)
{
   if b.Frequency % (b.Frame + b.Phase)
   b.Execute()
}
```

The phase is added to the frame to cater for the number of behaviours being greater than the amount of frames [4]. The scheduler will also execute the schedule plan and provide access to a mechanism for results and feedback. The observer pattern [14] can be used for the purpose of call back and notification, allowing a manager to report back. The interface can be created generally so that it may be implemented for the purpose of asynchronous call back, taking advantage of threading where available. Much like the rest of the design, the point is to be capable at the basic level and provide constructs so as to be extensible when further functionality is required.

Scheduling & Prioritisation

The logic of scheduling tasks will be wrapped into its own interface type so that it may be substituted for different logic at run time. A pseudo code listing of the scheduler strategy:

```
interface IScheduleStrategy{
   void Update(List<Behaviour>,float
frame)
}

class Scheduler {
   float frame
   List<Behaviour> list
   IScheduleStrategy strategy
   void Update() {
                                    frame++
            strategy.Update(list, frame)
   }
}
```

Using concrete strategy:

```
class Standard: IScheduleStrategy{
   void Update(List<Behaviour>,float
frame)
   {
    foreach(Behaviour b in listBehaviours)
{
   if b.Frequency % (b.Frame+b.Phase)
   b.Execute()
   }
   }
}
```

Now that the execution logic is wrapped into a class, any type of scheduling scheme may be employed. A manual-phasing scheme would require intervention from the programmer and knowledge of the code to be executed. In more complex scenarios, the constructs may be used to provide feedback by way of a return status. For example, two possible routines could be using mutually prime frames and timings taken for use Wright's Method for automatic phasing [4] and execution.

The core functionality can also be easily extended to handle hierarchical scheduling. The composite pattern [13] is a good choice for this. Depending on the programming language used, the abstract entities (scheduler/behaviour) can be implemented as an abstract class or interface type. This will allow behaviour and a scheduler to be used interchangeably and function as a hierarchy. A main schedulers list of behaviours, when executed, may in turn be sub-schedulers that hold their own list of behaviours to execute. This architecture is shown by way of a diagram in Fig. I. A hierarchical system could be expanded towards the use of behaviour trees and goal-oriented action planning [10].

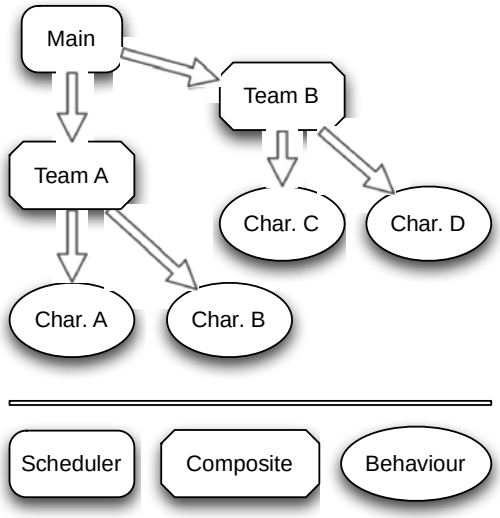


Fig. I Hierarchical scheduling system using composite classes. Inheriting a base class or implementing an interface can allow an object to fulfil the role of both scheduler and behaviour, effectively allowing schedulers to be executed as tasks.

Delimitations

The initial concept for a lightweight scheduler was born of the need for a test harness application for AI, using an automated unit testing library and mock objects. As such, the current design does not currently cater for parallelism, by way of threading or multi-core processing. The chosen language for initial implementation is C# using the XNA framework, although the design should be portable to any object-oriented (OO) language common to game programming E.g. C++. This language was chosen to take advantage of available rapid development tools. A small test harness application has been developed based on the Westworld Adventure game created by Mat Buckland [1]. The high level aim of the design is to encapsulate executable behaviours using a common interface. This behaviour can then be added to a scheduler object along with a given schedule, which will incorporate the behaviour to execute, the frequency of execution and a phase. The high level objects can be simple at first and later extended to incorporate more complex scheduling patterns. It is not intended to be a professional level tool but could easily be used for simple games.

Conclusions

General game AI frameworks, such as those developed by the AI Interface Standards Committee (AIISC) [15], have given rise to a paradigm shift in how developers consider implementing AI. It is commonplace to see graphical application programming interfaces (API) and abstractions related to rendering visuals, but less so in terms of AI. This is an obvious area for advancement as many AI tasks could fit into a common API and there is currently a lot of work in the AI community in moving towards an AI API. The lightweight scheduler as presented in this paper is designed to be the basis for an extensible framework for basic control flow and execution. While not part of a larger API, this scheduler design is intended to be compatible and adaptable to the general tasks of AI control. For the purpose of game development, it would be possible to extend the design presented in this paper to cater for multi-processor environments (language specific extensions) and software threads. The system could also support priority scheduling using basic computing data structures. The current design lacks any sort of load balancing or the ability to automatically detect and/or halt behaviours that have not completed successfully or failed to finish in a desired timescales.

References

[1] M. Buckland, *Programming Game AI by Example*. Plano, TX: Wordware Publishing Inc., 2005.
[2] A. J. Champandard, "Getting Started with Decision Making and Control Systems," in *AI Game Programming Wisdom 4*, S. Rabin, Ed., ed Boston, MA: Course Technology PTR, 2008, pp. 257-264.
[3] E. C. EA Tiburon, "Madden NFL 11," ed: EA Sports, 2010.
[4] I. Millington and J. Funge, *Artificial Intelligence for Games*, 2nd Edition ed. Burlington, MA: Morgan Kaufmann, 2009.
[5] Blizzard Entertainment, "Starcraft 2," ed: Blizzard Entertainment, 2010.
[6] R. Butt and S. J. Johansson, "Where do we go now?: anytime algorithms for path planning," presented at the Proceedings of the 4th International Conference on Foundations of Digital Games, Orlando, Florida, 2009.
[7] A. Nareyek, "AI in Computer Games," *Queue*, vol. 1, pp. 58-65, 2004.
[8] J. Orkin, "Agent architecture considerations for real-time planning in games.," in *Artificial Intelligence in Interactive and Digital Entertainment (AIIDE)*, Marina del Ray, CA, 2005.
[9] K. McGee and A. T. Abraham, "Real-time team-mate AI in games: a definition, survey, \& critique," presented at the Proceedings of the Fifth International Conference on the Foundations of Digital Games, Monterey, California, 2010.
[10] J. Orkin, "Three States and a Plan: The A.I. of F.E.A.R.," presented at the Game Developers Conference 2006, 2006.
[11] N. Lau, "Knowledge-Based Behaviour System - A Decision Tree / Finite State Machine Hybrid," in *AI Game Programming Wisdom 4*, S. Rabin, Ed., ed Boston, MA: Course Technology PTR, 2008, pp. 265-274.
[12] B. Schwab, *AI Game Engine Programming*. Boston, MA: Course Technology PTR, 2009.
[13] A. Shalloway and J. R. Trott, *Design Patterns Explained: A New Perspective on Object-Oriented Design*: Addison-Wesley Professional, 2001.
[14] E. Freeman, E. Freeman, B. Bates, and K. Sierra, *Head First Design Patterns*: O' Reilly \\& Associates, Inc., 2004.
[15] B. Yue and P. de-Byl, "The state of the art in game AI standardisation," presented at the Proceedings of the 2006 international conference on Game research and development, Perth, Australia, 2006.

Analyzing Performance of Counter-Based Broadcasting in Mobile Ad Hoc Networks

M. Deshmukh

Assistant Professsor, Pillai's college of Engineering, Mumbai, India
E-mail: manjudeshmukh@rediffmail.com

Abstract
Broadcasting is a fundamental service in mobile ad-hoc networks (MANETs). Flooding is used as a broadcast technique for route discovery in MANET. But it can result in high packet collision & redundancy which can degrade the network performance. Such a scenario is referred to as broadcast storm problem. Counter-based broadcasting (CBB) scheme has been proposed to overcome the broadcast storm problem in MANET. Counter-based approaches inhibit a node from broadcasting a packet based on number of copies of the broadcast packet received by the node within a random access delay time. It relies on the threshold value to decide whether or not to forward broadcast packet. In our study, counter-based threshold is dynamically adjusted based on host density in its neighborhood area. Simulation results of this study show the effect of threshold on the performance of proposed counter based flooding scheme.

Keywords: Counter based Broadcasting, CBB, Broadcast Storm Problem

Introduction

Wireless networks are becoming more and more important. People want their mobile and fixed devices to communicate with the hassle of wires. Preferably communication should be established automatically in an ad-hoc fashion. To achieve this, Mobile ad-hoc networks (MANETs) will be an important building block. Mobile ad-hoc networks (MANETs) are special type of wireless networks that comprises of wireless mobile nodes, which communicate with one another without relying on fixed infrastructure or central administration. The main advantage of this is communicating with rest of the world while being mobile. The distributed, wireless and self-configuring nature of MANET make them suitable for a wide variety of applications. These include critical military operations as well as disaster recovery scenario.

Broadcasting is a means of diffusing a message from a given source node to all other nodes in the network. Broadcasting is a fundamental operation in MANETs and a building block for most other network layer protocols. Most existing routing protocols proposed for MANET use flooding as a broadcast technique for route discovery.

One of the earliest broadcasting mechanisms proposed in the literature are simple or blind flooding where each node in the network retransmits a message to its neighbor upon receiving it for the first time. Although, flooding is simple and most commonly used for broadcasting in MANETs; it can result in high packet collision & redundancy which can degrade the network performance. Such a scenario is referred to as broadcast storm problem [1,2,3].

A counter-based method has been suggested in [1,4,5] as a means of reducing redundant broadcasts and alleviating broadcast storm problem. The counter-based method is based on counter c that records the number of times a host has received the same broadcast packet and is maintained by each host for each broadcast packet. When c reaches a certain threshold, the packet is dropped otherwise packet is retransmitted. The counter-based scheme can reduce the number of rebroadcasts, and as a result reduce the chance of contention and collision among neighboring nodes. Counter-based approaches inhibit a node from broadcasting a packet based on number of copies of the broadcast packet received by the node within a random access delay time.

This study introduces an efficient class of Counter-based Flooding scheme that has been proposed to overcome the broadcast storm problem in MANET. It relies on the threshold value to decide whether or not to forward broadcast packet. This is done based on locally available neighborhood information and without requiring any assistance of distance measurements or exact location determination devices. A straightforward method for gathering neighborhood information at a given node involves the periodic exchange of Hello packets between neighbors to construct a one-hop neighbors list at the nodes. The proposed algorithm is a combination of counter-based and knowledge-based methods.

The rest of this paper is organized as follows: In Section 2, we introduce the related work of broadcasting in MANETs. In section 3, we describe our proposed Counter-based Flooding scheme. In Section 4, we evaluate our approach and present the simulation results and compares current and proposed work. Section 5 concludes the paper and offers suggestions for future work.

Related Work

One of the earliest broadcasting mechanisms proposed in the literature is simple or blind flooding where each node in the network retransmits a message to its neighbor upon receiving it for the first time. Although, flooding is simple and most commonly used for broadcasting in MANETs; it can result in high packet collision & redundancy which can degrade the network performance. Such a scenario is referred to as broadcast storm problem [1,2,3].

Most sophisticated solutions have been proposed to alleviate the broadcast storm problem associated with blind flooding. Some of solutions inhibits some hosts from forwarding the broadcast packet for the sake of reducing redundancy and hence collision and contention. These solutions include probability-based, counter-based, distance based, area-based, cluster based and neighbor knowledge based schemes. In the probability based scheme [6,7], when receiving a broadcast packet for the first time, a node rebroadcasts the packet with a probability p; when p=1, this scheme reduces to blind flooding. The counter-based scheme [1,4,5] inhibits the rebroadcasts if the packet has already been received for more than a given number of times. In the distance-based scheme [1], a node rebroadcasts the packet only if the distance between the sender and the receiver is larger than a given threshold. An area-based scheme [1] uses pre-acquired location information of neighbors to make broadcasting decision. In neighbor knowledge-based schemes [8], a decision of retransmission of packet is made based on neighborhood information collected using periodic Hello packet exchange. Clustering is another method to select forwarding nodes as addressed in [1]. It groups nodes into clusters. A representative of each cluster is called as cluster head. A node that can communicate with a node in another cluster is called as gateway. Other nodes are called ordinary nodes. Nodes co-operate to elect cluster heads by periodically exchanging information. Cluster head broadcasts all other nodes in the same cluster. To rebroadcast packets to nodes in another cluster, gateway nodes are used. When the broadcast message is heard, if the host is a non-gateway member, the rebroadcast is inhibited and the procedure exits. If the host is either a head or a gateway, any of the probabilistic, counter-based, distance-based, and location-based schemes is used to determine whether to rebroadcast or not.

Broadcasting in MANETs is an active research area. The core problem is how to minimize the number of rebroadcast packets while maintaining reasonable latency and good reachability. Transmitting a large number of rebroadcasts does guarantee high reachability. However, it degrades the network throughput and potentially incurs long broadcast latency. Dispatching fewer rebroadcasts leads to lower bandwidth wastage, higher throughput and lower broadcast latency. However, sending too few rebroadcasts may cause a rebroadcast chain to be broken so that some hosts may never receive the broadcast packet resulting in lower reachability. On the other hand, rebroadcast algorithms may also affect the broadcast latency. For example, a host may dictate a random delay before sending a rebroadcast in order to reduce the chance of collisions. A long delay leads to high broadcast latency. A. Mohammed and M. Khaoua [4] focused on determining the best counter threshold value for counter-based approach. A. Mohammed, revealed in [4] that setting the optimal threshold values can optimize the performance of counter based flooding in terms of saved broadcast and end to end delay. Most counter based schemes assumed a counter threshold value 3 or 4. S.Y. Ni has concluded in [1] that a threshold value 3 or 4 can save many broadcasts in a dense region while achieving a delivery ratio comparable to blind flooding. On the other hand, larger threshold of greater than threshold value 6 will provide less saving of broadcasts in sparse region but behave almost like blind flooding in terms of reachability.

Y.C. Tseng, S.Y. Ni have proposed in [3] schemes to reduce redundant rebroadcasts and differentiate timing of rebroadcasts to alleviate this problem. They have shown that in most cases counter based flooding does not achieve high degree of reachability because each node has the same thresholds value to rebroadcast packets regardless of its surrounding, e.g. number of neighbors. The problem derives from the uniformity of the algorithm; every node has the same thresholds value to rebroadcast a given packet.

Proposed CBB Scheme

The idea of counter based scheme is based on the inverse relation between the expected additional coverage (EAC) and number of duplicate broadcast messages received [1]. A node is prevented from a retransmitting a broadcast packet when the EAC of the nodes rebroadcast is low. The idea of EAC is depicted in Figure 1.

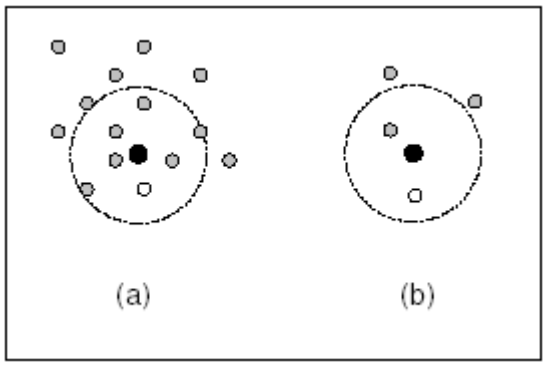


Figure 1: Expected Additional Coverage Example.

The hollow shaped nodes are source nodes that initiate the broadcast transmission and solid black nodes are nodes we use to clarify our ides ; we will refer to them as black-a and black-b. Apparently, the neighborhood density of black-a is higher than black-b. Therefore, the number of duplicate broadcast packets that would be received by black-a is higher as well. Moreover nodes within transmission range of black-a would have been reached by other forwarding nodes. Therefore, the EAC of black-a is lower than the EAC of black-b.

In counter based scheme, when a node receives a flooding packet, it will wait for random assessment delay (RAD). During this period the counter is counting the number of packets. The length of RAD is randomly chosen from uniform distribution between from 0 to Tmax seconds where, Tmax is the maximum delay time. When RAD expires, the counter does not reach threshold k, the node will retransmit the packet.

We have following remarks on existing counter-based broadcasting methods. The topology of MANET is often random and dynamic with varying degree of node density in various regions of the network as shown in Figure 2. The network may contain sparse and dense regions. In dense regions, multiple nodes share similar transmission range. Therefore the thresholds control the frequency of rebroadcasts

and thus might save network resources without affecting delivery ratios. Note that in sparse region there is much less shared coverage; thus some nodes will not receive all the broadcast packets unless the threshold parameter is low. Therefore, fixed counter threshold approach suffers from unfair distribution of C since every node is assigned the same value of C regardless of its local topological characteristics. Ideally, the threshold value *c* should be high if a node located in a dense region while relatively low if a node located in a sparse region. If *c* is too high reachability might be poor while if *c* is set too low, many redundant rebroadcasts might be generated. While using small threshold values provides significant broadcast savings. Unfortunately, the reachability will be poor. There exist a tradeoff between reachability and saved broadcast.

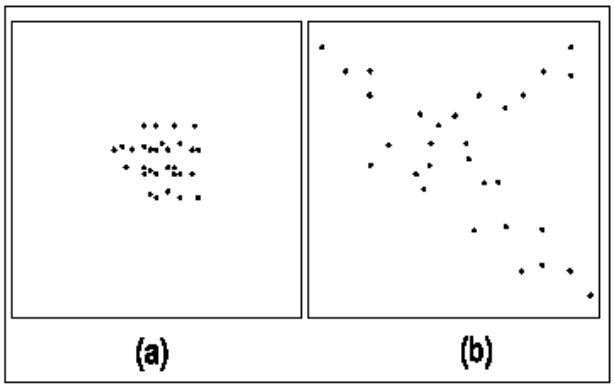


Figure 2: Example of changeable network topology (a) Dense region with 30 nodes (b) Same region with nodes forming several sets of sparse region

In order to achieve both high saved broadcast and high reachability when network topology changes frequently, the threshold should be set low for the nodes located in sparse regions and high for the nodes located in dense regions. The need for dynamic adjustments thus rises. Accordingly, sparse regions need a higher chance to rebroadcast than dense networks. This could be achieved by doing following modifications to fixed counter threshold approach. For dense region, a large threshold value C2 is used. For sparse region, a small threshold value C1 is used.

A high number of neighbors implies that the hosts in dense region, a low number of neighbors imply that the hosts in sparse region. That means we increase the counter-based threshold value if the value of the number of neighbors is too high. Similarly, we decrease the counter based threshold value if the value of number of neighbors is too high.

The proposed algorithm dynamically adjusts the counter based threshold value C at each mobile host according to the value of the local number of neighbors. The value of threshold changes when the host moves to a different neighborhood. A straightforward method for gathering neighborhood information at a given node involves the periodic exchange of Hello packets between neighbors to construct a one-hop neighbors list at the nodes. The proposed algorithm is a combination of counter-based and knowledge-based methods.

On hearing a broadcast packet m at node X for the first time, it does not immediately broadcast the packet. It waits for Random assessment delay (RAD). It finds the degree of X that is the number of neighbors of node X. If the degree of node X is less than the average number of neighbors (that means the current node is located in a sparse network), then the value of counter threshold (C1) is set to low value. If the degree of node X is greater than or equal to the average number of neighbors (that means the current node is located in a dense network), then the value of counter threshold (C2) is set to high value. During RAD, the counter is counting the number of repeated packets received. When the RAD expires, if counter c is less than threshold, the packet is broadcasted. Otherwise the packet is dropped. Additionally the values C1 and C2 are selected in a way that considers the expected additional coverage EAC. That is C2 (dense network threshold) should be in a way larger than C1 (sparse network threshold) in order to redundancy in a dense area.

We present an estimate of average neighbor number as the basis for the selection of threshold as given in Eq. (1). Let A be an area of an ad hoc network, N be the number of mobile nodes in the network and R be the transmutation range. The average number of neighbor's n can be obtained as shown below [6]

$$\bar{n} = \frac{(N-1) * 0.8 * \pi * R^2}{A}$$

Equation 1

Performance Evaluation

We evaluate the performance of our proposed algorithm using NS-2 simulator. NS-2 is a discrete event simulator targeted at networking research for wired and wireless networks [9]. This research work suggests and investigates the performance of new counter based flooding algorithms where the threshold values at a node is dynamically adjusted as per the node coverage distribution and movement using one-hop neighborhood information to increase reachability and saved rebroadcast.

Simulation Parameters

The parameters used in the following simulation experiments are listed in Table 1. Each node in the network has a constant transmission range of 250 meter. The MAC layer scheme follows the IEEE 802.11 MAC specification. The simulation is allowed to run for 900 seconds for each simulation scenario. We have used the broadcast mode with no RTS/CTS/ACK mechanisms for all packet transmissions including Hello and DATA packets. The movement pattern of each node follows the random way-point model. Each node moves to a randomly selected destination with a constant speed between 0 and the maximum speed. When it reaches the destination, it stays there for a random period and starts moving to a new destination.

Table 1

Parameter	Value
Network Area	800*800
Transmitter range	250 meter
Simulation Time	900 sec
Channel Bandwidth	2 Mb/sec
Pause time	2 ms
Maximum speed	20 m/sec
Mobility model	Random way point
Number of nodes	20,40,60,80,100, 120

Performance Measures

The performance of flooding algorithm can be measured by a variety of metrics [1,2,3,4]. A commonly used metric is the number of message re-transmissions with respect to the number of nodes in the network. In this work, we use *rebroadcast savings*, which is a complementary measure and is precisely defined below. The next important metric is *reachability*, which is defined in terms of the ratio of nodes that received the broadcast message out of all the nodes in the network. The formal definitions of these two metrics are given as follows [6].

***Saved Rebroadcasts (SRB)*:** Let *r* be the number of nodes that received the broadcast message and let and *t* be the number of nodes that actually transmitted the message.

The saved rebroadcast is then defined by *(r – t)/r*.

***Reachability (RE)*:** is the percentage of nodes that received the broadcast message to the total number of nodes in the network. For useful information, the total number of nodes should include those nodes that are part of a connected component in the network.

Performance Analysis

The objective of this work is to present the performance of proposed algorithm using fixed counter based broadcasting algorithm as well as blind flooding algorithm. Our main idea is to reduce the rebroadcasting number. To attain the objective, we vary the density by increasing number of nodes deployed over a fixed area of 800 * 800 m. The number of nodes has been varied from 20 to 120 in steps of 20 nodes with each node moving at a speed 0f 2 m/sec. To reduce traffic load, one node was randomly chosen to initiate the broadcast process at a sending rate of 2 packets per second.

Effect of Threshold

In this section, we evaluate the effect of threshold on the performance of proposed counter based flooding scheme. In this study, we evaluated the performance by changing the dense threshold to higher value over increasing number of nodes. The number of nodes has been varied from 20 to 120 in steps of 20 nodes. Figure 3 and Figure 4 presents the saved Broadcast and Reachability respectively, showing the effect of dense threshold.

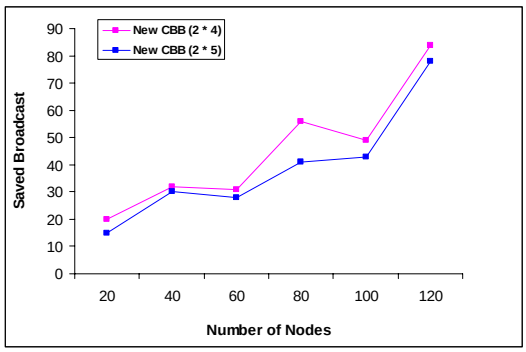


Figure 3: Saved Broadcast showing the effect of Threshold

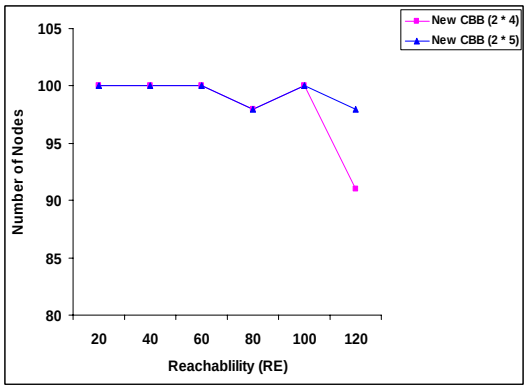


Figure 4: Reachability showing the effect of Threshold

The higher the dense threshold lower be the saved broadcast. But higher threshold value achieves more reachability as compared lower threshold in dense region. Lower the dense threshold higher is the saved broadcast, but results in poor reach-ability in dense region as compared to higher threshold.

Conclusion

The proposed counter based flooding algorithm in mobile ad hoc networks (MANETs) is devised to improve the saved Rebroadcast. The algorithm determines the counter threshold by considering the network density. In order to increase the saved rebroadcasts, the threshold of low density nodes is decreased while that of high density nodes is increased. We also evaluated the effect of threshold on the performance of proposed counter based flooding scheme. Lower the dense threshold higher is the saved broadcast, but results in poor reachability in dense region as compared to higher threshold. The higher dense threshold lower be the saved broadcast. But higher threshold value achieves more reachability as compared lower threshold in dense region. The optimal threshold value has significant effect on the overall performance of the proposed algorithm.

As a continuation of this research, further refining the threshold using more refined levels for nodes density regions, would lead to an improvement in the performance of proposed counter based flooding.

References

[1] S.Y. Ni, Y.S Chen, J. P. Sheu, The broadcast Storm Problem in a mobile ad hoc network, Wireless Networks, May 2002.

[2] B. Williams, T.Camp, Comparison of broadcasting techniques for mobile ad hoc networks, Third ACM International Symposium on Mobile Ad Hoc networking and Computing, June 2002.

[3] Y.-C. Tseng, S.-Y. Ni, and E.-Y. Shih. Adaptive Approaches to Relieving Broadcast Storms in a wireless multihop mobile ad hoc network, Proceedings of the 21st International Conference on Distributed Computing Systems,2001

[4] A.Mohammed, M. Khaoua, L.M. Mackenzie, Optimizing the Threshold value for Counter-based Broadcast scheme in MANETs, ISBN, 2007

[5] W.Peng, X. Lu, On the reduction of broadcast redundancy in mobile ad-hoc networks, MOBIHOC, 2000.

[6] M. B.Yassein, M.O. Khaoua, L.M. Mackenzie, Improving the Performance of Probabilistic Flooding in MANETs, IWWAN ,2006

[7] M. B.Yassein, M.O. Khaoua, S. Papanastasiou, Performace Evaluation of Flooding in MANETs in the Presence of Multi-Broadcast Traffic, IEEE 2005, 2005.

[8] J.Wu, F.Dai, Broadcasting in ad-hoc networks based on self-pruning, Infocom 2003, April 2003.

[9] K. Fall and K. Varadhan. The NS manual, the VINT project http://www.isi.edu

[10] M. Khalaf, Ahmed Y, A New Adaptive Broadcasting Approach for Mobile Ad-hoc Networks, IEEE 2010.

[11] Lewis M, M. Khaoua, Dynamic Probablistic Counter-based Broadcasting in MANET,IEEE 2010.

Enhanced Ant Colony based Routing in MANETs

[1]Mohammad Arif and [2]Dr. Tara Rani

[1]Research Scholar, Singhania University, Jhunjunu, Rajasthan, India
E-mail: arif_mohd2k@yahoo.com
[2]Assoc. Prof, NICE College of Technology, Agra, UP, India
E-mail: drtararani@rediffmail.com

Abstract
Mobile ad hoc network (MANET) is a collection of wireless mobile nodes dynamically forming a temporary network without the use of any preexisting network infrastructure or centralized administration i.e. with minimal prior planning. All nodes have routing capabilities and forward data packets for other nodes in multi-hop fashion. Nodes can enter or leave the network at any time, and may be mobile, so that the network topology continuously experiences alterations during deployment. The biggest challenge in MANETs is to find a path between communicating nodes. The considerations of the MANET environment and the nature of the mobile nodes create further complications which results in the need to develop special routing algorithms to meet these challenges. Swarm intelligence, a bio-inspired technique, which has proven to be very adaptable in other problem domains, has been applied to the MANET routing problem as it forms a good fit to the problem. In this paper, we have studied Ant Colony based routing algorithms i.e. two popular Ant based algorithms, AntHocNet and the Ant Routing Algorithm (ARA). A thorough analysis of ARA is carried out based on the effect of its individual routing mechanisms on its routing efficacy. The original ARA algorithm, although finds the shortest path between source and destination, is observed to not be competitive against other MANET algorithms such as AODV in performance criteria. Based on the analysis performed, modifications are proposed to the ARA algorithm. Finally, a performance evaluation of the original ARA and the modified ARA is carried out with respect to each other.

Keywords: MANET, Ant Colony, Routing, ARA**.**

Introduction

Mobile ad hoc networks (MANETs) are networks that are made up of a set of mobile devices. There are no designated routers, meaning that all nodes can serve both as end points of data communication and as intermediate relay points or routers. Ad hoc networks must also support communication between nodes that are only indirectly connected by a series of wireless hops through other nodes.

Nature of nodes and mobility of nodes are very important factors in performance of MANET. As devices in MANETs, have limited transmission range so in most cases they are not able to communicate directly with the destination device. Thus communication is relayed through intermediate nodes. Due to mobility of nodes in MANETs, nodes frequently come within the transmission range and go out of the range which interferes the routing of MANET. Thus to support the routing function, nodes frequently exchange data to become aware of the status of the network. MANETs are very powerful and are widely used in various real-world situations such as battle field scenarios, rescue operations and vehicular networks, where traditional network infrastructure is difficult or impossible to setup. Due to mobility of nodes the topology of the MANET changes continuously. Taking these points into consideration, some additional requirements are imposed on the Routing Algorithm. A MANET routing algorithm should not only be capable of finding the shortest path between the source and destination, but it should also be adaptive, in terms of – the changing state of the nodes, the changing load conditions of the network and the changing state of the environment.

MANET routing algorithms generally have three essential components: A route discovery mechanism, a route error correction mechanism, and a route maintenance mechanism. The route discovery mechanism finds initial routes between the source and destination nodes, the route maintenance mechanism maintains the routes discovered during the transmission of packets and the route error correction mechanism rebuilds routes when they fail.

MANET routing algorithms can be classified into three categories as proactive, reactive or hybrid [5]. Proactive algorithms try to maintain up-to-date routes between all pairs of nodes in the network at all times. The advantage is that routing information is always readily available when data need to be sent, while the main disadvantage is that the algorithm needs to keep track of all topology changes, which can become difficult when there are a lot of nodes or when they are very mobile. Examples of proactive algorithms are Destination-Sequence Distance-Vector routing (DSDV) and Optimized Link State Routing (OLSR) [9]. Reactive algorithms only maintain routing information that is strictly necessary: they set up routes on demand when a new communication session is started, or when a running communication session falls without route. This approach is generally more efficient, but can lead to higher delays as routing information is often not immediately available when needed. Examples of reactive routing algorithms include Dynamic Source Routing (DSR) [7] and Ad-hoc On-demand Distance-Vector routing (AODV) [6]. Finally, hybrid algorithms use both proactive and reactive elements, trying to

combine the best of both worlds. An example is the Sharp Hybrid Adaptive Routing Protocol (SHARP) [19].

As stated above, the traditional routing protocols face many problems due to the dynamic behavior and resource constraints in MANETs. To overcome this limitation, a routing protocol is required to have a self-organizing or an autonomous feature. An approach to achieve such feature is to use a biologically-inspired mechanism. In nature, many biological systems possess the ability to maintain their stable condition themselves regardless of the external influences or dynamic conditions. Ant colonies are complex biological systems that respond to changing conditions in nature by solving dynamic problems. Their ability of decentralized decision-making and their self-organized trail systems, have inspired computer scientists since 1990s, and consequently initiated a class of heuristic search algorithms, known as ant colony optimization (ACO) algorithms. These have proven to be very effective in solving combinatorial optimization problems, especially in the field of telecommunication. ACO is based on the ant foraging behavior, utilizing pheromone deposition as a means of evaluation for the travelled route. Rather than RREP and RREQ packets, 'forward' and 'backward ant' agents are sent across the routes, where the ant agents deposit pheromones at the nodes arrived. In the long term, this approach is used to determine the shortest path between the source and the destination.

In this paper we have focused on application of Ant Colony Optimization to the problem of MANETs. The Ant Algorithm mimics the behavior of ants in nature while they are searching for food. Particle swarm optimization is inspired by the behavior of flocks of birds as they fly in search of food. Bacterial foraging is yet another recent algorithm that simulates the behavior of bacteria searching for food. All these techniques are combinatorial in nature and when viewed in the perspective of optimization involve searching for the optimum solution in a given search space. It has been observed that, when these natural patterns are applied to complex engineering problems, they provide good solutions.

The rest of the paper is organized as follows: Section 2 presents descriptions of Ant Colony Optimization and its formulation. Section 3 explains Routing in MANET using Ant Colony System. Analysis of routing in MANET is illustrated in Section 4. In Section 5 we have proposed the modifications and Section 6 summarizes our contributions.

Ant Colony Optimization

Ant colony optimization (ACO) [2] is an optimization technique inspired by the exploratory behavior of ants while finding food. Ants start from their nest and find different paths to the food. In this context, the local information available to the ant is the path that it took to the destination. However a single ant is not aware of the complete topology of the environment. Under some appropriate conditions, they are able to select the shortest path among the few alternative paths connecting their nest to a food reservoir. While moving, ants deposit a volatile chemical substance called pheromone and, according to some probabilistic rule, preferentially move in the directions locally marked by higher pheromone intensity. Shorter paths between the colony's nest and a food source can be completed quicker by the ants, and will therefore be marked with higher pheromone intensity since the ants moving back and forth will deposit pheromone at a higher rate on these paths. According to a self-amplifying circular feedback mechanism, these paths will therefore attract more ants, which will in turn increase their pheromone level, until there is possibly convergence of the majority of the ants onto the shortest path. The volatility of pheromone determines trail evaporation and favors path exploration by decreasing the intensity of pheromone trails and, accordingly, the strength of the decision biases that have been built over time by the ants in the colony. The local intensity of the pheromone field, which is the overall result of the repeated and concurrent path sampling experiences of the ants, encodes a spatially distributed measure of goodness associated with each possible move. The pheromone acts significant stimuli since other ants are able to sense the pheromones deposited by each other, and they generally take the path of maximum pheromone concentration. This is how the ants progressively converge on a single optimum path between their nest and the food.

Shortest paths by Ants colonies

Many of the species of ants have a trail-following behavior when foraging [20]. While moving, individual ants deposit on the ground a volatile chemical substance called pheromone, forming in this way pheromone trails. Ants can smell pheromone and while choosing their way, they choose, in general, the paths marked by stronger pheromone concentrations. Also, they can be used by other ants to find the location of the food sources discovered by their nest mates.

The binary bridge experiment with same branch length

The binary bridge experiment [21] is shown in Figure 2.1. The nest of a colony of ants and a food source has been separated by a diamond-shaped double bridge in which each branch has the same length. Ants are then left free to move between the nest and the food source. The percentage of ants which choose one or the other of the two branches is observed over time. As a result it has been observed that after few minutes' ants tend to converge on a same path.

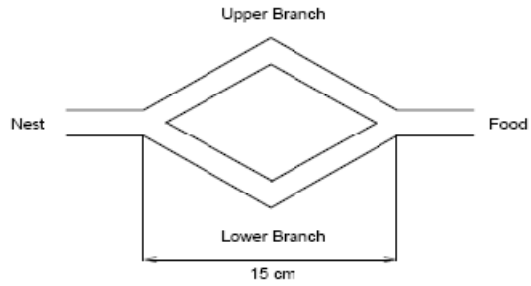


Figure 2.1: Binary bridge experiment.

The binary bridge experiment with different branch length

If the branches of the bridges are of different length, then due to the pheromone the majority of the ants in the colony choose the shortest path between the two available paths, as it is shown in fig 2.2 [22].

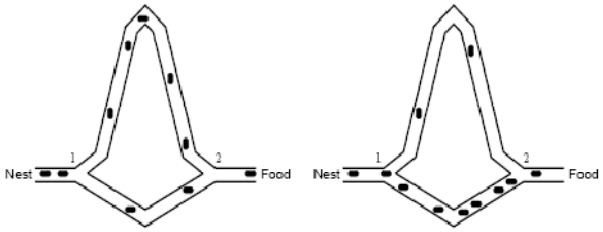


Figure 2.2: Experiment with a binary bridge with different branches length.

In this case, the first few ants arrive at the food source are those that traveled following the shortest branch. As the concentration of pheromone is higher on the shortest branch, this stimulates the same ants to choose again the shortest branch when moving backward to their nest. During the backward journey, additional pheromone is released on the shortest path. So, the concentration of pheromone on the shortest branch is higher than on the longest branch. Therefore, the choice of the shortest branch becomes more and more attractive for the subsequent ants at both the decision points.

Formulation of ACO

To formulate the ACO, the ants are modeled as artificial ants and the paths are represented as edges of a graph G. The formulation of ACO [2-4] as a combinatorial optimization problem can be done as follows:

$C = c_1, ..., c_n$ is a set of basic components. A subset S of components represents a solution of the problem; $F \subseteq 2^C$ is the subset of feasible solutions, thus a solution S is feasible if and only if $S \in F$. A cost function f is defined over the solution domain, $f: 2^C \rightarrow R$, the objective being to find a minimum cost feasible solution S*, i.e., to find S*: $S^* \in F$ and $f(S^*) \leq f(S), \forall S \in F$.

The search space S is defined as follows. A set of discrete variables, X_i, (i = 1, ..., n), with values $v_i^j \in D_i = \{v_i^1, ..., v_i^{|Di|}\}$, is given. Elements of S are full assignments, i.e., assignments in which each variable X_i has a value v_i^j assigned from its domain. The set of feasible solutions F is given by the elements of S that satisfy all the constraints.

In the Ant Colony Optimization, problems are usually modeled as a graph. Let G (V, E) be a connected graph with n = V nodes. Thus the components c_{ij} are represented by either the edges or the vertices of the graph. The objective of the problem is to find a shortest path between the source node V_s and destination V_d. Each edge of G maintains a value τ which denotes an artificial pheromone concentration value over that node which is modified whenever an ant transitions over it. To simulate the natural ant foraging process, three equations are used: *Pheromone evaporation, Pheromone increase, and Path selection*. If an ant currently at node i and transitions to node j:

$$\tau_{ij} = \tau_{ij} + \Delta\tau \qquad (2.1)$$

τ_{ij} is the artificial pheromone concentration over link j at i. The artificial Pheromones gradually evaporate over time, which is modeled by:

$$\tau_{ij} = (1 - \lambda) * \tau_{ij} \qquad (2.2)$$

Where $(1-\lambda)$ is called the pheromone decrease constant. At each node the ant has to make a decision about the next hop over which to travel. To simulate the exploratory behavior of ants the artificial ant makes a stochastic decision based on probabilities of the next hop. The probability of an ant transitioning to node j from node i at node d, where Ni represents a set of neighbors, is calculated by the equation:

$$p^d_{ij} = \{ (\tau^k_{ij} / \Sigma_{j \in N_i} \tau^k_{ij}) \text{ if } j \in N_i \} \text{and} \{ d_0, \text{otherwise.} \} \qquad (2.3)$$

Where k is called the route selection exponent and determines the sensitivity of the ant algorithm to pheromone changes.

The Ant Algorithm

In ACO, artificial ants build a solution to a combinatorial optimization problem by traversing a fully connected construction graph, defined as follows. First, each instantiated decision variable $X_i = v^j_i$ is called a solution component and denoted by c_{ij}. The solution is constructed by incrementally choosing the components from the Graph G(V, E). As mentioned before, the components can be associated with either the vertices or the edges of the graph.

Each component has a pheromone value associated with it τ_{ij}. The ants move through the graph and at each node probabilistically choosing the next component to add to the solution determined by the pheromone value of the components. The ant also deposits an amount of pheromone on the component depending on the quality of solution found. The ACO algorithm as described by [2-4] is shown in Algorithm 1.

Algorithm 1 ACO Meta heuristic

Require: parameters
1: while Iterations not complete do
2: Construct Solutions;
3: Update Pheromones;
4: Daemon Actions; {optional}
5: end while

Construct Solutions, chooses a subset of the set of components C. The solution begins with an empty partial solution $s^p = \phi$ and then at each construction step a feasible component is added to s^p. Daemon Actions are usually used to perform centralized actions that cannot be performed by a single ant and that may be problem specific. Update Pheromones serves two tasks: To increase the pheromone values of the components which are good, and to decrease the pheromone values of the components which are bad. The pheromone decrease is achieved through evaporation. Many different algorithms have been proposed with different pheromone update equations.

Routing in MANET using Ant Colony System

Routing protocols are classified into reactive [5, 6, 7], proactive [8, 9] and hybrid [10, 19] algorithms. Reactive protocol is also known as on-demand routing protocols, reactive routing protocols have been proposed with an aim to reduce the overhead caused by flooding of control packets.

This is achieved by maintaining routing information only for the active routes, rather than maintaining all the routes periodically. Therefore, route discovery is initiated 'on demand' when required. This protocol consists of two phases: (i) route discovery and (ii) route maintenance. In proactive routing protocols, each node attempts to maintain a consistent view of the network, which is done by periodically broadcasting its routing information to every other node within its neighborhood. They are classified into two, which are (i) link state routing and (ii) distance vector routing. In proactive algorithms each node broadcasts control information (called HELLO packets) about route information, that it has, to other nodes periodically, and the nodes which receive that information update their routing tables. Hybrid routing protocols were introduced with an aim to combine the advantages of proactive and reactive routing protocols. In hybrid protocols, the network is partitioned into zones. A proactive routing method is used within each zone while a reactive routing method is used for inter-zone communication. With this method, the overhead is reduced as the inefficiency of the proactive approach is limited only within the zone, while reactive routing enables efficient connectivity across zones. Ant Colony optimization, a bio-inspired meta-heuristic [7] has been applied to the MANET routing problem resulting in algorithms [18–21] to improve MANET performance. The technique is based on ant-agents modifying their environment to guide other ant-agents through the shortest path, a process called stigmergy. These algorithms have been shown to be very adaptive and responsive to changing environmental conditions in other domains and hence are a good fit for the MANET routing problem.

In this paper, we have described one of the popular Ant Colony algorithm i.e. Ant Routing Algorithm (ARA).

The ARA Algorithm

ARA is a purely reactive MANET routing algorithm. It does not use any HELLO packets [15] to explicitly find its neighbors.

Routing Mechanisms

When a packet arrives at a node, the node checks to see if routing information is available for destination d in its routing table. In ARA the route discovery is done either by the FANT (forward ant) flood technique [2] or FANT forward technique [12]. In the FANT flooding scheme, when a FANT arrives to any intermediate node, the FANT is flooded to all its neighbors. If found, it forwards the packet over that node, if not, it broadcasts a forward ant (FANT) to find a path to the destination. By introducing a maximum hop count on the FANT, flooding can be reduced. In the FANT forwarding scheme, when a FANT reaches an intermediate node, the node checks its routing table to see whether it has a route to the destination over any of its neighbors. If such a neighbor is found, the FANT is forwarded to only that neighbor; else, it is flooded to all its neighbors as in the flood scheme. In ARA, a route is indicated by a positive pheromone value in the node's pheromone table over any of its neighbors to the FANT destination. When the ant reaches the destination it is sent back along the path it came, as a backward ant. All the ants that reach the destination are sent back along their path. Nodes modify their routing table information when a backward ant is seen according to number of hops the ant has taken. When a route is found the packet is forwarded over the next hop stochastically according to equation 2.3. The results for the route discovery mechanism reveal an interesting trend. The FANT forwarding technique does better in situation of high mobility, that is, in situations having a lower pause time. However in cases of lower mobility, the FANT flood technique does better in the metrics of packet delivery ratio, throughput, delay and jitter. One more thing is being observed that in lower mobility situations, the FANT flood technique causes a lot of overhead, and increases the time required to find a route to the destination.

Route Maintenance

In ARA the route is maintained through the adjustment of pheromone values of the links present in the node routing tables. Whenever a particular link is selected as the next hop, the pheromone value of that link for the destination of that packet is incremented by a constant value in the routing table, according to equation 2.1. Pheromone values also made to constantly decrease according to equation 2.2. The authors of [12] studied and classified the various pheromone update functions used in ant algorithms.

Route Error Correction

In order to ensure delivery of the packet, the algorithm may need to correct the link if it fails due to mobility of nodes. In general, two mechanisms are widely used for route error correction: local route repair [12], and route error back-propagation [11]. In local route repair, if a link at a particular node fails, the algorithm buffers the packet and sends out a new FANT to discover a route to the destination. Once a route is found, the packet is forwarded over that route. In the route error back-propagation mechanism, if a link error occurs at a node and another route to the destination does not exist at that node, a ROUTE-ERROR packet is sent to the previous node in the forwarding chain. This is repeated until the ROUTE-ERROR packet reaches the source and then a route discovery process is initiated. The route correction mechanism used in this case is the error back-propagation algorithm and for route maintenance the discrete pheromone decay equation is used.

Proposed modifications to ARA

ARA and AODV are compared by the author in [18] and ARA is found better than AODV. Since ARA is a reactive protocol, that is why it is used in such situations where mobility of nodes are higher. In this section, we have proposed the modifications to the algorithm by which the potential of ARA will increase in high mobility scenarios. Pheromone updates play a critical role in the performance of the ant algorithm. In ARA algorithm, initial pheromone value is computed by number hops during the route discovery. This method may not be suitable when nodes are mobile. Pheromone equations are classified in different categories. Two of them are the Classic pheromone filter, where route quality is not taken into consideration, for example the original ARA pheromone equation, and the Gamma pheromone filter, which takes time and route quality into consideration.

Simulation Parameters

Algorithm is implemented on Qualnet version 4.0 [17] and simulations for each "choice" are run on the Qualnet simulator. The simulations are conducted on an area of 1000m x 1000m. Node mobility is restricted to a maximum speed of 10 m/s and according to the random waypoint [9] mobility model. 802.11b is used as the underlying MAC layer protocol with a propagation limit of -111dB. Tcp-lite is used as the application layer protocol. Simulation time for each instance of an experiment is 300s. For each experiment the sample size is 5. Constant bit rate connections, configured between the nodes are 18. The random seed for the simulation is initially set to 300. Through each experiment various performance metrics of the algorithm are measured in terms of Packet Delivery Ratio, End-to-End delay, Jitter, and Throughput at receiver node. Through experiential observations the maximum hops for the FANT is set to 10. Other parameters include a pheromone decrease constant of 0.4, a pheromone increase constant of 0.6, a decrease interval of 5s, and a route select exponent of 3. Measurements are taken by varying pause time, which is indicative of the mobility of the network. Pause times are taken to be 0s, 30s, 60s, 90s, 120s, and 150s. Lower pause times indicate a greater mobility of the nodes in the network.

The Time Metric

Taking path quality into consideration we develop a type of Gamma Pheromone filter for ARA to update pheromone values as Gamma Pheromone filters show a better performance over the classic pheromone filters [15]. The modified pheromone update equation sets the initial value of the pheromone as:

$$\tau^{d}_{ij} = 2 / (hops + t) \quad (5.1)$$

The pheromone update is done as per Equations (2.1) and (2.2). Where τ^{d}_{ij} denotes the pheromone concentration over link (i, j) for a destination d. t denotes the time interval between the sending of a forward ant and the receipt of the backward ant, and hops is the total number of hops made by the ant. The inclusion of time in the equation creates a pheromone gradient from source to the destination point depending on the time it takes for the backward ant to reach the node that forwarded it. In the case of only FANT hops being taken into consideration, many paths with a similar gradient are formed; however the time metric creates a marked difference in the path gradient and thus the packet would be randomly forwarded over the path with the greatest pheromone gradient. This metric is thus expected to produce better results than if only number of hops is considered.

Pause Time	Message Delivery Ratio	
	Orig. ARA	Prop. Algo
0	0.972	0.985
30	0.981	0.988
60	0.976	0.980
90	0.980	0.985
120	0.957	0.980
150	0.970	0.990

Figure 5.1: Message Delivery Ratio.

Pheromone decay process

We have proposed that the pheromone decay should be discrete process rather than a continuous one. As the pheromone decreases asynchronously after a particular time interval, so the discrete process shall allow more pheromone to be available on the routes so that routes live longer.

Pause Time	Throughput	
	Orig. ARA	Prop. Algo
0	2.015	2.022
30	2.012	2.028
60	2.009	2.014
90	2.013	2.018
120	1.943	2.006
150	2.006	2.028

Figure 5.2: Throughput.

Route Selection Exponent

In original ARA algorithm, the route selection is done by equation 2.3 and uses a route select exponent of k = 1. We have proposed that the route select exponent to the ARA algorithm should be k = 4. This increases the sensitivity of the algorithm to changes in pheromone values, making it more adaptive in nature.

Routing mechanisms

In original ARA algorithm, flooding technique was used. But we have simulated the proposed algorithm using forwarding technique. So, if a route exists from a node to the destination, the FANT is forwarded over that route instead of flooding. This will reduces the overhead during the route discovery process. Maximum number of hops is increased. The maximum number of hops is set through experiential observation of a reasonable amount of time it takes for the FANT to reach the destination from the source. Due to this delay may increase but performance tested to be increased.

Pause Time	Message Delay	
	Orig. ARA	Prop. Algo
0	5.4	3.2
30	3.5	3.8
60	4.2	4.5
90	2.5	2.9
120	4.7	4.5
150	4.2	5.2

Figure 5.3: Message Delay.

Simulation Results and Discussion

The performance is measured in terms of Packet Delivery Ratio, Throughput, End-to-End delay and Jitter for various values of pause time respectively.

The observations show the efficiency of the modifications to the ARA algorithm. For the delivery ratio, throughput, and jitter metrics, the proposed ARA algorithm performs better than the original ARA algorithm. As shown in figure 5.1, 5.2,

5.3 and 5.4 all the performance metrics are enhancing except the delay.

Pause Time	Jitter	
	Orig. ARA	Prop. Algo
0	3.8	3.0
30	3.6	2.8
60	3.9	3.2
90	3.0	2.8
120	5.1	3.1
150	3.8	2.6

Figure 5.4: Jitter.

Proposed ARA shows an advantage over the original ARA algorithm due to the inclusion of the time metric and this advantage is clearly shown in the case of delay. In our proposed algorithm, the route selection exponent value k=4 makes the ant route selection equations more sensitive to changes in pheromone values. The changes in pheromone values indirectly indicate the topology of the MANET, and are helpful in the selection of ant routes. As we have used forwarding scheme during route discovery, it will help in finding the routes faster.

Conclusion

The foraging behaviour of the ant colonies has been extensively investigated for more than 50 years and has been explored the remarkable trail systems achieved through robust, decentralized communication. Their collective intelligence has been shown to be one of the best examples of self-organization. Early research revealed their ability to detect shortest paths in static environments, whereas recent research discovered fundamental mechanisms in the foraging systems for the dynamic systems as well.

The main objective of this paper was to develop an algorithm which works well under certain constrained conditions. The proposed routing algorithm design in this paper is a step forward towards that goal. Through the simulations routing mechanism is being analyzed and it has been found that the proposed technique is working well in high mobility scenarios. The ARA algorithm is modified and it is observed, through various simulation based experiments, that modified ARA performed better in comparison to the original ARA in terms of varying mobility. Ant Colony algorithms are very adaptive to the changing environments. However, their performance must be improved further and they must be able to solve the problems of heterogeneous networks.

References

[1] Al Huda Amri and et. al. Scalability of manet routing protocols for heterogeneous and homogenous networks. Computers and Electrical Engineering, 2008.

[2] Dorigo M. and G. Di Caro. Ant colony optimization: a new meta-heuristic. In Proceedings of the Congress on Evolutionary Computation, 1999.

[3] V. Maniezzo. Exact and approximate nondeterministic tree- search procedures for the quadratic assignment problem. INFORMS Journal of Computing, 11(4):358–369, 1999.

[4] Dorigo M., G. Di Caro, and L. M. Gambardella. Ant algorithms for discrete optimization. Artificial Life, 5(2):137–172, 1999.

[5] E.M. Royer and C.K. Toh. A review of current routing protocols for ad hoc mobile wireless networks. In IEEE Personal Communications, volume 6, April 1999.

[6] 6] C. Perkins. Ad hoc on-demand distance vector routing. Internet- Draft, draft-ietf-manet-aodv-00.txt, November 1997.

[7] DSR: The Dynamic Source Routing protocol for multi-hop wireless ad hoc networks, chapter 5, pages 139–172. Addison- Wesley, 2001.

[8] C.E. Perkins and P. Bhagwat. Highly dynamic destination- sequenced distance vector (dsdv) for mobile computers proc. of the sigcomm 1994 conference on communications architectures, protocols and applications. pages 234–244, Aug 1994.

[9] T. Clausen and P. Jacquet. Optimized link state routing protocol (olsr). RFC 3626: Optimized link state routing protocol (OLSR), Oct 2003.

[10] Z.J. Haas and M.R. Pearlman. The zone routing protocol (zrp) for ad-hoc networks. IETF MANET working group, Internet Draft, June 1999.

[11] M. G¨unes, Sorges U, and I. Bouazizi. Ara - the ant-colony based routing algorithm for manets. In proceedings of the 2002 ICPP Workshop on Ad Hoc Networks (IWAHN 2002), pages 79–85. IEEE Computer Society Press, August 2002.

[12] G.A. Di Caro, F. Ducatelle, and Gambardella L.M. Anthocnet: An adaptive nature-inspired algorithm for routing in mobile ad hoc networks. European Transactions on Telecommunications, Special Issue on Self-organization in Mobile Networking, 16(5), October 2005.

[13] M.Roth and S.Wicker. Termite: A swarm intelligence routing algorithm for mobile wireless ad-hoc networks, stigmergic imization. Studies in Computational Intelligence, 34:155–185, 2006.

[14] O. Hussein and T. Saadawi. Ant routing algorithm for mobile ad-hoc networks. In The International Performance Computing, and Communications Conference (IPCCC), Phonix, Arizona, April 03.

[15] M. Roth and S. Wicker. Asymptotic pheromone behavior in swarm intelligent manets. In Proceedings of the Conference on Mobile and Wireless Communication Networks (MWCN 2004), pages 335–336, 2004.

[16] J. Broch, D. A. Maltz, D. B. Johnson, Y.C. Hu, and J. Jetcheva. A performance comparison of multihop wireless ad hoc network routing protocols. In Proceedings of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'98), pages 85–97, 1998.

[17] The qualnet 4.0 programming manual.
[18] M. Gunes, M. Kahmer, , and I. Bouazizi. Ant-routing-algorithm(ara) for mobile multi-hop ad-hoc networks - new features and results. In Proceedings of the 2nd Mediterranean Workshop on Ad-Hoc Networks (Med-Hoc-Net'2003), Mahdia, Tunisia, 25-27, June 2003.
[19] V. Ramasubramanian, Z. J. Haas, and E. G. Sirer. Sharp: A hybrid adaptive routing protocol for mobile ad hoc networks. In Proceedings of The Fourth ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc), 2003.
[20] B. H¨ olldobler and E.O.Wilson. The Ants. Springer-Verlag, Berlin, Germany, 1990.
[21] J.-L. Deneubourg, S. Aron, S. Goss, and J.-M. Pasteels. The self-organizing exploratory pattern of the argentine ant. Journal of Insect Behavior, 3:159–168, 1990.
[22] S. Goss, S. Aron, J. L. Deneubourg, and J. M. Pasteels. Self- organized shortcuts in the Argentine ant. Naturwissenschaften, 76:579–581, 1989.

Energy Efficient Routing Protocol for Dual Transmission in WHSNs

[1]Kusum Lata, [2]Ashutosh Dixit and [3]Soni Chaurasia

[1&2]Department of Computer Science &Engineering, YMCA University of Science & Technology, Faridabad, Haryana, India
E-mail: ranapoo@gmail.com, dixit_ashutosh@rediffmail.com
[3]Department of Computer Science &Engineering, Lingyas University, Faridabad, Haryana, India
E-mail: 28soniya@gmail.com

Abstract
Wireless Sensor Networks (WSNs) has known beforehand big changes in data gathering, processing and dispersing for monitoring specific applications such as emergency services, disaster management, and military applications etc. Wireless sensor network consists of a group of sensor nodes that are distributed in wide area which are interconnected without wires.

In this paper, an Energy Efficient Routing protocol for Dual Transmission for wireless heterogeneous sensor network with multiple sensing unit and two transmission unit has been proposed. The proposed algorithm has been designed for WHSN in which one sensor uses real time service for the transmission of data of one sensing unit while other sensing unit uses a best effort service for the transmission of data of the other sensing unit in a node. This protocol overcomes the problem related to event capturing and real time data transfer with high energy consumed. The Protocol was simulated using OMNET++4.0 simulator on Linux platform on Pentium IV machine. ERDT was compared with EEFS and REFS Protocol. Simulation results show that ERDT outperforms EEFS and REFS in terms of network life, energy efficiency & network life time.

Keywords: WSN; WHSN; Data Redundancy; Best Effort Service; Real Time Service.

Introduction (Heading 1)

A Wireless Sensor Network (WSN) consists of number of autonomous sensors which are widely distributed in wide area to monitor physical or environmental conditions, such as temperature, sound, motion, vibration, pressure or pollutants and to cooperatively pass their data through the network to a main location. WSN has many application such as Health monitoring, military application, etc.[2].Sensor nodes are having capability of gathering information, processing and communication. Battery or power supply is required to complete sensor node. So algorithm has been designed for WHSN with multiple sensing units and two transmitters in sensor node called Energy Efficient Routing Protocol for Dual Transmission.

In the next section is describes the issues involves in designing routing protocols, the third section describes the related work, fourth section describes about the proposed algorithm and last section describes about the conclusion.

Design issue involved in routing protocol

Some of the factors which influenced the design of routing protocols are discussed below:-

Node Deployment

Node deployment can be random, deterministic or self - organizing. For deterministic deployed networks the routes are pre-determined, however for random deployed networks and self-organizing networks route designation have been a challenging subject.

Energy Consideration

Since the life-time of the WSN depends on energy resources and their consumption by sensors, the energy consideration has a great influence on route design. The power consumed during transmission is the greatest portion of energy consumption of any node.

Data Delivery Model

Data delivery model depends on the application and can be continuous, event-driven, query-driven, or hybrid [3, 4]. In continuous model of delivery, each sensor sends the data periodically. In event-driven and query driven data delivery models, the transmission is triggered by an event or a query generated by the sink. Hybrid model is a combination of continuous, event driven and query-driven data delivery models.

Data Aggregation

Since the sensors are densely deployed by definition, the data gathered from each node are correlated. Therefore data aggregation or in other words data fusion decreases the size of the data transmitted.

Fault Tolerance

WSNs are prone to failures; some of the nodes may fail or be blocked by physical interference, physical damage, or lack of power. The routing protocol has to be dynamic; failures of specific nodes should not affect network operation.

Scalability
WSNs may consist of hundreds, thousands or more nodes. Any protocol including routing protocols should manage this huge number of nodes.

Network Dynamics
Most of the proposed networks are considered to be stationary; however for some application areas WSNs in which some or all nodes are mobile are required. Routing protocols for such networks must cater for mobility requirements.

Quality of Services
Some of applications require QoS as especially there exist some time-critical applications. The relevance of the data expires within some period. For such applications the routing protocols should be designed according to the requirements. However, the general trend is to attribute more importance to energy awareness than QoS requirements.

Related work
There are many research work done which is based on minimizing of energy consumption in a sensor network. The sensor network lifetime is directly related to the energy consumption. If the energy consumed by sensor node in sensor network is low then it means that the network lifetime is increases. Heterogeneous wireless sensor network (HWSN) consists of sensor nodes of different capability like one sensor node with high power and second is with low power. It may possible that most of data send by sensor node with high power so energy consumption is more here.Therefore lot of work is also done to improve battery life with the help of data aggregation technique. Here we introduce a new technique to improve the network lifetime and remove redundancy of data packet. Wireless sensor networks using Divisible Load Theory (DLT) [6] is used for scheduling work load. This technique is used because battery power of sensors is limited so it is desirable that it complete a task as quickly. But it uses a single sensing unit in a sensor node which consumes more energy in scheduling as compared to that of multiple sensing units in a sensor node. MSUS (multiple sensing unit scheduling) [7] is used to minimize the event-misses and energy in wireless sensor network. This algorithm provides a best power state based on the priority and timing requirements. On target coverage in wireless heterogeneous sensor networks with multiple sensing units [8], introduces the concept of the target coverage problem in wireless heterogeneous sensor networks (WHSNs) with multiple sensing units. This approach uses proposed two heuristic but distributed schemes, REFS and EEFS. Both schemes increases the network lifetime but there is a problem that if sensing attribute (parameter) is increase then decrease the network lifetime. So that we have proposed routing protocols in WHSN with multiple sensing unit and two transmitters in our proposed algorithm.

Proposed Algorithm
The algorithm has been designed for dual radio transmission called Energy efficient Routing protocol for dual transmission for providing QoS in Wireless Heterogeneous sensor networks. The Energy efficient Routing protocol for dual transmission has been designed for sensor nodes having multiple sensing unit and two transmitters. The proposed routing protocol named as Energy efficient Routing protocol for dual transmission. The proposed protocol is used to reduce the energy consumption as heterogeneous sensors consume more energy. The proposed algorithm has been designed for WHSN in which one sensor uses real time service for the transmission of data of one sensing unit while other sensing unit uses a best effort service for the transmission of data of the other sensing unit in a node. The data of one sensing unit (real time service) uses a high power radio for transmission while for best effort transmission (data of other sensing unit) a low power radio is used. The sensor nodes coordinate the transmission of data among themselves timely and achieve load balancing among the sensor network. This protocol overcomes the problem related to event capturing and real time data transfer with high energy consumed as compared to the HWSN. In Heterogeneous wireless sensor network (HWSN) consists of many sensor nodes with different ability, such as different computing power and sensing range. WSN consists of sensor nodes with different abilities, such as various sensor types and communication /sensing range, thus provides more flexibility in deployment. For example, Let us consider WSN in which nodes are equipped with different kinds of sensors to give different kind of sensing services. Besides WSN have also two types of sensor nodes such as one sensor nodes with high radio power have higher process throughput and higher communication/sensing range and the one sensor nodes with low radio power have limited process throughput and communication/sensing abilities.

This scheme is based on the Tree based approach where each sensor is detecting the event and sends aggregated data to its aggregated node or parent node. The aggregated node directs data packets towards the sink. This protocol assumes the existence of sensors nodes (S1, S2, S3 and S4) and two transmitters (T1 and T2) for sensor nodes. All the sensor nodes are assumed to be same and have four different types of sensing unit and two transmitting unit. In one sensor node, the sensing unit say S1 senses the event and sends data to the sensor node through low power radio (Tx1) and this time all other sensing unit are in sleep mode and same procedure for all other sensing unit. And if at the same time sensing unit say S3 also senses the event i-e sudden change in environment and sends data to the sensor node through high power radio (Tx2), at this time, All other sensing units continue to send the data through low power radio (Tx1). The sensor nodes should be able to transfer data to the sink in real time. The ERDT protocol is mainly focused on the data delivery aspect of real time data and data is aggregated at every sensor node that may be child node or parent node. The advantages of ERDT protocol is real time event capturing and real time data transfer and avoids redundant data transmission.

Let us consider that one sensing unit senses the room temperature, at the same time the other sensing unit senses the humidity data. If the temperature data is most critical data then it will first send temperature data when event is detected. The temperature data is based on real time service for the transmission of data. Other sensor senses the humidity data and uses the best effort service for the transmission of data.

The proposed protocol considers packet deadline, energy

of the forwarding nodes and congestion at intermediate nodes to deliver real-time traffic and best effort traffic. It also reduces data redundancy (duplicate data packets) at the source node to increases reliability using data aggregation method.

The main important aim of data aggregation is to collect the most critical data from the different sensor nodes and transmit data to the sink in an energy efficient manner with minimum data latency. Data latency is important in many applications such as environment monitoring, where the freshness of data is also an important factor. Data aggregation has many trade-offs among all parameters like data latency, data accuracy and energy efficiency. Energy-efficient data algorithms are that satisfy the latency of a specific application are imperative for the long life of the network. Energy efficiency of sensor networks depends on a variety of factors such as network architecture, data aggregation mechanism and routing algorithm. The routing algorithm is a mechanism that delivers aggregated data to the sink node through aggregating tree. The mechanism has each aggregating node send at least one message to the sink node and make sure redundant data is not sent. The network architecture needs to support the structure which can adapt to network changes in which a node joins and leaves to control static topology in wireless sensor networks.

System Architecture

The system model is working as Tree Based Routing Protocol as shown in figure 1.1 Let us consider wireless sensor network in which all sensor nodes are distributed uniformly in different regions such as Region1, Region2, Region3 as shown in figure 1 and the entire nodes in the network form a tree with different level. In Regions, the sensor node senses the data and aggregated node is a node that sends aggregated data to next Region. In figure 2 is shown the sensor node architecture, all sensor nodes know its location and sends data to its parent node or aggregated node. Let us assume all sensor nodes know predefine range. All nodes are communicated to each other within its range. The child node sends the data to parent node or aggregated node which is most nearest to the other region. Every node maintains a routing table to exchange Beacon messages (HELLO) and forward a data packet based on the routing table

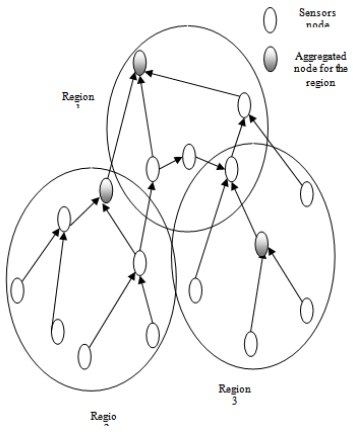


Figure 1: Data clustering tree for data aggregation.

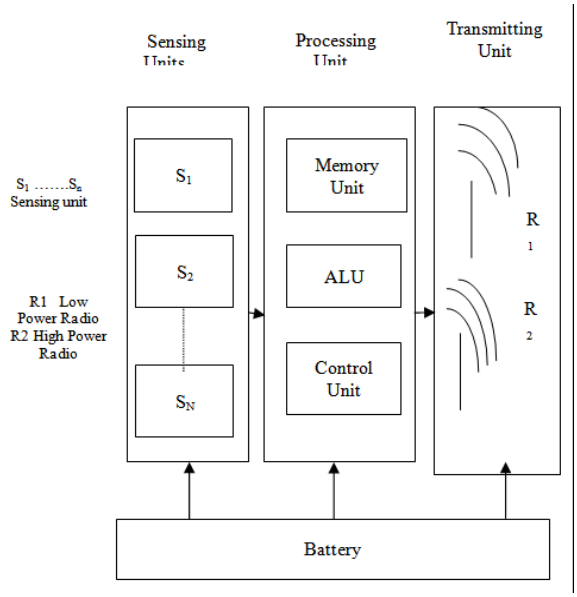


Figure 2: Sensor Node Architecture.

Working Method

In WHSN, a sensor node has multiple sensors to sense the different parameter such as light, humidity and temperature etc. In proposed algorithm, it is based on multiple sensing units and two transreceiver. The both two trans-receiver is working as best effort service and real time service. The low power transmitter sends the data with best effort service. And the low power radio means consume less energy as compared to high power radio. The low power radio range is less as compared to the high power radio. The high power transmitter is based on real time service. There are S_1, S_2, …., S_n sensing unit in one sensor node. When Sensor S_1 is sensing the environment at this time all the sensing unit is in sleep mode. And sensors S_1 send the data to low power radio Tx1 and after that S_2sense the environment and all the other is in sleep mode. And it sends the data to low power radio $Tx_1$and so on. But if the sudden change in environment is observed, at this time the randomly wake up the sensing unit and send its data to the high power radio Tx_2. At this time the low power transmitter is not in off mode.

Event detection

We assume the following parameters are known:-

GPS location is known to all sensor nodes.

All sensor nodes are aware of Sensing unit Ids.

The Area is previously allocated for all sensor nodes where all sensor nodes are distributed.

Two Radio trans-receivers (High power and Low power)has uniform transmission power throughout the network.

Sensor node send the HELLO message to the base station in the form of

<Node_Id, S_Id, cor, En, D, T>

Where

Node_Id is unique ID for node identification.

S_Id is unique ID for sensing unit identification

cor denote the location in form of(x, y)are the coordinates

En denotes the energy of the node.

T is timestamp of the packet.

D is coming packet that wants to send.

Scheduling of Data Transmission

When Sensor S1 is wake up, all other sensor is in sleep mode. S1is detecting the event and sent through Tx1. If detected event is same as the previous event, now this time **S1discard the data.** And now S2 is wake up and its work same as S1and so on. If the data is not same, at this situation, we have given a time slot to the entire sensing unit.

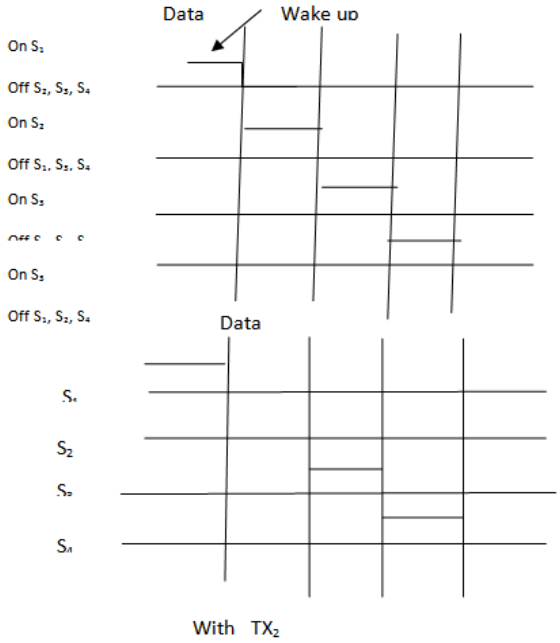


Figure 3: Data Timing Diagram.

ERDT Algorithm

The proposed algorithm is based on real time service and best effort service. The sensing unit S1 detects the event and check the Data, if new data (NR) is not equal to previous data (PR) then the S1 transmits data through low power radio to the aggregated node or parent node or cluster head. But if the new data (NR) is same as previous data (PR) then discard this data and turn off the sensing unit S1. In parallel the sensing unit S2 senses the event, check the new data(NR) is not equal to previous data (PR). Then S2 is sending the data to the aggregated node or parent node or cluster head through the high power radio. The same method is applied for all remaining S3, S4 sensing units. The transmission depends on α, which is application specific value and can be adjusted according to the application to ensure a minimum or maximum rate guaranteed for different sensor readings. If α is equal to zero then the S2 is turned off, the value of α is important for transmission of best effort service data via a high power radio and sensing unit S1 sends data thought high power radio (Tx2) to the cluster head or aggregated node or parent node. The cluster head or aggregated node or parent node is sent the aggregated data to the sink. The cluster head or aggregated node or parent node uses the average function for aggregation. The cluster head or aggregated node or parent node sends the aggregated data to the base station. The terminologies are used in ERDT algorithm are shown in table 1 and ERDT algorithm is shown in table 2.

Table 1: Terminology used in ERDT Algorithm.

Tx1 is low power radio
Tx2 is high power radio
0 represent low priority data
1 represents high priority data
N represents Number of sensing unit in sensor node
T represents time
S is sensing unit α represent variation in time

Table 2: ERDT algorithm.

ERDT Algorithm

```
for (i=1 to N)
Si is sense the environment.
Si store the data // store the data in queue
If (Si==ON && Tx1==ON && T=α && priority==0)
  for (j = α; j>=0;j--)
Send the data through Tx1 to Base station (BS) or parent node: DataTx1
Pr [ ] =DataTx1
sense new data
 Else if (Pr [ ] == Nr []) then
     Discard the New Data;
     Si is in sleep mode and Si+1is wake up
Else if (Priority =1, Tx2==ON, Sj==ON)
 Send the high priority data through theTx2to the base station or parent node: DataTx2
 End if
End For
Aggregating node or parent node is receiving data from the child node.
If (data receiving from child node)
  for (i=1 to i= no of child nodes)
  Check the data
Else if (DataSi is from Si)
   Store data into queue
   Apply aggregation function on the new data which is store in a queue
   Transmit this new data to the base station i.e. DataAggTx1
Else if (DataSi is from Sj)
   Store data into queue
   Apply aggregation function on the new data which is store in a queue
   Transmit this new data to the base station i.e. DataAggTx2
End else if
End else if
```

Algorithm Description

1. Initially Sensing unit Si (for i=1 to N, where N is no. of sensing unit in one sensor node) sense the environment and store sensed data.
2. It check that if Siis wake up i-e ON and transceiver with low power is ON and low priority data is there and set the time Tis equal to α. (Time slots are assigned to all sensing unit)
3. Send the data through Tx1 i-e low power transceiver to base station and marked as Previous Data. And sensing unit Si sense data from environment and marked as New Data.
4. Repeat step 3 until time is not expired
5. If Previous Data and New Data are same then discard the new data and transfer the control to other sensing unit for sending data. Otherwise send the high priority data through Tx2 i-e high power transceiver.
6. Repeat step 1, 2, 3, 4, 5 until no. of sensing unit is not covered.
7. Every node sends its child list to its father node. According to the child list, the parent node sends a TDMA slots to its child node. In its schedule the child node can send its data to the parent node.
8. If the child node has the data, then it will forward its aggregated data to its parent node in its time slot (TDMA slot) otherwise it will send a nack data to its parent node. The parent node will aggregate its data with its children data and send aggregated data to its parent node. Then, finally the node nearer to the base station sends its data to the base station.
9. Data is come from child node DataTx1& DataTx2 to the parent node. Parent node applies the aggregation on the data. Send the data to the base station: DataAggTx1 & DataAggTx2

Simulation Result and Disscussion

In this Section, we are having evaluation of the performance of our proposed algorithm we have compared ERDT algorithm with REFS and EEFS in our simulation. The simulator used for simulation is OMNET++ 4.0. In our simulation, we have considered following parameter:-

1. The number of nodes chosen was between 20 to 150 for simulation purpose.
2. All nodes are placed randomly placed.
3. Each sensor's location is represented by Cartesian coordinates.
4. The base station was located at (Xmax/2, Ymax/2) for simulation.

The graph plotted between Numbers of packets generated vs. Number of nodes for ERDT algorithm as shown in figure 5.1. The sensors S1, S2, S3, S4 send data packet via low power transmitter (best effort service) within given time but if any sensing unit sense sudden change in environment than it is wake up and send those data packet via high power transmitter (real time service).Figure 5.1 shows that as the numbers of nodes increase the number of packets generated also increase.

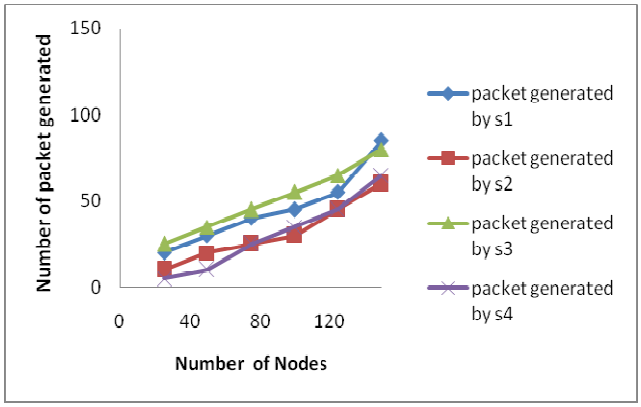


Figure 4: Number Of Packet Generated Vs. Number Of Nodes.

The graph is plotted between Average energy consumed by the nodes vs. Number of nodes for ERDT protocol (our proposed protocol) and REFS protocols is shown in figure 5.2. The energy consumption of ERDT protocol is less compare to REFS protocol. This is because of the fact that REFS consume more energy for sending redundant data to the sink. Our proposed protocol consumes less energy as compare to REFS. This is because of fact that it does not send the redundant or duplicate data. Our proposed protocol is more energy efficient as compared to the REFS protocol.

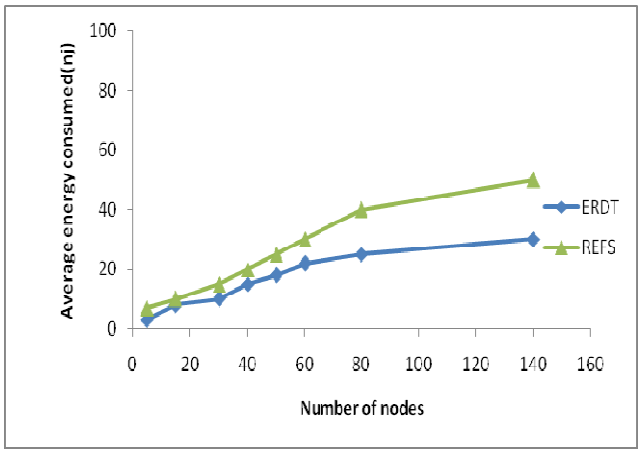


Figure 5: Average energy consumed Vs. Number of nodes.

The graph plotted between network life times vs. numbers of nodes for ERDT and EEFS protocols is shown in figure 5.4.the simulation results show that the performance of ERDT protocol is better than EEFS protocols. The simulation result shows that as number of nodes increase the network life time also increases. The performance of proposed protocol is better when the number of nodes increase because of the fact that as the number of the nodes increase, number of control messages decreases hence the performance of ERDT protocol better in terms of network lifetime.

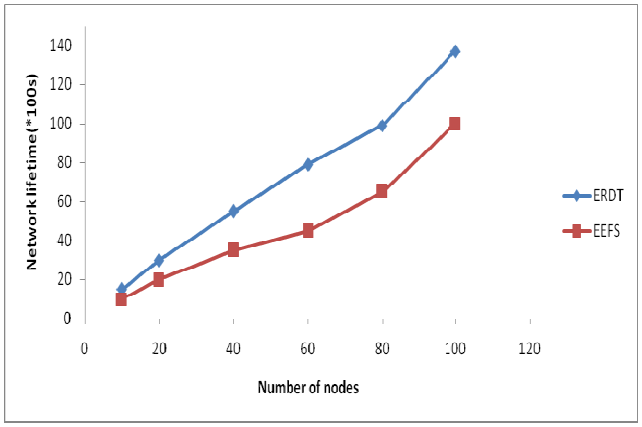


Figure 6: Network life time Vs. Number of nodes.

Conclusions

Wireless heterogeneous sensor networks with multiple sensing have enhanced performance over Wireless Heterogeneous sensor network because of reduces duplicity of data at the nodes. In HWSNs, Every sensor nodes are having one sensing unit sensing different attributes. HWSN uses more sensor nodes for sensing all attributes and here are two types of sensor nodes are one is low power sensor node and other is high power sensor node and it may be possible that most of data send by high power sensor node so thereby increasing the energy consumption. Our proposed protocol (ERDT) is an energy efficient protocol for wireless heterogeneous sensor networks with multiple sensing units and two transmission units. Our Proposed protocol has been designed to be more energy efficient as compared to the EEFS. Simulation results show that our proposed protocol for wireless heterogeneous Sensor Networks with multiple sensing units and two transmission units. on a single node consume less amount of energy as compared to the REFS. The simulation results also show that, in our proposed protocol the most critical packets for an application are sent to the parent node after removal of the duplicate packet. The removal of redundancy on the node and aggregation at the aggregated node or child node or parent node increases the life time of the network. Our proposed protocol (ERDT) minimizes the energy consumption in sensor network and maximizes the network lifetime.

Simulation results also show that the ERDT protocol for Wireless heterogeneous sensor nodes with multiple sensing units and two transmitters performs better than REFS. The ERDT protocol is more energy efficient than REFS.

References

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks, " IEEE Communications Magazine, Volume: 40 Issue: 8, August 2002, pp.102-114.

[2] Th. Arampatzis, J. Lygeros, and S. Manesis, “A Survey of Applications of Wireless Sensors and Wireless Sensor Networks, ” Proceedings of the 13th Mediterrean Conference on Control and Automation, pp.719-724, June 2005.

[3] J. N. Al-Karaki and A. E. Kamal, “Routing Techniques in Wireless Sensor Network: A Survey, ” IEEE Wireless Communications, pp.1-37, 2004.t

[4] K. Akkaya, and M. Younis, “A Survey on Routing Protocols for Wireless Sensor Networks, ” In: Ad Hoc Networks, vol. 3, pp.325-349, 2005.

[5] C. K. Liang, Y. J. Huang, and J. Da Lin, “An Energy Efficient Routing Scheme in Wireless Sensor Networks, ” 22nd International Conference on Advanced Information Networking and Applications-workshops (AINAW), pp.916-921, March 2008.

[6] H. Kang, P. Guan, X. Liu, and X.Li, “Utility-Based Divisible Sensing Task Scheduling in Wireless Sensor Networks, ” Proceedings of the International Conference on Wireless Networks(ICWN), pp.342-348, 2007.

[7] R. Poornachandran, H. Ahmad, and H. C¸ am, “Energy-Efficient Task Scheduling for Wireless Sensor Nodes with Multiple Sensing Units”, Proceedings of the International Workshop on Strategies for. Energy Efficiency in Ad-hoc and Sensor Network (IWSEEASN), pp.409-414, April 2005.

[8] TK. Ping Shih, H, ChangChen, C. MinChou, and B. JunLiu, “On target Coverage in Wireless Heterogeneous Sensor Networks with Multiple Sensing units”, Journal of Network and Computer Applications, vol. 32, pp.866–877, 2009.

[9] K. Akkaya, and M. Younis, “A Survey on Routing Protocols for Wireless Sensor Networks, ” In: AdHoc Networks, vol. 3, pp.325-349, 2005.

[10] Vivek Katiyar, Narottam Chand, Surender Soni “A Survey on Clustering Algorithms for Heterogeneous Wireless Sensor Networks”, Int. J. Advanced Networking and Applications, Volume: 02, Issue: 04, pp: 745-754 (2011).

A Survey on Various Propagation Model for Wireless Communication

[1]Pooja Prajesh and [2]R.K. Singh

[1]Asst. Professor, GRDIMT, Dehradun, India
[2]Professor, KEC, Dhawarahat, India

Abstract
Signal Propagation is used for wired or wireless communication. It is depend upon terrain, frequency of operation, height of mobile, base station and other dynamic factor. Propagation models predict the mean signal strength for an arbitrary transmitter-receiver (T-R) separation distance[5]. In this paper, Empirical propagation models such as Okumura, Hata, and Lee model has been surveyed exhaustively.

Keywords: Path Loss, Okumura model, Hata model and Lee model

Introduction

In Wireless communication signal is transmitted by transmitting antenna and received by receiving antenna, any distortion in signal strength at receiver is known as path loss. Propagation model are useful for predicting the signal attenuation or path loss between the transmitter and receiver. This path loss information may be used as a controlling factor for wireless communication system performance to achieve the perfect network planning [1].

The Propagation model is generally of two types: Empirical (statistical) models and Physical (Deterministic) models. In this paper empirical models are considered. Statistical methods (also called stochastic or empirical) are based on fitting curves with analytical expressions that recreate a set of measured data. Among the most commonly used such methods are *Okumura Model*, *Hata Model,* and *Lee's Model* [3]. The Empirical or statistical models are suitable for both macro cell and micro cell.

Path Loss Models

Okumura Model

The Okumura model [5] is empirical model to measure the radio signal strength in urban areas. The model was built by the collected data in Tokyo city. This model is applicable for frequencies in the range of 150 MHz to 1950 MHz and distance of 1 km to 100 km. it can be used for the base station antenna heights ranging from 30m to 1Km.To determine path loss using Okumara's model, the free space path loss between the points of interest is first determined and then the value of A_{mu} (*f, d*) is added to it along with correction factors according to the type of terrain.

The expression of the model

$$PL(dB)=L_F+A_{mn}(f, d)-G(h_{te})-G(h_{re})-G_{AERA} \quad (1)$$

Where

PL is path loss [dB], L_F is Free space path loss [dB] Amn (f, d) is Median attenuation relative to free space [dB], $G(h_{te})$ is Base station antenna height gain factor [dB], $G(h_{re})$ is Mobile station antenna height gain factor [dB], G_{AREA} is Gain due to the type of environment [dB], h_{te}: transmitter antenna height [m] hre: Receiver antenna height [m], d is Distance between transmitter and receiver antenna [km]

$G(h_{re}) = 10 \log_{10} (h_{re}/200)\ h_{re} < 3m$
$G(h_{re}) = 20 \log_{10} (h_{re}/200)\ 10m > h_{re} > 3m$
$G(h_{te}) = 20 \log_{10} (h_{te}/3)$

Okumura Model is considered to be among the simplest and best in terms of accuracy in predicting the path loss for early cellular system. The major disadvantages of this model are its slow response to rapid changes in terrain profile. Therefore the model is fairly good in urban and suburban areas, but not good for rural areas.

Hata Model

Hata model [13] is basically an empirical model based on Okumura model where some correction factor are included and it is valid from 150 MHz to 1500 MHz. Hata represented the Urban area propagation loss as the standard formula along with additional correction factor for application in the other situations such as suburban, rural among others. The computation time is short and only four parameter are required in Hata model. The path loss in dB for the urban areas is given by:

$$PL(dB) = 69.55 + 26.16 \log_{10}(f_c) - 13.82 \log_{10}(h_{te}) - a(h_{re}) + (44.9-6.55\log_{10} h_{te})\log_{10} D \quad (2)$$

Where

f_c = Frequency from 150 MHz to 1500 MHz, h_{te} = The effective base station antenna height (30m to 200m), h_{re}= The effective mobile antenna height (1m to 10m), D = The transmitter-receiver (T-R) distance in km, $a(h_{re})$ = The correction factor for effective mobile antenna height. For a small to medium sized city, the mobile antenna correction factor is given by

$$a(h_{re}) = (1.1\log f_c - 0.7)h_{re} - (1.56\log f_c - 0.8)$$

For a large city, it is given by

$$a(h_{re}) = 8.29(\log 1.5h_{re})^2 - 1.1 \text{ for } f_c < 300MHz$$

$a(h_{re}) = 3.2(\log 11.75)^2 - 4.97$ for $f_c > 300$MHz

To obtain the path loss in suburban area, the Hata standard formula is modified as

$$PL\ (dB) = PL(\text{ Urban}) - 2[\log(f_c/28)]^2 - 5.4 \quad (3)$$

Although Hata's model does not have any of the path specific correction which are available in Okumura model. This model is well suited for large cell mobile system, but not personal communication [5][12].

ECC-33Model

The ECC-33 model is developed by Electronic communication committee (ECC). This is generally used for FWA (Fixed Wireless Access) system. The path loss is defined as [10].

$$PL\ (dB) = A_{fs} + A_{bm} - G_b - G_r \quad (4)$$

Where

A_{fs}, A_{bm}, G_b and G_r are the free space attenuation, the basic median path loss, the BS height gain factor and the terminal height gain factor. They are the individually defined as

$$A_{fs} = 92.4 + 20\log_{10}(D) + 20\log_{10}(f)$$
$$A_{bm} = 20.41 + 9.83\log_{10}(D) + 7.894\log_{10}(f) + 9.56[\log_{10}(f)]^2 \quad (5)$$

$$G_b = \log_{10}(h_b/200)\{13.958 + 5.8[\log_{10}(D)]^2 \quad (6)$$

And for medium city environments,

$$G_r = [42.57 + 13.7\log_{10}(f)][\log_{10}(h_r) - 0.585] \quad (7)$$

Where

f is the frequency in GHz, D is the distance between Transmitter and Receiver in km, h_b is the BS antenna height in meters and h_r is the CPE antenna height in meters. The predictions using the ECC-33 model with the medium city option are compared with the measurements taken in suburban and urban environments [3][6].

COST-231Model

COST-231 model was devised as an extension to the Hata-Okumura model, The COST-231 model is designed to be used in the freq range 1500MHz to- 2GHz. This model contains corrections factor for urban, suburban and rural (flat) environments. The basic equation for path loss in dB is,

$$PL\ (dB) = 46.3 + 33.9\log_{10}(f) - 13.82\log_{10}(h_b) - a_{hm} + (44.9 - 6.55\log_{10}(h_b))\log_{10} D + c \quad (8)$$

Where

f is the frequency in MHz, D is the distance between AP and CPE antennas in km, and h_b is the AP antenna height above ground level in meters. The parameter c_m is defined as 0 dB for Medium sized city and suburban environments and 3dB for urban environment. All the parameters are

f=1500MHz to- 2GHz, h_{te} = 30m to 200m

h_{re} = 1m to 10m, d = 1km to 20 km [3], [12]

Lee Model

Lee's path loss model is based on empirical data chosen so as to model a flat terrain. Large errors arise when the model is applied to a non-terrain. However, Lee's model has been known to be more of a "North American model" than that of Hata. The propagation loss calculated as:

$$PL\ (dB) = 124 + 30.5\log_{10}(D/D_0) + 10k\log_{10}(f/f_c) - \alpha \quad (9)$$

Where,

D is in km, f and fc is in MHz, k = 2 for fc < 450 MHz and in suburban/open area and 3 for fc > 450 MHz and in urban area, D_0 = 1.6 km. f is the transmitted frequency, D is the Transmitter- Receiver distance and α_0 is a correction factor to account for BS and MS antenna heights[2].

Walfisch and Bertoni Model

The COST-231 model does not have the impact of diffraction from rooftops and buildings. A model which uses diffraction to predict average signal strength at street level is known as Walfisch-Bertoni model. The model considers the path loss to be the product of three factors:

$$L = P_0 Q^2 P_1 \quad (10)$$

Where

P_0 is the free space path loss for isotropic antennas, Q^2 gives the signal power reduction due to buildings which provides shadow the receiver at street level, and $P_1$is based on signal loss from the rooftop to the street due to diffraction.

In dB, the path loss is given by,

$$L = L_0 + L_{rts} + L_{ms} \quad (11)$$

Where

L_0 represents free space path loss, L_{rts} is the "rooftop-to-street diffraction and scatter loss", and L_{ms} is diffraction loss due to building[5].

Longley rice model

The Longley – rice model is generally used for point to point communication systems and it has a frequency range from 40MHz to 100GHz, over different types of terrain. The median transmission loss is predicated using the path geometry of terrain profile and the refractivity of the troposphere. The Longley-rice propagation prediction model is also referred to as the *ITS irregular terrain model.*

The Longley-Rice method generally operates in two modes, the path-specific parameters can be easily determined when the detailed terrain path profile is available and prediction is called a point-to-point mode prediction. On the other hand, the Longley-Rice method provides techniques to estimate the path-specific parameters, if the terrain path profile is not available, and such a prediction is called *area mode* prediction. One shortcoming of the Longley-Rice model is that it does not provide a way of determining correction due to environmental factor in the immediate vicinity of the mobile receiver, or consider correction factors to account for the effects of building and foliage. Further, multipath is not considered[5].

Stanford University Interim (SUI) Model

The proposed standards for the frequency bands below 11 GHz contain the channel models developed by Stanford University, namely the SUI models. Note that these models

are defined for the Multipoint Microwave Distribution System (MMDS) frequency band in the USA, which is from 2.5 GHz to 2.7 GHz. Their applicability to the 3.5 GHz frequency band that is in use in the UK has so far not been clearly established[6]. The SUI models are considered into three types of terrains, namely A, B and C. Type A is associated with maximum path loss and is appropriate for hilly terrain with moderate to heavy foliage densities. Type C is associated with minimum path loss and applies to flat terrain with light tree densities. Type B is characterized with either mostly flat terrains with moderate to heavy tree densities or hilly terrains with light tree densities. The basic path loss equation with correction factors is presented by [7], [8],

$$PL = A+10\gamma\log_{10}d/d_0+X_f+X_h+s \text{ for } d > d_0 \quad (12)$$

where

d is the distance between the AP and the CPE antennas in meters, d_0 = 100 m and s is a log normally distributed factor that is used to account for the shadow fading owing to trees and other clutter and has a value between 8.2 dB and 10.6 dB [7]. The other parameters are defined as,

$A = 20 \log_{10} (4 \pi d_0/ \lambda)$

$\gamma = a - bh_b + c/h_b$

Where,

The parameter h_b is the base station height above ground in meters and should be between 10 m and 80 m. The constants used for a, b and c are given in Table I. The parameter γ is given above which is equal to the path loss exponent. For a given terrain type the path loss exponent is determined by h_b.

Table I

Model Parameter	Terrain A	Terrain B	Terrain C
a	4.6	4.0	3.6
b(m^{-1})	0.0075	0.0065	0.005
c(m)	12.6	17.1	20

Numerical Values for the Sui Model

The correction factors for the operating frequency and for the CPE antenna height for the model are [7]

$X_f = 6.0 \log_{10} (f/2000)$

and

$X_h = -10.8 \log_{10}(h_r /2000)$ for terrain types A& B

$= -20.0 \log_{10} (h_r /2000)$ for Terrain type C

where

f is the frequency in MHz and h_r is the CPE antenna height above ground in meters. The SUI model is used to predict the path loss in all three environments namely rural, suburban and urban.

Egly propagation model

Egly is simplified model that assumes gently rolling terrain with average hill heights of approximately 50 feet, Because of this assumption, no terrain elevation data between the transmit and receive facilities is needed. Instead, the free space propagation loss is adjusted for the height of the transmit and receive antennas above ground. As with many other propagation models, Egli is based on measured propagation paths and then reduced to mathematical model. In case of Egli, the model consist of a single equation for the propagation loss[9].

$$A=117+40\log D_{mile}+20\log F-20\log(H_T*H_R) \quad (13)$$

Where

A is the attenuation in dB(between dipole), D is the path distance in miles, F is the frequency in Mega Hertz, H_T is the transmitter antenna height above ground level(AGL)in feet, H_R is the receiver antenna height above ground level in feet.

The typical equation used for Free Space loss between half wave dipole antenna (in dB) is

$$A_{FS}=32.27 + 20\log D_{miles} + 20\log F_{MHZ}$$

To isolate the propagation of the loss attributable to Egli consideration, subtract the free-space portion from the computed Egli attenuation:

$$A_{Eg}=A-A_{FS}=84.73+20\log D_{mils}20\log(H_T*H_R) \quad (14)$$

If the value of A Egli is zero or less, then free space valued is used. The Egli model should not be used in such type of areas like areas of rugged terrain, significant obstructions etc. Egli says it is limited to those areas which are similar to plain earth, such as relatively short over water and very flat barren land paths.

COST 231 Walfish-Ikegami (W-I) Model

This model is a combination of J. Walfish and F. Ikegami model. The COST 231 project further developed this model. Now it is known as a COST

231 Walfish-Ikegami (W-I) model. This model is most suitable for flat suburban and urban areas that have uniform building height. Among other models like the Hata model, COST 231 W-I model gives a more precise path loss. This is as a result of the additional parameters introduced which characterized the different environments. It distinguishes different terrain with different proposed parameters. The equation of the proposed model is expressed in, [4]

For LOS condition

$$(PL)_{LOS} = 42.6 + 26\log(d) + 20\log(f) \quad (15)$$

And for NLOS condition

$$(PL)_{NLOS} = \{L_{FSL}+ L_{rts}+ L_{msd}\} \quad (16)$$

for urban and suburban

This is the extended version of COST-231 this model consists rooftop losses and building losses.

Bullington model

This model is used to compute the diffraction loss over multiple knife edges. It defines the new effective obstacle at the point where the line of sight from two antennas crosses. This model many practical applications in urban and rural areas [11].

Epstein-Peterson model

This Epstein-Peterson model is similar in nature to the Bullington model but the exception is that it takes to draw line of sight between revelent obstacles, and to add the diffraction loss at each obstacles. However this model does not take urban losses into account and 10 dB or more must be added to the calculated loss in urban areas [11].

Conclusion

In this paper we surveyed different types of propagation model with their path loss equation. some of them model are used in urban, suburban area but some are in rural areas. For example Hata-Okumura are better in suburban areas and the Longley-rice model in rural areas.

References

[1] H.R. Aderson, "Fixed Broadband Wireless system Design" John Wiley & co 2003.

[2] F.D.Alotaibi and A.A.Ali April 2008, "Tunning of Lee path loss model based on recent RF measurement in 400MHz conducted in Riyadh city, Saudi Arabia" The Arabian Journal for Science and Engg. Vol 33, no. 1b pp 145-152.

[3] K.Ayyappan, P.Dananjayan "Propagation model for highway in mobile communication system" may 2007.

[4] H.K.Sharma, S.Sahu, S.sharma, "Enhanced Cost231W.I.Propagation Model in Wireless Network" *International Journal of Computer Application(0075-8887) Volume 19-No.6, April 2011.*

[5] T.S.Rappaport, "Wireless Communications", Pearson Education, pp.150-154, Second Edition.

[6] Abhayawardhana, V.S., *et al*., 2003, "comparison of Empirical propagation Path loss Model for Fixed Wireless Access Systems". Project funded by Ofcom, UK

[7] V. Erceg, K. V. S. Hari, *et al.*, "Channel models for fixed wireless applications, " tech. rep., IEEE 802.16 Broadband Wireless Access Working b Group, January 2001.

[8] V. Erceg, L.J.Greenstein, *et al.,* "An Empirically Path loss model for Wireless channels in suburban environments" IEEE Journal on selected areas of Communications, vol. 17, pp.1205-1211, July1999.

[9] The Egli Model is described in "Radio Propagation above 40MC Over Irregular Terrain, *(Proceeding of the IRE, vol.45, Oct. 1957, pp1383-1391).*

[10] Purnima.K.Sharma.et.al. "Comparative Analysis of Propagation Path Loss Models with Field Measured data" international journal of Engineering Science and Technology, Vol.2(6), 2010, 2008-2013.

[11] "Communication study" NTIA Report 82-100*, 1968* and NBS Tech Note, Vols, 1 and 11, 1967.

[12] Armoogum.V, Soyjaudah.K.M.S, FogartyT.andMohamudallyN., "Comparative study of Path loss using existing models for Digital Television Broadcasting for Summer, Mauritius Vol. 4, pp 34-38, May Season in the north of Mauritius", Proceeding of Third Advanced IEEE International Conference on Telecommunication 2007.

[13] M. Hata, "Empirical formula for propagation loss in land mobile radio services, " *IEEE Transactions on Vehicular Technology*, vol. vol. VT-29, pp. 317–325, September 1981.

Authors Biography

Mrs. Pooja Prajesh was born on 30th June 1978 in Roorkee, Uttrakhand (India). She received her M.Tech.degree in Digital Communication from Uttrakhand Technical University (U.K.), India. She is a Associate Member of the AMIETE. She has published several Research papers in national and international journals/conferences. She is presently research scholar in Uttarakhand Technical University, Dehradun (India). Her present research interest is in Wireless Communication.

Dr. R.K. Singh Professor, KEC, Dwarahat, Almora, He is member of academic staff of Kumaon Engineering College, Dwarahat, Almora, where he is a professor in the department of Electronics and Communication Engineering. Dr. Singh has given his contribution to the area of Microelectronics, Fiber Optic Communications, and Solid State Devices. He has published several research papers in seminar/conference and journal papers. He is member of several institutional and educational and educational bodies. Before joining Kumaon Engineering College, Dwarahat, he has worked in Birla Institute of Technology and Sciences (BITS), Pilani, and Central Electronics Engineering Research Institute (CEERI) Pilani. At present he is serving as OSD, in newly established Technical University of Uttarakhand known as Uttarakhand Technical University, Dehradun (INDIA).

Implementation of ANT Swarm Intelligence over Mobile Autonomous Robots with Customized Wireless Communication Model

K. Uma Rao, Akshay D. and Sridhar S.

RNS Institute of Technology, Bengaluru, India
E-mail: umarao_k@yahoo.co.in

Abstract
This paper presents a unique and cost effective mode of implementing ANT Swarm intelligence over MAR (Mobile Autonomous Robots). It also provides a mode of communication for moderate or less dense swarm of MAR's. RF communication system is used for data transfer between MAR's, as it provides a cost effective solution compared to other standard available wireless communication methods. Infrared transceivers based circuits are developed that enables the MAR's to follow one another without active mode of communication and hence decreasing the complexities related to wireless communication to a very large extent. This paper tries to give a right blend of active and passive mode of communication between MAR's enabling them with ANT like behavior in the least complex and inexpensive mode.

Keywords: Swarm Robots, Ant Swarm Intelligence, Mobile Autonomous Robots

Introduction

Robots are extensively being used to do a huge verity of tasks, as they promise far greater efficiency and consistency compared to the conventional methods. As the applications ranges from agriculture to military, providing them with intelligence of their own will surely add more to the advantages that it is already displaying. Swarm intelligence is a new approach which tries to provide the much required intelligence for multiple robot coordination to achieve a single given task. Swarm intelligence can be defined as a distributed problem solving technique based on self-organization theory and inspired by collective social behavior [1]. Compared to conventional robots swarm robots exhibits far greater reliability, efficiency and flexibility but the fact that in more than a decade not much of artificial swarm behavioral implementation have been carried out speaks for the subtle difficulties it is associated with. Each MAR (Mobile Autonomous Robot) should exhibit autonomous behavior with a check on coordination among the other MAR's. Thus, Communication plays a very important role to accomplish complex tasks in swarm robots' scenarios.

Various communication techniques are adopted by MAR's for having a reliable data transmission, some of them are Bluetooth [3], RFID tags, Ultrasonic and Infrared [3]. These methods have their own advantages and disadvantages which is vastly dependent on the scenario of MAR usage. This paper explores a new unique and cost effective mode of implementing ANT Swarm intelligence over MAR's using a perfect blend of both passive and active mode of communication. The technique uses a combination of customized Infrared circuit for passive communication and short range RF for active communication.

Ants are very fascinating creatures having the ability to setup a very complex colony structure in which every member has a role to play. Each Ant in the colony has its own task to be completed and yet the group as a whole appears to be highly organized. Ants form and maintain a line to their food source by laying a trail of pheromone (which is a chemical) to which other members of the same species are very sensitive to. This characteristic of Ants is tried and implemented over the developed MARs. In swarm scenarios this phenomenon is termed as "Follow the Leader". Lot of optimization techniques are implemented for formation of the line as well as finding the shortest distance between the starting to the final point. Infrared light is emitted from each of the MAR to which the MAR behind it is attracted to, this forms the passive method of forming and maintaining the line and RF module forms the active mode of communication, using which the MARs can interact with one another. Information about the surrounding environment explored by each MAR can be a very important and vital in accelerating the process of solving a given task. A proper active mode of communication could help in accelerating the process of solving a given problem and also help in evolution of each MAR making it better and more efficient.

Passive Robot following using IR Sensors

A customized IR (Infrared) Sensor circuit is developed which is used by the MARs to follow one another. Infrared light is an electromagnetic radiation with longer wavelength than visible light. The IR wavelength is between 750nm and 1mm [3636]. The IR is divided into three bands [3636/10]:

IR A: 700nm – 1400nm
IR B: 1400nm – 3000nm
IR C: 3000nm – 1mm

In this paper IR transmitter and receiver works at a wavelength of 940nm falling in the IR A band. The passive robot following technique uses 2 IR components; IR emitter

(TSAL5100) and an IR phototransistor (TEFT4300). These components are provided in plastic packages. TEFT4300 is a sensitive and high speed phototransistor. It is suitable for sensing nearby IR radiations with fast response time. TSAL5100 is a standard IR emitter diode, having the required features for the application under consideration.

IR emitter circuit

IR emitter circuit is formed by an array of strategically placed TSAL5100. Every MAR is fitted with an IR emitter circuit at its back. The circuit was designed keeping into consideration the physical dimensions of robot, the disperse angle (or Angle of half intensity) and lumen output of the IR led for the given current. It is very evident from Fig. 1 that at half intensity we get the maximum spread and as our application demanded for distance and not spread the IR LED were fed with maximum forward continuous current which is 220mA.

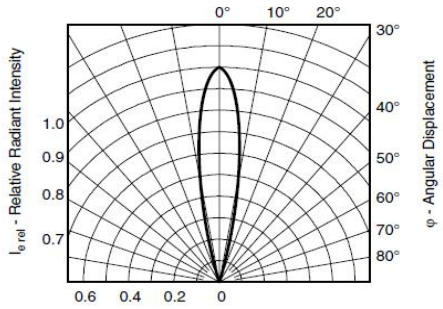


Figure 1: Angle of half intensity.

The IR emitter circuit is a x b in dimension having 6 IR emitters LED separated by a distance of 1 cm each. Each IR emitter LED had a series current limiting resistor and a pot connected to it and there are 6 of such series combinations connected in parallel. By default the circuit is designed such that maximum continuous current, the IR emitter can withstand is allowed to flows through it, the pot connected in series is used to adjust the range and intensity as required. This circuit is used to create a band of IR light source which substitutes the pheromone (The chemical Ants actually use for following each other). The created IR band is approximately 6cm wide and it is used by the MAR behind it to analyze the movement and behavior pattern of the considered MAR. Fig. 2 shows the placement of the IR emitter circuit on the MAR.

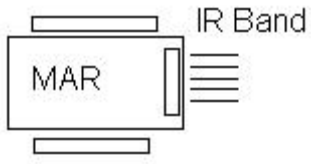


Figure 2: Placement of IR emitter on MAR

IR receiver circuit

IR receiver circuit is formed by 2 phototransistors (TEFT4300). As the name goes the phototransistors are devices that get forward biased once IR rays fall over them. The circuit basically consists of 2 voltage divider circuits placed a little more that 6 cm apart. IR receiver circuit is placed at the front of the MAR, which detects the movement of the MAR in front of it. From Fig. 3 it is seen that the spread of +-30 can be achieved using TEFT4300. The output of the voltage divider using a TEFT4300 will usually not give the distinct variation in output voltage that is required. A simple signal condition circuit is developed to get the required distinct voltage difference between the detection of IR band of the other MAR.

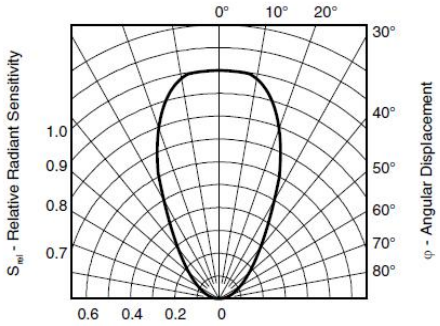


Figure 3: Angle of half intensity.

The signal conditioning circuit basically consists of a LM358 (dual comparator IC). The inverting terminal of the IC is fed with the output of the TEFT4300 circuit and a reference voltage is fed to the non inverting terminal of one of the comparator in LM358. The reference voltage is generated using yet another voltage divider circuit consisting of a resistor and a pot. By varying the pot we can adjust the reference voltage that is fed to the non inverting terminal of the comparator. The output of the IC is a more clearly define logic high (~5V) or logic low (~0V) which is understood by the microcontroller. Fig. 4 shows the placement of the IR receiver circuit on the MAR.

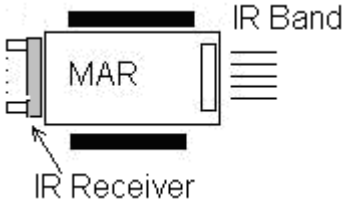


Figure 4: Placement of IR receiver on MAR.

Following the leader

Except the leader all the MARs have an IR emitter at its back and an IR receiver at its front. The circuit is placed such that by default the receiver of one MAR is out of the IR band generated by the IR emitter circuit placed on the MAR in front of it. When the leader or the MAR in front of the considered MAR drifts in one direction it shall be detected by the receiver circuit and it shall send a signal to the microcontroller to drift in the same direction. This is described in more detail by considering the following cases.

***Case1*:** In this case the MARs in a straight line and the receivers are out of the IR band that is been generated by the IR emitter which is placed in the MAR in front of it. Both the receivers of the considered MAR reads a logic 0 (output of the sensor is 0V). Under such circumstance the MAR is coded to move straight. Fig. 5 gives a diagrammatic representation of the position of MAR and the sensors.

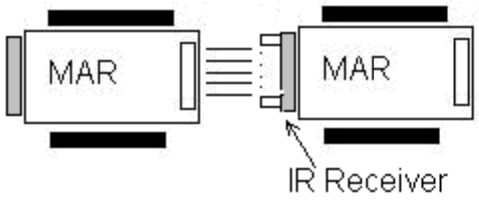


Figure 5: Default position.

***Case2*:** In this case the MARs in front drifts towards the left and the left receiver detects the movement of the robot in front of it and sends a signal to the controller. The left receiver of the considered MAR reads a logic 1 (output of the sensor is 5V) and right receiver reads a logic 0 (output of the sensor is 0V). Under such circumstance the MAR is coded to move towards left till default position is reached. Fig. 6 gives a diagrammatic representation of the position of MAR and the sensors.

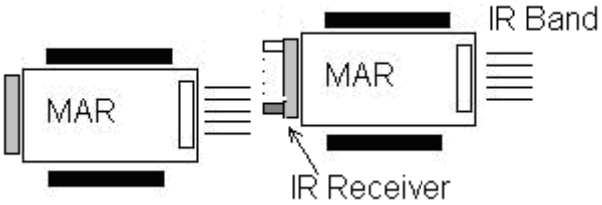


Figure 6: Left Drift.

***Case3*:** In this case the MARs in front drifts towards the right and the right receiver detects the movement of the robot in front of it and sends a signal to the controller. The right receiver of the considered MAR reads a logic 1 (output of the sensor is 5V) and left receiver reads a logic 0 (output of the sensor is 0V). Under such circumstance the MAR is coded to move towards right till default position is reached. Fig. 7 gives a diagrammatic representation of the position of MAR and the sensors.

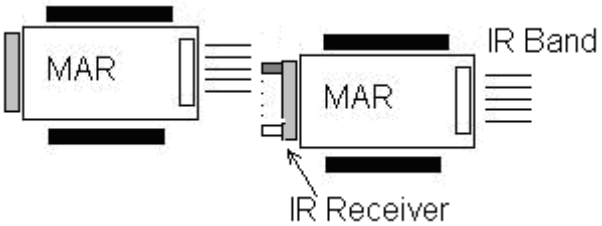


Figure 5: Right Drift.

Active RF based Communication

RF modules were used for active communication between the robots. At first the information to be sent should be coded as the selected transmitter cannot directly transmit the data. Hence HT12E encoder was used for doing the same. Fig. 6 shows circuit diagram that represents the connection of the encoder and transmitter.

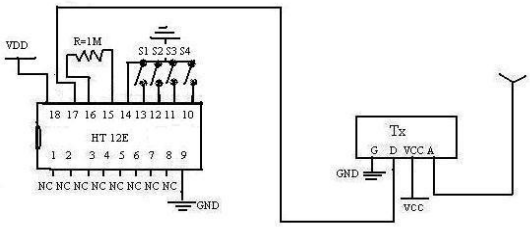


Figure 6: Circuit diagram of transmitter.

The transmitter basically has 4 pins; the first goes to ground, the second gets the input from the encoder, the third is connected to VCC and the last one will be connected to an antenna.

Similarly on the receiving end we need a receiver and the output of the receiver must be decoded. A HT12D decoder is used to decode the information. The output of the decoder will then go to the microcontroller. Fig. 7 gives the circuit diagram showing the connections between the decoder and receiver.

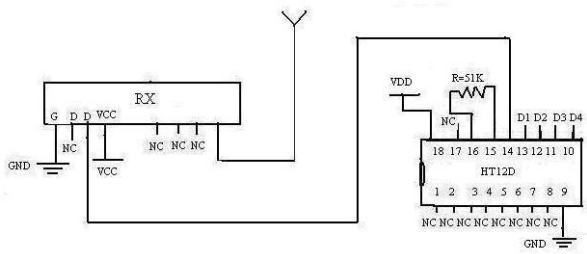


Figure 7: Circuit diagram of receiver.

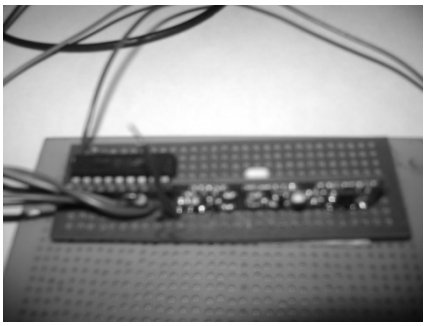


Figure 8: Receiver on the robot.

Similar to the transmitter the RF receiver has 2 pins for VCC, ground and a pin that gives the received information. This is then given to the decoder which in turn gives us the required data. Data transmission and reception is done at a

frequency of 315MHz.

Each MAR is equipped with a RF transmitter and a RF receiver. The mode of communication is selected to be broadcast and all the RF modules work at the same frequency of 315MHz. A customized protocol is used by the MAR to understand and communicate efficiently. This paper deals only with the broadcast mode of operation though other singular transmission techniques have been implemented over the MAR which is essential for solving few scenarios.

RF protocol

A separate protocol was developed to meet the unique requirements of the MARs. Each frame of information that is been transmitted consists of 4 parts. And each part is formed by a nibble (4 bits, selected transmitter restricts from having a byte of information to be transmitted).

The first part is the start command indicating the start of transmission. To avoid congestion due to simultaneous transmission of information from various MARs, before transmission the MAR waits and listens for a fixed period of 25ms and if it receives nothing then it starts sending the complete frame but on the other hand if it receives information during that 20ms, transmission is postponed giving more preference for analyzing the received data. Hex value of 2 indicates the start of transmission.

The second part indicated the address or identity of the sending MAR. This is very important as it is used to identify the source of the information being received. Information received from master is given more importance compared to information received from other MARs and also while transmission as all the transmitters work at the same frequency the sending MAR will receive a copy of the information its transmitting, in which case the data will be rejected but the received start frame and address frame is used to check if transmission is successful, if not the entire frame is retransmitted. As all the frames are send in form of nibbles the maximum number of MARs it can handle is 16. And hence the address range from 0H to 15H.

The third part involves the data to be transmitted. There are 16 different commands than can be transmitted, each having very specific information to be conveyed. This is used to send information related to the explored area to the other MARs, sending emergency commands in terms of breaking the line and going off track, and various other features can be added depending on the demands of the scenario under consideration.

The final part is the stop part indicating the termination of the transmission which is represented by a value 4H. The frame structure can be seen in Fig. 9

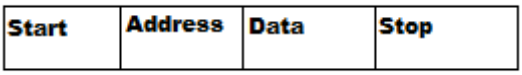


Figure 9: Transmission Frame.

The selected RF Transmitter and receiver is a low cost module with a short range than enable us to coordinate with small or less dense swarm of robots.

Conclusions

In this paper we have presented the implementation of ANT like behavior over MAR (Mobile Autonomous Robots) using a combination of IR and RF modes of passive and active communication. The emphasis was on developing an inexpensive solution for deploying a small or moderate swarm of robots. Wireless communication being a necessity in this field and it being expensive makes this very difficult to implement. This paper brings out a perfect blend of passive leader following technique and a dedicated active mode of communication for realizing the same.

References

[1] "Evolutionary Swarm Intelligence Applied to Robotics", Sidney N. Givigi, Jr., and Howard M. Schwartz, Proceedings of the IEEE International Conference on Mechatronics and Automation, Niagara Falls, Canada , July 2005

[2] "Implementation of RF Communication with TDMA Algorithm in Swarm Robots", Sagar Bhandari, Prasanna Gautam, 2008

[3] "A Short Range Infrared Communication for Swarm Mobile Robots", Farshad Arvin, Khairulmizam Samsudin and Abdul Rahman Ramli, 2009 International Conference on Signal Processing Systems

[4] "Coordination with the leader in a robotic team without active communication", Ming Cao, Changbin Yu and Brian D. O. Anderson, 17th Mediterranean Conference on Control and Automation, Greece, June 24-26, 2009

[5] "An analysis of Collective Movement Models for Robotic Swarms", W.A.F.W. Othman, B.P. Amavasai, S.P.McKibbin and F.Caparrelli, EUROCON 2007 the Int Conf " Computer as a tool"

[6] "On the implementation of a robotic SWARM test-bed", Xiaolei Hou and Changbin Yu, Proceedings of the 4th International Conference on Automation Robots and Agents, Feb 10-12, 2009.

Proposed Bluetooth Protocol for Short Range Communication

Kamani Krunal C., Kathiriya Dhaval R. and Ghodasara Yogesh R.

Asst. Prof., Director (IT), Assoc. Prof.

Abstract
We present the interface for developing new design for proximity wireless transaction system a new routing method is based on the concept of Bluetooth relay transmission based on Java (J2ME with wireless plug-in) programming, in which it is possible that mobile devices (mobile phones, PDAs) can interact with each other in a proposed architecture formation.

For example in piconets, where one device is a Slave in both Piconets, we can able to send message to any node though the node which is common in both the networks.

Therefore, we have design a new protocol in our program. Building on that a device C1 could send a message to device New, which is not within the range of C1, via devices C3, which are within 10 m range of each other. So C1 is in range of New, C3 and works as bridge.

Index Terms: Bluetooth, Scatternets, Piconet, Routing, Protocol.

Bluetooth Concepts

Bluetooth is an emerging standard for wireless connectivity. It specifies a system — not just a radio — that encompasses the hardware, software framework, and interoperability requirements. And, the radio system is optimized for mobility. In other words, Bluetooth primarily specifies a cable-replacement technology that targets mobile users in the global marketplace.

Bluetooth technology was intended to hasten the convergence of voice and data to handheld devices, such as cellular telephones and portable computers. Through the efforts of its developers and the members of the Bluetooth Special Interest Group (SIG), it is now emerging with features and applications that not only remain true to its original intent, but also provide for broader uses of its technology.

In this paper, we provide you new basic design and structure to develop and launch new proximity transaction technology for wireless short range communication and designing of transaction system. We keep the minimum technical jargon to give you a detailed, thorough, yet understandable look at Bluetooth and its world. In this paper, we address the different objectives, to help you get started on the road to implementing Bluetooth technology and proximity wireless transaction system design:

Introduction

Bluetooth is an open standard specification for a radio frequency (RF)-based, short-range connectivity technology that promises to change the face of computing and wireless communication. It is designed to be an inexpensive, wireless networking system for all classes of portable devices, such as laptops, PDAs (personal digital assistants), and mobile phones. It also will enable wireless connections for desktop computers, making connections between monitors, printers, keyboards, and the CPU cable-free. [1]

Currently we are having the concept of scattenet, but we are not able to establish that in current environment using Bluetooth. Furthermore in current scenario direct communication is possible but relay communication is not possible. [2][3]

We have design a new interface in J2ME which is placed over RFCOM layer of Bluetooth architecture and this interface enables the communications among the nodes which are not in the range of each other and it also enable to communicate on multiple available paths. In short the proposed design extends the range of short range network, in which of one of the node works as bridge through which node can connect with remote node which is not in the range.

Bluetooth architecture

Figure 1 shows a simplified Bluetooth architecture.

The radio layer defines the requirement of the Bluetooth transceiver devices operating in the 2.4 GHZ ISM band. It is the lowest layer in the Bluetooth architecture, and uses pseudo-random hopping sequences with a fast hopping rate of 1600 hops per second.

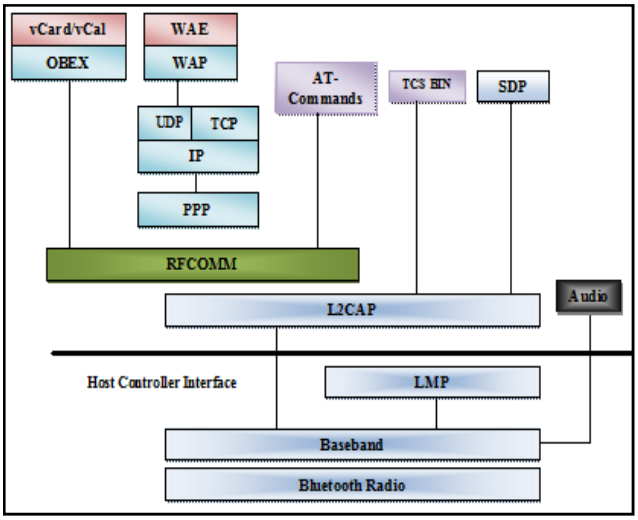


Figure 1: Bluetooth Architecture.

The baseband layer is similar to the physical layer in general network architecture. The baseband manages many things like physical channels, links, error correction, hop selection and so on. Here we will mostly concentrate on link. The set up of connection between two devices need two steps: inquiry procedure and page procedure. The Bluetooth device uses inquiry procedure to know the address of the other device. Then the device performs the page procedure to connect to the other one using the returned address.[4]

LMP (Link Manage Protocol) takes care of link configuration and authentication. The authentication uses a link key. If two devices do not have common link key, it can be created from a PIN. Besides, the link key can be changed temporarily, thought this change can only be valid for the session. Another important use of LMP is its support for name request to another device. The name consists of a maximum of 248 bytes according to the UTF-8 standard. If you want to close the connection between two devices, just use LMP to detach them.

HCI (Host Control Interface) is the interface to LMP and baseband, and it can access to hardware status and control registers. All protocols above HCI are software based. L2CAP (Logic Link Control and Adaptation Protocol) is the lowest layer of such protocols. The connection primitives it provides are: set up, configure and disconnect.

The Bluetooth protocol stack can be divided into four layers according to their purpose including the aspect whether Bluetooth SIG has been involved in specifying these protocols. Table 1 shows the protocols belong into the layers.[5]

The Bluetooth technology provides both a point-to-point connection and a point-to-multipoint connection. In point-to-multipoint connections, the channel is shared among several Bluetooth units. In point-to-point connections, only two units share the connection.

Bluetooth protocols assume that a small number of units will participate in communications at any given time. These small groups are called piconets, and they consist of one master unit and up to seven active slave units. The master is the unit that initiates transmissions, and the slaves are the responding units. This type of Bluetooth network can have only one master unit.

Table 1: the Protocol and Layers in Bluetooth Protocol Stack.

Protocol layer	Protocols in the stack
Bluetooth Core Protocols	Baseband, LMP, L2CAP, SDP
Cable Replacement Protocol	RFCOMM
Telephony Control Protocols	TCS Binary, AT-commands
Adopted Protocols	PPP, UDP/TCP/IP, OBEX, WAP, vCard, vCal, IrMC1, WAE

Bluetooth Networking

If several piconets overlap a physical area, and members of the various piconets communicate with each other, this new, larger network is known as a scatternet. Any unit in one piconet can communicate in a second piconet as long as it serves as master for only one piconet at a time. [6]

Figure 2 shows the intercommunication between two nodes in different piconets.

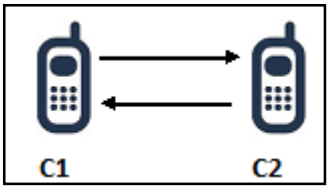


Figure 2: Inter Communcation in Piconet.

Proposed proximity wireless transaction system design

The proposed proximity wireless transaction system is to design a transaction system to deliver data packets in Bluetooth scatternet.

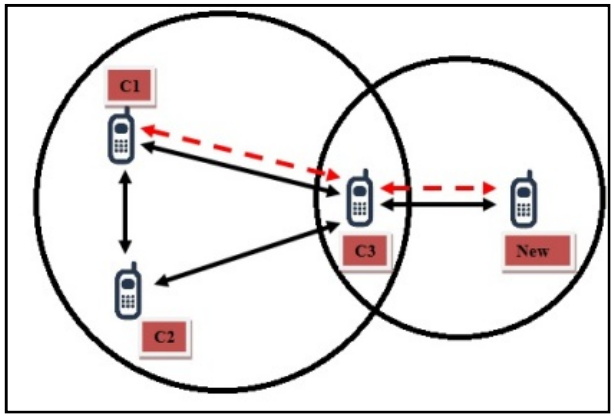


Figure 3: New proposed architecture design, C3 as Bridge.

In proposed design extends the range of short range network, in which of one of the node works as bridge through which node can connect with remote node which is not in the range. For example C1, C2 and C3 are in one network and New node is n different network. C3 is a common node though which we can able to communicate among C1 to New via C3 as shown with dotted line in figure 3. [4][7]

Furthermore we can send message C1 to C2 directly as well as C1 to C2 via C3 in the same network.

1. It is possible to transfer data/messages between scatternet.
2. Each device has membership information to deliver/forward data/message.
3. Performance analysis of the model.(transfer rate, packet loss etc…)[3]

Implementing Transaction System with J2ME Emulation Environment

To achieve the goal following tools were used.

- IDE : Eclipse Java EE IDE for Web Developers. Version: Indigo Release. Build id: 20110615-0604
- JDK : 1.4

- EclipseME 1.7.9 provides cross-platform support for developing J2ME midlets within the Eclipse IDE. Supported functions include development, pre verification and emulator launching and debugging. (EclipseME J2ME Development Tools for Eclipse).
- J2ME: J2ME Wireless toolkit 2.2.

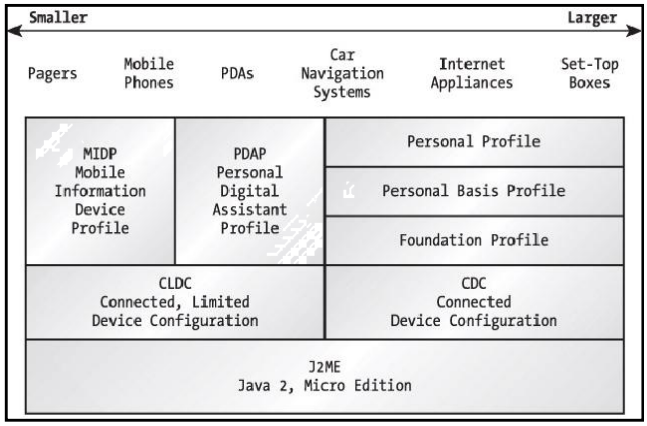


Figure 4: J2ME Architecture.

J2ME isn't a specific piece of software or specification. All it means is Java for small devices. Small devices range in size from pagers, mobile phones, and personal digital assistants (PDAs), all the way up to things like set-top boxes that are just shy of being desktop PCs.[9]

J2ME is divided into configurations, profiles, and optional APIs, which provide specific information about APIs and different families of devices. A configuration is designed for a specific kind of device based on memory constraints and processor power. It specifies a Java Virtual Machine (JVM) that can be easily ported to devices supporting the configuration. It also specifies some subset of the Java 2 Platform, Standard Edition (J2SE) APIs that will be used on the platform, as well as additional APIs that may be necessary.

Profiles are more specific than configurations. A profile is based on a configuration and adds APIs for user interface, persistent storage, and whatever else is necessary to develop running applications.

Optional APIs define specific additional functionality that may be included in a particular configuration. The whole caboodle—configuration, profile, and optional APIs—that is implemented on a device is called a stack. For example, a possible future device stack might be CLDC/MIDP + Mobile Media API. See the section later on platform standardization for information on JSR 185, which will define standard J2ME stacks. [10]

When we performed practically using the eclipse emulator the action performed in cellular environment like handshaking, acknowledgment, connection status, etc on different layers performed in the background is shown here in table -2.

Table 2: Snippets of code in emulator.

```
Running with storage root
temp.DefaultColorPhone1312364447594
ChatMain: invoke startApp()
ChatMain: invoke commandAction. command=Chat
ChatMain: set local nick name to c3
NetLayer: invoke init()
Print Local Device 000026E768A8
Name: WirelessToolkit
MajorDevice:Phone
MinorDevice:Cellular
ServiceClass:
NetLayer: invoke query()
Print Service Record (# of element: 4)
Print Service Record URL
btspp://000026E768A8:1;master=false;encrypt=false;authentic
ate=false
DataElement[ServiceAvailability] 255
NetLayer: invoke deviceDiscovered name=WirelessToolkit
DataElement[ProtocolDescriptorList] 48 (# of element: 2)
NetLayer: invoke deviceDiscovered name=WirelessToolkit
  DataElement[ProtocolDescriptorList] 48 (# of element: 1)
NetLayer: invoke inqueryCompleted
    DataElement[ProtocolDescriptorList] L2CAP
  DataElement[ProtocolDescriptorList] 48 (# of element: 2)
    DataElement[ProtocolDescriptorList] RFCOMM
    NetLayer: invoke serviceSearchCompleted: 16
NetLayer: SERVICE_SEARCH_COMPLETED
NetLayer: BlueChat service
url=btspp://0123456789AF:1;master=false;encrypt=false;auth
enticate=false
NetLayer: a new active EndPoint is established.
name=WirelessToolkit Address ::::0123456789AF
EndPoint: invoke putString 0 c3
Sender: sending signal 0 string 'c3' to WirelessToolkitTarget
Device:All
Reader: waiting for next signal from WirelessToolkit
NetLayer: search service on device WirelessToolkit
Reader: read in  HANDSHAKE_ACK name c1 from
WirelessToolkit
```

Result

In eclipse environment, nodes in different network can able to communicate with each other. We have taken 3 different nodes C1, C2 and C3 in one network. Up to now these nodes are able communicate with each other is same network directly. But now after using our interface they can able to communicate with each other with the help other node in same network. For example C1 can communicate with C2 via C3 that is represented in figure 5 and table -2 by C2(C3).[11]

Furthermore, we have added new node with name New as shown in Figure 5 in different network. New node is directly connected with C3, which is common in both networks. Now C1 and C2 can able to communicate with New node of another network with the help of C3 as shown in table 3 in bold-italic font New(C3). [12][13]

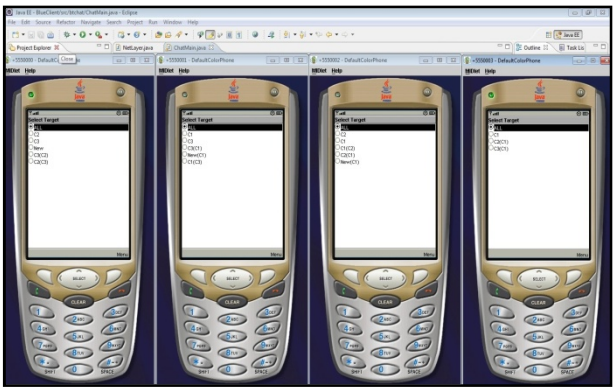


Figure 5: Nodes in Eclipse Emulation.

Table 3: Connectivity among different nodes.

Nodes →	C1	C2	C3	New
Connection	C2	C1	C1	***C3***
	C3	C3	C2	***C1(C3)***
	C3(C2)	C1(C3)	***New***	***C2(C3)***
	C2(C3)	C3(C1)	C2(C1)	
	New(C3)	***New(C3)***	C1(C2)	

Summary and Conclusion

In this paper a preview of new design for proximity wireless transaction system which makes Bluetooth a proficient. To summarize this research the first step was to study and analyze the current architecture. The second phase was to find out an appropriate technology which enables both nodes to communicate directly. Our proposed routing method is based on the concept of Bluetooth relay transmission based on J2ME programming, which is an efficient method for encoding source route paths in Bluetooth.

References

[1] http://en.wikipedia.org/wiki/Bluetooth
[2] http://www.techxperts.co.in/
[3] http://www.bonrix.net
[4] http://www.wirelessdevnet.com/channels/bluetooth/features/bluetooth.html
[5] Kamani,K.C. Kathiriya,D.R. Virapariya, P.V. and Ghodasara,Y.R. (2010) an Protocol Usage Model & An Architecture Design for Bluetooth – A Short Range Transaction System. *International Journal of Advance Research in Computer Science*, Vol. 1 No. 4.
[6] I. Stojmenovic and N. Zaguia, *"Bluetooth scatternet formation in ad hoc wireless networks".*
[7] http://iossoftwares.com/
[8] Kamani,K.C. Kathiriya,D.R. Virapariya, P.V. and Ghodasara,Y.R. (2011) Routing Mechanism in Bluetooth: A Short Range Communication Technology. In *proceedings of the National Journal of System and Information Technology (NJSIT).*
[9] Java: The Complete Reference, Seventh Edition - Herbert Schildt.
[10] J2ME: The Complete Reference - James Keogh
[11] Java Eclipse Tutorials: http://www.vogella.de/articles/Eclipse/article.html
[12] Java Eclipse Tutorials: http://javaprogrammingforums.com/java-jdk-ide-
[13] tutorials/253-beginners-eclipse-tutorial-how-run-your-first-eclipse-java-application.html

Authors Biography

Prof. Krunal Kamani is currently working as an Assistant Professor (Computer Science) Information Technology Center, Anand Agricultural University, Anand.

He is having M.Phil. (Computer Science) from Madurai Kamaraj University, MCA from Saurashtra University, Rajkot & M.Sc. (IT) from Punjab Technical University, Punjab degree in the field of computer.

His publication includes 4 papers in international journal, 4 papers in national journals and 15 papers in national conferences/seminars.

He received 3rd rank for the best research paper at the national seminar at Shri M & N Virani Science College – Rajkot in the year 2006. (Email: kamanikrunal@aau.in)

Dr. Dhaval Kathiriya is currently working as a Director (Information Technology) at Anand Agricultural University, Anand He is Ph. D. in Computer Science from Saurashtra University, Rajkot His publication includes 4 books, 6 international research papers, 2 projects attended, and 11 national/state level conferences / seminars on various topics of IT. He is also involved in syllabus designing of MCA and M.Sc. (I.T.) programmes of Saurashtra University and Kadi Sarva Vishwavidyalaya. He is recognized Ph.D. guide in Computer Science at Kadi Sarva Vishva Vidyalaya, Singhania University, Rajasthan and Mevad University, Rajasthan. (Email: dit@aau.in)

Dr. Yogesh Ghodasara is currently working as an Associate Professor at the College of Agricultural Information Technology, Anand Agricultural University, Anand. He has completed his Ph.D. (Computer Science) in 2009 from Saurashtra University, Rajkot. His publication includes 5 international journal and 5 papers in national conferences/seminars. He was also involved in syllabus designing of BCA and B.Sc. (I.T.) programmes of Saurashtra University. (Email: yrghodasara77@yahoo.co.uk)

Mitigating Timing and Side-Channel Attack for Secure Data Communication in MANETs

[*]Manpreet Singh, [**]Sanjeev Rana and ***Sonia Arora

**Prof, Deptt. of Comp Sc. & Engg., Mullana, India*
E-mail: dr.manpreet.singh.in@gmail.com
***Assoc. Prof., Deptt. of Comp Sc. & Engg., MMU, Mullana, India*
E-mail: sanjeevrana1@gmail.com
****Deptt. of Comp Sc. & Engg., APIIT, Panipat, India*
E-mail: sonia@apiit.edu.in

Abstract
Over the past decade or so, there has been rapid growth in wireless and mobile applications technologies. Exchanging sensitive information over unprotected wireless links with unidentified and untrusted endpoints demand the deployment of security in MANETs. However, lack of infrastructure, mobility and resource constraints of devices, wireless communication links and other unique features of MANETs induce new challenges that make implementing security a very difficult task and require the design of specialized solutions [1]. In literature, no. of security mechanisms using RSA exist but RSA itself suffers from number of attacks i.e. timing attacks, adaptive chosen cipher attacks etc. In this paper, we proposed security enhancement mechanism to foil both timing attack and adaptive chosen cipher attacks for secure data communication.

Keywords: RSA, timing attacks, adaptive chosen cipher text attacks, AODV, DSR

Introduction

A MANET consists of mobile nodes, a router with multiple hosts and wireless communication devices. MANET have various characteristics as: the network topology may change randomly and rapidly at unpredictable times, and may consist of both bidirectional and unidirectional links[1]; Wireless links will continue to have significantly lower capacity than their hardwired counterparts[5]; some or all of the nodes in a MANET may rely on batteries or other exhaustible means for their energy.[6]; mobile wireless networks are generally more prone to physical security threats than are fixed- cable nets [1][4].

Because of the features listed above, the mobile ad hoc network will need more robust security scheme to ensure the security[7]. The various security issues in MANET are as: Link Level Security, Secure Routing, Key Distribution, Privacy [3].

Motivation

Cryptography is one of the approach to cope with the security issues in MANETs, In cryptography, RSA (which stands for Rivest, Shamir and Adleman) is an algorithm for public-key cryptography [2]. The RSA algorithm involves three steps: key generation, encryption and decryption [8].

Key generation

The keys for the RSA algorithm are generated in the following way:

1. Choose two distinct prime numbers p and q.
2. Compute n = p*q. n is used as the modulus for both the public and private keys.
3. Compute $\varphi(n) = (p - 1)*(q - 1)$, where $\varphi(n)$ is Euler's totient function.
4. Choose an integer e such that $1 < e < \varphi(n)$ and $\gcd(e,\varphi(n)) = 1$, i.e. e and $\varphi(n)$ are co-prime. e is released as the public key exponent. e having a short bit-length and small Hamming weight results in more efficient encryption – most commonly 0x10001 = 65537. However, small values of e (such as) have been shown to be less secure in some settings.[5]
5. Determine $d = e–1 \bmod \varphi(n)$; i.e. d is the multiplicative inverse of e mod $\varphi(n)$.

This is often computed using the extended Euclidean algorithm. d is kept as the private key exponent.

Encryption

Alice transmits her public key (n,e) to Bob and keeps the private key secret. Bob then wishes to send message M to Alice. He first turns M into an integer m, such that $0 < m < n$ by using an agreed-upon reversible protocol known as a padding scheme. He then computes the cipher text c corresponding to $c = m^e \pmod n$ and transmits c to Alice.

Decryption

Alice can recover m from c by using her private key exponent d via computing $m = c^d \pmod n$. Given m, she can recover the original message M by reversing the scheme.

Problem Definition

Many security solutions using RSA exist but they all exhibit problems as RSA suffers from attacks as discussed below:

***Timing Attacks* :** If the attacker Eve knows Alice's hardware in sufficient detail and is able to measure the decryption times for several known cipher texts, she can deduce the decryption key d quickly.

Adaptive chosen cipher text attacks(CCA2) : CCA2 is an interactive form of chosen-ciphertext attack in which an attacker sends a number of cipher texts to be decrypted, then uses the results of these decryptions to select subsequent cipher texts.

***Side – channel Attacks* :** side-channel attacks require technical knowledge of the internal operation of the system on which the cryptography is implemented

To overcome these problems we introduced the concept of hashing within RSA in our proposed solution.

Proposed Scheme

The objective is to develop an algorithm for secure data transmission using public key cryptography in MANET's. We used RSA which is a public key cryptography algorithm for encryption and Hash function for message authentication . We assume that Keys are distributed by KDC initially before deployment of Mobile nodes. Proposed solution used following notation described below:

$MANET_n$	MANET
M_n	Mobile node
N	Number of Mobile Nodes
P	Global Key Pool for all Mobile Nodes and KDC
K_i	Key Pool for Mobile Node
E	Encryption Algorithm
D	Decryption Algorithm
M	Message
C	Cipher
N_{Id}	Node ID
KDC	Key Distribution Center(offline)
U_i	Public Key
PT	Plan Text

Implementation of Proposed Scheme

Various steps involved in implementation of proposed scheme are:

Key Generation and initialization

Before the deployment of mobile nodes, KDC generates public key for each mobile node which is stored in the global key pool P and key pool of mobile node K_i. G_i key can be used as a common key for a particular group of mobile nodes. After key distribution KDC will become offline

1. KDC generates: : $U_i = Pi \,||\, Q_n$
2. $M_n \rightarrow K_i = U_i$
3. $P \rightarrow K_i = U_i$

Neighbor selection and Data transfer

After these private and public keys are distributed, Sender encrypts the data with its public key and sends it to the receiver. Receiver receives the message and decrypts it with private key and it also checks the hash code of the message before accepting it. Without using the keys nodes cannot transmit the data. When sender S has some data to send to its neighbor, it initiates the process of mutual authentication. Receiver receives the message and calculates E mod N. If this is a valid public key, only then data transfer starts. After every data transfer, keys are updated so this scheme prevents from various types of attacks which we discussed earlier. Every node follows the steps given below:

1. get_initial_Data(Prime Numbers p,q) from KDC and calculate n=get_product(p,q)
2. E=get_KeysCalculated (n, M_n)
3. M_n ->Dn= get_KeysGenerated(n, M_n)
4. Before data transfer sender repeat the following steps:
   a. authentic = Mutual_Authentication(sender, receiver)
   b. if (authentic) then
      $C = M^E \bmod n$
      Send_Data(C , receiver)

   Else
   Send_Data(false)
   End if

5. Receiver repeat the following steps:

$PT = C^D \bmod n$
HashCode =get_Hash(PT)
If (HashCode) then
accept_Data(PT)
else
accept_Data(false)
End if

Secure Data Communication and Authentication

Nodes can also update the key pairs as per following steps:

If (old_key(M_{ni} -> E_i == M_{nn} -> E_j)) then

Calculate n=p*q

Check_e_=Calculate M_{ni} -> E_i mod n

Check_d_=Calculate M_{nn} -> D_j mod n

If (Check_e_== M_{nn} -> E_j) then authentic=true else authentic=false

If (authentic) then keyUpdate(M_{ni} ->keypair, M_{nn} -> keypair)

End if

Performance Evaluation

We implemented our algorithm in NS-2 simulator.

Cryptography algorithm (RSA) is implemented in C++ and we used its interface in TCL for data transfer. We extended the NS-2 by adding RSA script in NS-2

Experiment Setup

Simulation is performed for various parameters as:

Table 1: MANET's Simulation Parameters.

Parameters	Value
Routing Protocols	AODV/DSR
Security Mechanism	RSA
No. of Nodes	10,20,30,40,50,60
Pause Time	5,10,15,20 ms
Simulation Time	10 min
Sampling Interval Time	0.5 sec

Performance Evaluation

To evaluate the performance of routing protocols, we used Throughput and End-to-End delay metrics to compare the performance of the selected protocols.

Throughput

The throughput is defined as the total amount of data a receiver receives from the sender divided by the time it takes for the receiver to get the last packet.

The graphs given below shows the comparison of AODV and DSR with RSA on the basis of throughput by taking various values for pause time and the no. of nodes

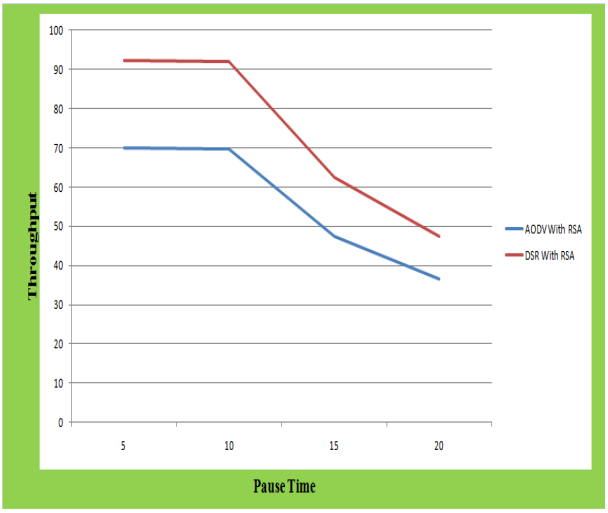


Figure 1: Comparison of throughput between AODV and DSR with RSA (considering pause time).

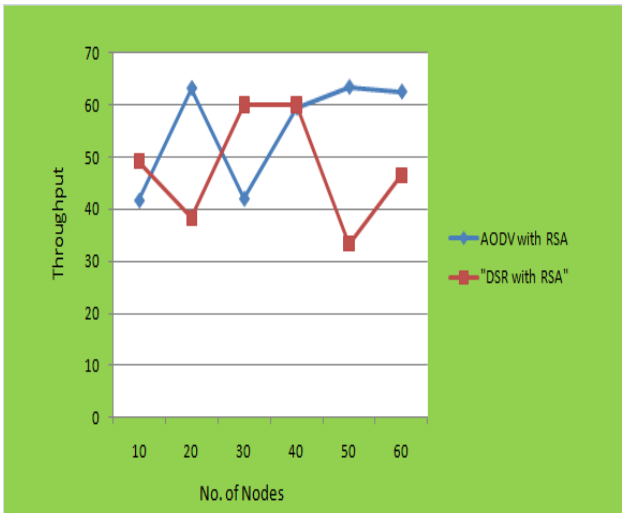


Figure 2: Comparison of throughput between AODV and DSR with RSA (considering nodes).

From the graphs shown above it is clear that, **DSR** protocol with RSA has better throughput as compared to AODV protocol with RSA when we consider the mobility of nodes whereas AODV protocol has better throughput as compared to DSR protocol when no of nodes is taken into consideration

End-to-end delay

End-to-end delay indicates how long it took for a packet to travel from the source to the application layer of the destination.

The graphs given below shows the comparison of AODV and DSR with RSA on the basis of End-to-End delay by taking various values for pause time and the no. of nodes

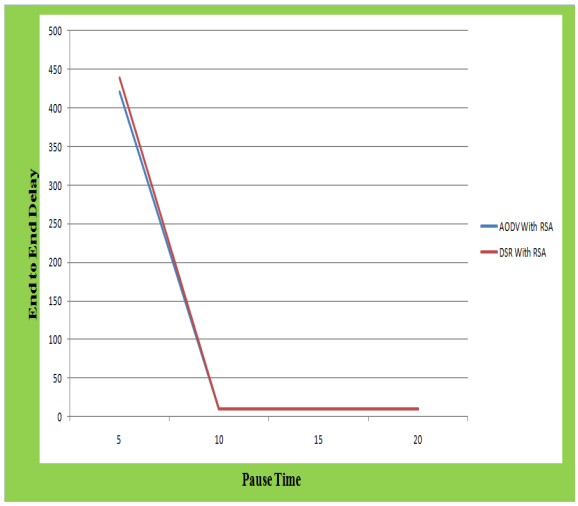


Figure 3: Comparison of End to End Delay between AODV and DSR with RSA(considering pause time)

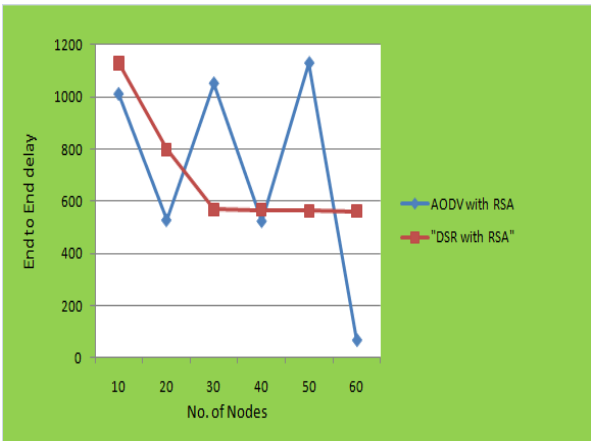


Figure 4: Comparison of End to End Delay between AODV and DSR with RSA(considering nodes)

From the above shown graphs we can conclude that **DSR** protocol with RSA has approximately same end to end delay time as in AODV protocol with RSA when we consider pause time but DSR protocol has the highest end to end delay than AODV protocol incase no of nodes is considered.

Conclusion

The concept of hashing within RSA and updation of keys after every data transfer has prevented the system against various attacks like like timing attacks etc. On the basis of performance of protocols , we can say that our algorithm works better with DSR instead of AODV i.e DSR is better performer than AODV in context of mobility of nodes. Whereas AODV protocol is best performer as compared to DSR and DSR protocol is the average performer in context of no. of nodes

References

[1] Erdal Çayırcı, "Security in Wireless Ad Hoc and Sensor Networks", A John Wiley and Sons, Ltd, Publication, Ist Edition-2009

[2] Rivest, R.; A. Shamir; L. Adleman (1978). "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems" (http://theory.lcs.mit.edu/~rivest/rsapaper.pdf).Communications of the ACM 21 (2): 120–126. doi:10.1145/359340.359342 (http://dx.doi.org/10.1145%2F359340.359342). http://theory.lcs.mit.edu/~rivest/rsapaper.pdf.

[3] Preetida Vinayakray-Jani, "Security within Ad hoc Networks", Position paper, PAMPAS Workshop, Sept. 16/17 2002, London

[4] Sudip Misra l Isaac Woungang, Subhas Chandra Misra, Editors, "Guide to Wireless Ad Hoc Networks," Springer-Verlag London Limited 2009

[5] Carlos T. Calafate, Juan-Carlos Cano, Pietro Manzoni, Manuel P. Malumbres, "A QoS architecture for MANETs supporting real-time peer-to-peer multimedia applications",, Polytechnic University of Valencia (UPV), Valencia, Spain, ISM.2005.18

[6] Juan-Carlos Cano,Pietro Manzoni, "A Performance Comparison of Energy Consumption for Mobile Ad Hoc Network Routing Protocols", Universidad Polit?cnica de Valencia, MASCOT.2000.876429

[7] Panagiotis Papadimitraos and Zygmunt J. Hass, "Securing Mobile Ad Hoc Networks", in Book The Handbook of Ad Hoc Wireless Networks (Chapter 31), CRC Press LLC, 2003.

[8] Rajorshi Biswas, shibdas Bandyopadhyay, Anirban Banarjee, "A fast implementation of the RSA algorithm using the GNUMP library", national workshop on cryptography 2003, at Anna University, Channai.

Segment-aware Cooperative Caching for Peer-assisted Media Delivery Systems

Chamil Kulatunga and Dmitri Botvich

[1]*School of Computing, Asia Pacific Institute of Information Technology, Union Place, Colombo 02, Sri Lanka*
E-mail: chamil@apiit.lk
[2]*Telecommunications Software and Systems Group, Waterford Institute of Technology, Cork Road, Waterford, Ireland*
E-mail: dbotvic@tssg.org

Abstract
The quantity and the size of video contents were exponentially grown in the recent years. Griping this burden, video delivery systems are also evolving into different business directions like wall-garden IPTV networks and managed TV services over the public Internet such as GoogleTV and AppleTV. Therefore the limited network resources should be used effectively to minimise network congestion and to improve Quality of Experience (QoE) to the end-users. Such networks typically use a proxy-based architecture and it has recently gained an added advantage as an energy saving communication paradigm. Segment-based cooperative proxy caching and peer-to-peer networking has been used to enhance the performance and the capacity of such systems. However, the unreliable nation of the peer nodes has not been integrated into the caching algorithms. This paper proposes an efficient caching algorithm for proxy-assisted video delivery systems with peer-support. It uses the segment number in the cache replacement polices. Early segments are cached in the reliable proxy servers while the late segments are accessed from the unreliable peer nodes. We have simulated the new algorithm for performance evaluations and the results demonstrate that the new approach contributes to a significant performance improvement in such content delivery systems.

Keywords: video delivery, IPTV, proxy-assisted systems, cooperative caching, peer-to-peer, segment-aware

Introduction

Proxy-assisted hierarchical video delivery systems [1] have recently gained a significant attention from the research community as an energy saving communication paradigm [2]. Improvement of efficiency of such content delivery systems is also a main research objective owing to the video contents are becoming bulky with High Definition (HD) and Three-Dimensional (3D) videos. At the same time production and distribution of high quality videos have turned into effortless with the availability of inexpensive production equipments and widespread deployments of broadband technologies. As a solution to minimise concurrent streaming sessions at the central servers (reducing the required computing power and network bandwidth), such systems use proxy servers located closer to the clients. They cache extremely popular contents expecting high hit-rates. Those proxy servers also work together in a cooperative manner to increase the cache space and fair load balancing [3].

However, the high-end proxy servers could be expensive and difficult to locate and maintain in the remote sites. As a result, they may provide limited network resources at a high cost. At the same time peer-to-peer networks are becoming well accepted and are on the way making it more conservative and network friendly mechanism to be accepted by the ISPs [4]. Usually peer nodes have abandoned computing power and storage space. Therefore combination of proxy-assisted content delivery networks with peer-support becomes a natural candidate to improve the system capacity and ultimately the Quality of Experience (QoE) to the end-users.

This architecture has already been realised in some IPTV networks with Set-Top-Boxes (STB) as peer nodes providing immense storage space to the video delivery system (Figure 1). Video Head Office (VHO) acts as the central content management server with the accessibility to all the video libraries. Video Sub Offices (VSO) located regionally are functioned as proxy servers, which are not able to store all the video libraries available. In general IPTV networks are wall-garden networks while the networks downwards from the VSO may use the public broadband access network via the Digital Subscriber Loop Access Multiplexers (DSLAM).

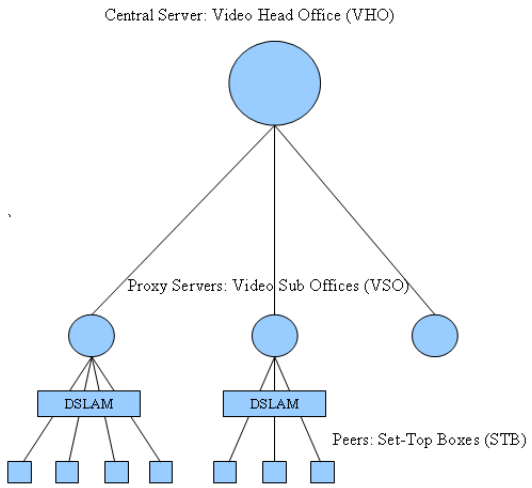


Figure 1: Proxy-assisted hierarchical content delivery architecture

Bulkiness of multimedia contents is an anxious parameter for caching algorithms in making replacement decisions of the proxy-assisted systems. Therefore partial caching (i.e. segmentation of contents) is necessary in multimedia in oppose to web caching and several algorithms have been proposed in the literature [5]. However, there are no much research works done to propose optimum algorithms for proxy-assisted networks with peer-support.

Even though the peer nodes have excessive computing power and storage space, they are loosely coupled to the network providing only an unreliable service. Considering this as a major performance factor we propose a caching algorithm for hybrid delivery systems using the segment number as a key parameter. Predominantly the early segments will be served by the reliable (but restricted) proxy servers while the late segments will be served in the background by the unreliable peers. With the developments of high-bandwidth broadband technologies like ADSL2+, Fibre To The Home (FTTH) this becomes an effective mechanism in content deliveries since peer-churns cannot be compensated with the high bandwidth.

When a late segment (being downloaded from a peer node) is interrupted due to a peer-churn, the disturbed segment can still be discovered and acquired from another location. As shown in Figure 2, segment number 8 (S_8) is downloaded from a peer node while the early segments are downloaded (from the proxy server) and played. Since the segment 7 (S_7) has no enough resources to download in the first try, it will be downloaded with a time lag. But it still has an adequate time to complete the download before the playback point.

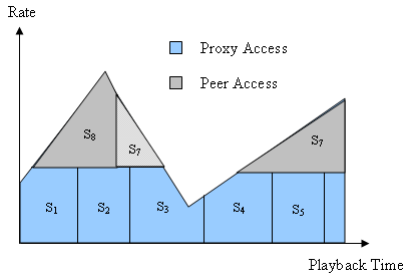


Figure 2: Reliable access form the proxy servers and unreliable access from the peers

The rest of the paper is organised as follows. Section II presents the related works and Section III discusses cooperative caching algorithms used in PROP (Collaborating and coordinating PROxy and its P2P clients) protocol. Section IV details the elements of our proposed algorithm. Simulation results are presented in Section V and Section VI concludes the paper.

Related Works

Due to the large size of multimedia contents, partial caching has become essential and there are two domains a content can be partitioned [5]. One is in space (quality) domain. Layered media caching, multiple version caching and hotspot caching [6] are some of the proposals we can find in the literature. However, due to the complexity of content arrangements and uncertainty of viewing the entirety of a content, content partitioning in time-domain has become more victorious than quality-domain.

Prefix caching [7] caches the prefix of a content at proxy servers and the rest is delivered from the central media server. This method significantly reduces the start-up delay at the end-users. A proxy caching algorithm for video contents with uniform segments has been first proposed by R. Rejaie et al. [8]. K. Wu et al. [9] proposes exponential segments, which minimises the control overhead of equal segments. Exponential segments also support that many users only watch the early parts of a video stream. Lazy segmentation has been proposed by S. Chen et al. [10] as a more intelligent and adaptive approach for segment-based caching.

Caching algorithms having cooperation among the proxy servers (cooperative caching) have improved performance of hierarchical content delivery systems compared to greedy caching [3]. Some proposals suggest dividing the storage space into two parts namely cooperative and greedy. Highly popular segments are stored in the greedy section without considering the availability at other proxy servers. Other segments are stored in the cooperative part based on their availability at other proxy servers. If a segment is available at many proxy servers, it owns a lesser priority than a segment which is available only at the concerned proxy server. As a result, less popular segments take a chance to stay within the systems.

PROP [11] has proposed a segment-based cooperative caching algorithm for proxy-assisted content delivery systems with peer-support. This has improved performance of pure proxy-based delivery systems. Peer nodes cooperatively store segments by considering their availability at other peer nodes. However, PROP does not propose to cooperate between the proxy servers. Proxy servers cache the segments simply according to the popularity of the contents. We describe the PROP algorithm and our proposed segment-aware extension in detail in the following sections.

Prop Protocol In Hybrid Systems

Servicing a Media Segment

The PROP protocol has been proposed to apply for a network having the logical topology shown in Figure 1. Proxy servers act as the bootstrap site between the central content management server and the peer nodes. Either the proxy server, a designated peer or a separate server should be used as an index (discovery) server to determine the available locations of the segments within the network (i.e. the content identity and the location identity). When a client needs a segment, it first looks up in its own cache. If the segment is not found in its own, then the client sends a request to the index server. The index server searches its database to identify a location of the requested segment at a peer node. If the segment is available at a peer node, the actual availability of the segment is verified. Then a message is sent to the requested client to receive the segment from the available peer node.

If the index server can not find the segment at a peer node under its overlay domain, the request is forwarded to the proxy

server. At this stage the proxy server looks up the segment in its own cache. If the segment is available at the proxy server, it starts sending the segment to the client. If the segment is not available at the proxy server, then the segment is requested from the central server. The retrieved segment is then delivered to the client and at the same time the segment is cached at the proxy server.

Segment Replacement Policies

Popularity (p) of a segment (either at a proxy server or at a peer node) in PROP is calculated using the equation given below. The important factor here is that it uses a combination of Least Recently Used (LRU) and Least Frequently Used (LFU) approaches. First part represents the frequency of using the segment while the second part represents the freshness of its use.

$$Popularity\ (p) = \frac{S_{sum} / S_0}{T_r - T_0} \quad \min\left(1, \frac{T_r - T_0 / n-1}{t - T_r}\right)$$

Here, T_0 is the time when the segment is first accessed, T_r is the time the segment is last accessed, S_{sum} is the cumulative bytes the segment has been accessed, S_0 is the segment size in bytes and n is the number of requests received for the segment.

PROP applies only the popularity of a segment to replace it at a proxy server. Cooperation between proxy servers has not been implemented in their design. Lowest popularity segments will be removed first from the proxy server.

If only the popularity is used in the caching algorithms to determine which segments to keep and which segments to delete, higher popular segments are cached by all the entities (i.e. proxy servers and peer nodes). Therefore it is highly desirable to have some cooperation between those entities to improve the availability of a moderately popular segment in the lower hierarchy of the distribution tree (i.e. increasing the caching space). PROP implements cooperative caching among peer nodes. It includes the number of replications of a segment in a utility function so that the utility is inversely proportional to the number of replicated segments. Therefore when the numbers of replications are increased its utility value is decreased. As a result, increasing of replications leads pushing such segments towards the bottom of the priority list in the caching algorithm and will be removed first from a peer node.

Therefore the utility (u) function (given below) of PROP is based both on the popularity and the number of replications of the segment within the system.

$$Utility\ (u) = \frac{p_u}{r^{\alpha}}$$

Where $p_u = (\log p - \log p_{\min})(\log p_{\max} - \log p)$, r is the number of replications of the segment in the system, α is a constant (α = 2), p_{min} (minimum popularity of contents) and p_{max} (maximum popularity of contents) are maintained at the proxy server and broadcast to the peers when it is required to update the utility of a segment.

Segment-Aware Cooperative Caching

PROP does not use the segment number of a content as a parameter for caching either at the peer nodes or proxy servers. Therefore early and late segments occupy the storage space randomly. Hence some early segments will not find a space at the proxy servers and will have to stay on the peer nodes. This increases the unreliability of the streaming session when an early segment is accessed from a peer node closer to its playback point. The serving peer could leave (peer-churn) or congested. Therefore it could be better to avoid this uncertainty by acquiring the late segments early (at a lower rate) using excessive bandwidth of the client peers. Uplinks usually have low bandwidths and peer-churns depend on user behaviours. Therefore this unreliability issue can not be solved merely with enhanced technologies but efficient content management strategies are required.

In this protocol we introduce the timeliness (timely divided segment number) within a content as a parameter for caching decisions. It has been designed to store early segments in the reliable proxy servers and late segments in the unpredictable peer nodes. The availability of segments at different locations of the hierarchical delivery system can be illustrated as shown in Figure 3. Since this allows to access early segments at a higher speed and late segments in parallel at a lower speed without affecting the main stream, we recommend to use a novel congestion control algorithms, which are designed to use only the excessive bandwidth. Such protocols are being under standardisation at the Low Extra Delay Background Transport (LEDBAT) working group of the Internet Engineering Task Force (IETF).

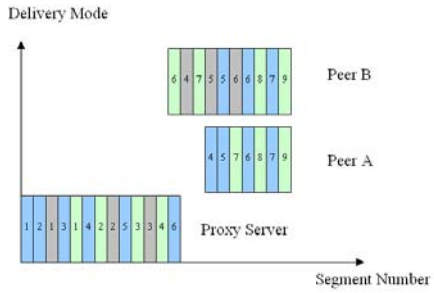


Figure 3: Locations of reliable and unreliable segment caching

In order to achieve the desired performance, we include the segment number (S_i) to the utility function defined in PROP and apply both on the proxy server and the peers. Our algorithm implements the cooperative caching between proxy servers as well. Equation (1) is used as the utility function (u_{proxy}) at the proxy servers for segment replacements. r_{proxy} in the equation (1) is the number of replications of a segment in other proxy servers (i.e. the same hierarchical level).

$$u_{proxy} = \frac{p_u}{r_{proxy}^{\alpha} \log S_i} \qquad (1)$$

Utility function at the peer nodes (u_{peer}) is used to prioritise the segments with late segment numbers. Therefore the defined utility function including the maximum segment number of a selected content (S_{max}) is given in equation (2).

r_{peer} in the equation (2) is the number of replications of a segment in the peer nodes under a proxy server.

$$u_{peer} = \frac{p_u}{r_{peer}^{\ \alpha} \log(S_{\max} - S_i)} \qquad (2)$$

Simulation Results

Simulation Environment

We developed a simulator using C++ to evaluate the performance of the proposed caching algorithm named Segment-aware Cooperative Caching with Peer-support (SCCP). Simulations were conducted in the proxy-assisted hierarchical content delivery network with peer-support (Figure 1). The discrete-event flow-level simulator used 1 minute simulation time granularity. All the results related to the new protocol were compared with PROP. Minimising load at the central content management server and reliability of the streaming session were quantified under different uplink bandwidths and peer-churn rates.

A single VHO (i.e. the central server) has been implemented with 10 VSOs (i.e. proxy servers). We have selected the number of client STBs that could reasonably be served by a single VSO in a small scale IPTV network. The system re-evaluates the necessary parameters in 1 minute intervals in our design for simplicity.

An equal combination of Standard Definition (SD) and High Definition (HD) video contents has been used. The size of the SD video contents were chosen randomly and uniformly distributed between 180 MB and 1080 MB. This corresponds to video content durations between 30 minutes and 3 hours respectively. It represents that 1 minute playback time equals to 6 MB size of stored content. We considered that IPTV delivers professional contents like soap operas starting at half an hour and movies going up to 3 hours. HD content sizes were selected to be four times higher than the SD contents of the same range of duration. We have selected the bandwidth requirement for SDTV and HDTV according to the content size and their duration.

Due to the unpredictability of the available storage space at a STB (i.e. due to pre-recorded contents by the user and sometimes the pre-populated contents by the IPTV provider), available storage spaces were distributed uniformly between 100 MB and 600 MB. Uplink bandwidths at user STBs were selected as a mixture of low-end broadband connections like ADSL, medium level connections like Cable and high-end connections like FTTH. 20% of the STBs were having 1 Mbps downlink and 10 kbps uplink bandwidths, 30% were having 5 Mbps downlink and 20 kbps uplink bandwidths, 30% were having 10 Mbps downlink and 50 kbps uplink bandwidths and remaining STBs were having 50 Mbps downlink and 100 kbps uplink bandwidths.

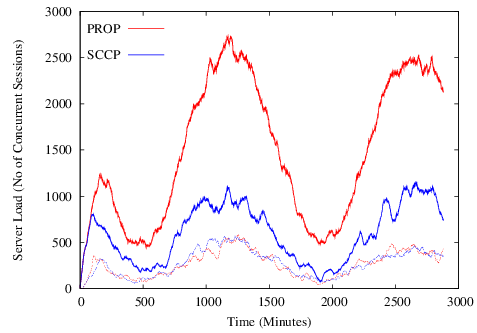
TV watching time of a user per day was selected randomly in between 15 minutes and 6 hours, which approximately represents the widely surveyed average TV watching time of 24 hours per week. Among them, only 10% of the users watched that duration in one-go while the others watched in 2 - 4 spells.

We have selected an approximate sine-shaped daily demand pattern for contents in IPTV for simplicity, which represents peak demand around 7.00 pm and off-peak at 7.00 am. According to this pattern, request generation rates were calculated assuming one user making 10 requests per day on average. The number of active STBs at a time is also selected according to the demand pattern and assuming one user views continuously an average of 1 hour.

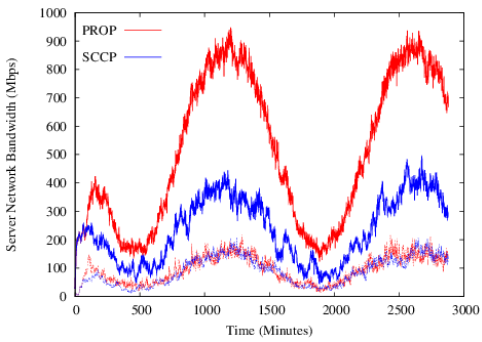
We modelled the popularity of contents using zipf distribution as zipf is found to be approximate content popularity in IPTV networks [12] (skew = 0.5 in our case).

Performance Evaluation of SCCP

We simulated the hybrid system according to the traffic and system specifications presented in the previous section. We measured the loads of the central server and a proxy server in terms of number of concurrent streaming sessions and the used uplink bandwidth. Figure 4 shows the loads at the VHO (using thick continues lines) and a selected VSO (using thin dotted lines) for PROP and SCCP protocols. According to Figure 4(a), SCCP significantly reduces the server load in terms of number of concurrent sessions compared to PROP in peak hours of the day. It also shows that the proxy usage has not been increased. Therefore the number of sessions from the peer nodes should have been increased due to their ability of serving late segments at low rates in SCCP. Figure 4(b) illustrates the same behaviour for the uplink bandwidth usage of the two servers.



(a) Concurrent sessions

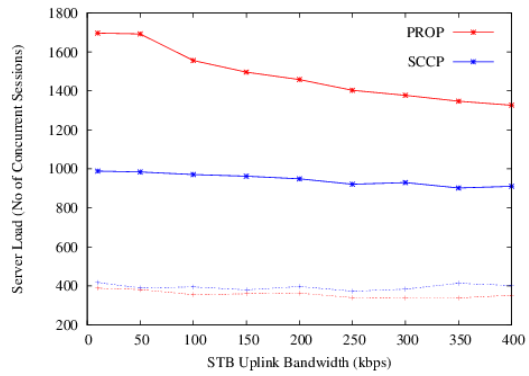


(b) Uplink bandwidths

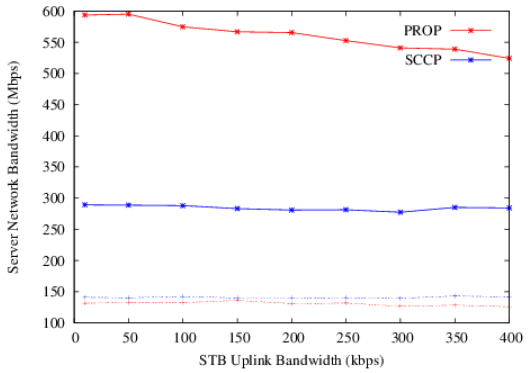
Figure 4: Loads at the central server and a proxy server.

Then we have simulated the system with different STB uplink bandwidths. Figure 5 shows that SCCP gains a consistent performance improvement for all the different STB bandwidths. The load at the central server decreases when the

STB uplink bandwidth becomes lager in the case of PROP. This is because, peers in PROP can serve more segments taking away the load from the central server leaving the unreliability of such sessions on peer-churns.



(a) Concurrent sessions



(b) Uplink bandwidths

Figure 5: Central server and proxy server loads for different STB uplinks

Then we have changed the peer-churn rate (i.e. the probability of a peer to leave while serving another peer) and measured the number of session disruptions by peer-churns. According to Figure 6, SCCP shows a significant improvement over PROP in terms of session disruptions due to the P2P overlay. When the peer-churn rate is increased, the disruptions are increased in both PROP and SCCP.

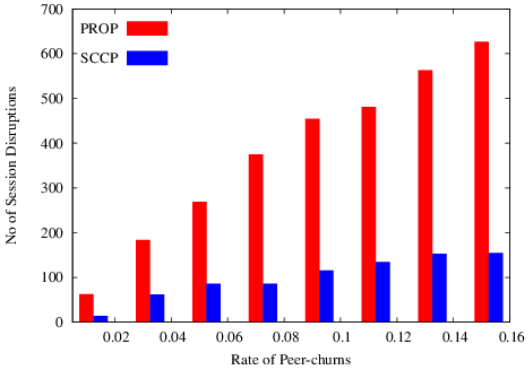


Figure 6: P2P session disruptions for different peer-churn rates

Conclusions

P2P networking paradigm has been recognised by the IETF to re-design as a non-aggressive and ISP-friendly network service in the Internet. Content distribution networks like IPTV have limited proxy capacities to tackle the emerging high quality videos. Also the proxy supported content delivery systems are becoming lucrative in the future due to its energy efficiency. Therefore acquiring the abandoned peer support in such networks is vital. We have developed a cooperative caching algorithm to effectively use peer node capacities minimising the affect of their unreliable nature. We have introduced the segment number as a key parameter to stream early segments from the proxy servers and late segments from the unreliable peer nodes. Simulation results show that it has significantly improved the performance over the PROP protocol which does not consider segment number as a parameter in its cache replacement policies.

Acknowledgement

This work was supported by the projects “FutureComm: Serving Society” funded by Higher Education Authority (HEA) in Ireland under the PRTLI scheme.

References

[1] H. Fabmi, M. Latif, S. Sedigh-Ali, A. Ghafoor, P. Liu, L. H. Hsu, “Proxy Servers for Scalable Interactive Video Support”, IEEE Computer Magazine, vol. 34(9), pp. 54-60, September 2001

[2] Bruce Nordman and Ken Christensen, “Proxying: The Next Step in Reducing IT Energy Use”, IEEE Computer Magazine, vol. 43(1), pp. 91-93, January 2010

[3] Ni Jian and D. H. K.Tsang, "Large-scale Cooperative Caching and Application-level Multicast in Multimedia Content Delivery Networks", IEEE Communications Magazine, vol. 43(5), pp. 98-105, May 2005

[4] J. Seedorf, S. Kiesel, M. Stiemerling, “Traffic Localization for P2P-Applications: The ALTO Approach”, IEEE Conference on Peer-to-Peer Computing (P2P’09), September 2009

[5] J. Liu and J Xu, “Proxy Caching for Media Streaming over the Internet”, IEEE Communications Magazine, vol. 42(8), pp. 88-94, August 2004

[6] Z. L. Zhang et al., “Video Staging: A Proxy Server-based Approach to End-to-end Video delivery over Wide Area Networks”, IEEE Transactions of Networking, vol. 8(4), pp. 429-442, 2000

[7] S. Sen, J. Rexford and D. Towsley, “Proxy Prefix Caching for Multimedia Streams”, IEEE INFOCOM, 1999

[8] R. Rejaie et al., “Proxy caching Mechanism for Multimedia Playback Streams in the Internet”, Web Caching Workshop, 1999

[9] K. Wu, P. S. Yu and J Wolf, “Segment-based Proxy Caching of Multimedia Streams”, ACM World Wide Web Conference, 2001

[10] S. Chen, H. Wang, X. Zhang, B. Shen and S. Wee, "Segment-based Proxy Caching for Internet Streaming Media Delivery", IEEE Multimedia Magazine, July 2005

[11] L. Guo, S. Chen, S. Ren, X. Chen, S. Jiang, "PROP: A Scalable and Reliable P2P Assisted Proxy Streaming System", Conference of Distributed Computing Systems (ICDCS), March 2004

[12] L. Breslau, Pei Cao, Li Fan, G. Phillips and S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications", INFOCOM'99, March 1999

Host Based Intrusion Detection Architecture for Mobile Ad Hoc Networks (MANETs): Proposed Architecture

Sunil Kumar and Kamlesh Dutta*

Department of Computer Science and Engineering
National Institute of Technology, Hamirpur, H.P.-177 005, India
E-mail: sunilkaushik27@gmail.com, kdnith@gmail.com

Abstract
A mobile ad hoc network (MANET) is a self-configuring infrastructureless *network* of mobile autonomous mobile nodes connected by *wireless* links that form a dynamic, purpose-specific, multi-hop radio network in a decentralized fashion. In a MANET, the nodes themselves implement the network management in a cooperative fashion and thus, all the network members share the responsibility for this. The wireless and distributed nature of MANETs in conjunction with the absence of access points, providing access to a centralized authority, poses a great challenge to system security designers and finally makes them susceptible to a variety of active and passive attacks *due to its limited physical security*, ***dynamically changing network topology, energy constrained operations and lack of centralized administration***. Things are getting worst when some nodes getting hijacked or compromised, make this network to stop from the smooth workings as prevention methods (cryptography techniques) alone are not sufficient o make them secure as these methods does not prevent the node from capture therefore detection should be added as another defense before an attacker can violate the working of the system. This paper proposes the host based intrusion detection architecture based on data mining to identify the malicious node and provide security support and we have divided our architecture in to three main modules Audit Data Collection Module, Intrusion Detection Module and Response module to continue the smooth workings of this network.

Keywords: MANETs, Data Mining, IDS, ADP, DM, IE, ES, CM, KE, RM, HID, MDM, ADM, SCM etc.

Introduction

A *mobile ad hoc network* [1, 2, 3, 4, 5,] is a s*elf-configuring infrastructureless* network consisting of mobile nodes (Laptop, Personal Digital Assistants (PDAs) and wireless phones) with routing capabilities where each node operate both as host as well as router to forward the packets to each other, with the characteristics of self-organization and self-configuration which enable it to form a new network quickly as shown in figure 1. MANET has following distinct characteristics:

- Weaker in Security
- Device size limitation
- Battery life
- Dynamic topology
- Bandwidth and slower data transfer rate

Apart from these limitation MANETs has many extensive application like: Military communication and operations, Automated battlefields, Search and rescue operations, Disaster recovery, Policing and fire fighting.

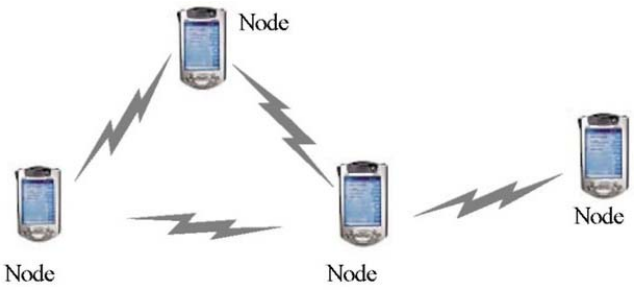


Figure 1: Mobile Ad Hoc Network.

MANETs are highly vulnerable to several types of attacks, due to its characteristics: dynamic topology, infrastructure less, vulnerable of channels, vulnerable of Nodes and lack of strong line of defense. These mobile nodes can be malicious or selfish. Therefore, deploying security in mobile ad hoc networks is important [6].

The first line of defense using approaches such as authentication, access control, encryption, and digital signature by using different verification and encryption methods is to prevent attacks, however, from past experiments have shown that encryption and authentication used as security prevention method are not sufficient So, to resist against attacks, a second wall is needed which is Intrusion Detection (ID) and cooperation enforcement mechanisms that are monitoring activities for policy violation in mobile ad hoc networks. Intrusion detection systems are used to detect misuse and anomalies. These two mechanisms should act together to ensure high security requirements [7].

An intrusion can be defined as process of monitoring activities in a system that attempt to compromise the integrity, confidentiality or availability of a resource with the following functionalities [6, 8]:

- Continually monitor activities (packet traffic or host behavior)
- Automatically recognize suspicious, malicious, or inappropriate activities
- Trigger alarms to system administrator

The intrusion detection system should not introduce a new weakness in the MANET, should run continuously and remain transparent to the system and the users of the system, should use as little of the system resources as possible to detect the intrusions and it must be fault tolerant in the sense that it must be able to recover from system crashes, hopefully recover to the previous state, and resume the operations before the crash [9].

Apart from detecting and responding to intrusions, IDS should also resist subversion. It should monitor itself and detect whether it has been compromised by an attacker. It is desired that there should be fewer false positives and false negatives alarms.

The existing IDS architectures for MANETs fall under three basic categories [10]:

1. Stand-Alone Architecture
2. Cooperative Architecture
3. Hierarchical Architecture

The stand-alone architectures [7, 11, 12] use an intrusion detection engine installed at each node utilizing only the node's local audit.

The cooperative architectures include an intrusion detection engine installed in every node, which monitors local audit data and exchanges audit data and/or detection outcomes with neighbouring nodes in order to resolve inconclusive [12].

The hierarchical architectures amount to a multilayer approach, by dividing the network into clusters. Specific nodes are selected to act as cluster-heads and a more comprehensive engine are running on these nodes and undertake various responsibilities and roles in intrusion detection that are usually different from those of the simple cluster members where only lightweight local intrusion detection engine that performs detection only on local audit data.

Depending upon the detection engine used in the architecture, the intrusion detection engines can be classified in three categories [10]:

1. Anomaly / Behavior-based IDS
2. Signature / Misuse / Knowledge-based IDS
3. Specification based IDS.

1. Anomaly/ Behaviour Based Detection: In this detection system, a baseline profile of normal system activity is created which rely on nodes' behavior in the system. Any activity in the system that is a deviation from the baseline is treated as a possible intrusion.
2. Signature / Misuse / Knowledge Based Detection: In this detection technique, decisions are made on the basis of knowledge/signature of a model of the intrusive process and which rely on a predefined set of patterns to identify attacks.
3. Specification Based detection: Specification-based detection defines a set of constraints that describe the correct operation of a protocol or a program and monitors the execution of the program with respect to the defined constraints. This technique may provide the capability to detect previously unknown attacks with a low false positive rate, when there is deviation of correct operation of a protocol.

Each method has their distinct advantages and disadvantages as well as suitable application areas of intrusion detection.

When considering the area being the source of data used for intrusion detection, another classification of intrusion detection systems can be used in terms of the type of the protected system. There is a family of IDS tools that use information derived from a single host (system) — host based IDS (HIDS) and those IDSs that exploit information obtained from a whole segment of a local network (network based IDS, i.e. NIDS).

Existing Systems

H.–Y. Chang et al.[13] proposed a real-time knowledge-based network intrusion systems which accumulate knowledge about attacks, examine traffic, try to identify patterns indicating that a suspicious activity is occurring and developed specifically for OSPF [10] for detecting the link-state routing protocol attacks. These systems are particularly attractive because of their high accuracy detecting suspicious activity and low false alarm rates. But this approach can be applied against known attack patterns only and the utilized knowledge base needs to be updated frequently to maintain the accuracy of detecting the malicious activity. Farooq et al.[14] proposed a "signature based intrusion detection technique ", in which they assume that they knows the signature of the attack. D. Sterne et al.[15] presented a cooperative intrusion detection architecture that facilitates accurate detection of MANET-specific and conventional attacks. The architecture is organized as a dynamic hierarchy in which detection data is acquired at the leaves and is incrementally aggregated, reduced, and analyzed as it flows upward toward the root. The nodes at the top are responsible for security management functions. Ioanna Stamouli [16] proposed RIDAN architecture which uses timed finite state machine to formally define attack against the AODV routing process. It uses a knowledge based methodology to detect the intrusion. RIDAN operates locally in every participating nodes and observe the network traffic. This model can able to detect resource consumption attack, Sequence number attack and dropping routing packet attack. The finite state machine has been used to detect attacks on the DSR protocol has been proposed by P. Yi, Y. Jiang, Y. Zhong, and S. Zhang in 2005[17].A finite state machine model of local AODV behavior was proposed B. Wang, S. Soltani, J. Shapiro, and P. Tan in 2006[18]. In 2003, Bo Sun, Kui Wu, Udo W. Pooch, presented a IDS model for MANET based on Markov Chains[20].In 2004, A. Cardenas, , V. Ramezani, J.Baras presented a statistical framework based on Hidden Markov Model (HMM) that allows the incorporation of prior information about the normal behaviour of the network and of the network[21].In 2004, A. Patcha and J. Park, proposed an

IDS based on Game Theory which falls under the category of multi-stage dynamic non-cooperative game with incomplete information[19].In 2006, H. Hassan, M. Mahmoud, and S. El-Kassas proposed a specification-based IDS for AODV protocol[22].

From the literature, it was found that

- Most research concentrates on the onstruction of operational IDSs, rather than on the discovery of new and fundamental insights into the nature of attacks and false positives.
- It is very common to focus on the data mining step, while the other KDD steps are largely ignored.
- Much research is based on strong assumptions that complicate practical application.
- Up to now, data mining in intrusion detection focuses on a small subset of the spectrum of possible applications.

In our proposed architecture we are going to add Expert system and data mining module to enhance the capability of the classifier module to detect the defined attack pattern with minimum amount of time as previously mined attack pattern is there and detect the behavior of the system with more accuracy producing high positive alarms and low false alarms.

Proposed Architecture

Data mining is frequently used to designate the process of extracting useful information from large databases. In this view, the term knowledge discovery in databases (KDD) is used to denote the process of extracting useful knowledge from large data sets using data mining as a particular step. Specifically, the data mining step applies so-called data mining techniques to extract patterns from the data. Additionally, it is preceded and followed by other KDD steps, which ensure that the extracted patterns actually correspond to Data Mining for Intrusion Detection useful knowledge. Here, we broadly outline some of the most basic KDD steps [32] :

1. Understanding the application domain: First is developing an understanding of the application domain, the relevant background knowledge, and the specific goals of the KDD endeavor.
2. Data integration and selection: Second is the integration of multiple (potentially heterogeneous) data sources and the selection of the subset of data that is relevant to the analysis task.
3. Data mining: Third is the application of specific algorithms for extracting patterns from data.
4. Pattern evaluation: Fourth is the interpretation and validation of the discovered patterns. The goal of this step is to guarantee that actual knowledge is being discovered.
5. Knowledge representation: This step involves documenting and using the discovered knowledge.

Data mining techniques can be differentiated by their different model functions and representation, preference criterion, and algorithms [25].The main function of the model that we are interested in is classification, as normal, or malicious, or as a particular type of attack [26][27].

It is a fairly recent topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition.

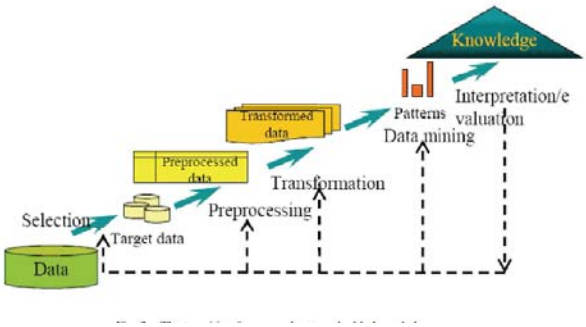


Figure 2: KDD Process [32].

Here are a few specific things that data mining might contribute to an intrusion detection project [22, 23]:

- Remove normal activity from alarm data to allow analysts to focus on real attacks
- Identify false alarm generators and ”bad” sensor signatures
- Find anomalous activity that uncovers a real attack
- Identify long, ongoing patterns (different IP address, same activity)

To accomplish these tasks, data miners employ one or more of the following techniques:

- Data summarization with statistics, including finding outliers
- Visualization: presenting a graphical summary of the data
- Clustering of the data into natural categories
- Association rule discovery: defining normal activity and enabling the discovery of anomalies
- Classification: predicting the category to which a particular record belongs

In our proposed Host based architecture we have added some additional modules (Expert System, Knowledge Expresser, Data Miner and Interference Engine) to enhance the capability of Classifier module as Data mining attempts to extract implicit, previously unknown, and potentially useful information from data, so that it can detect the attacks with minimal amount of time and provide useful information to the ADM and MDM module, to increase the effectiveness and scalability of this proposed architecture. There is one module previously mined data module in the architecture that access the previous stored signatures of different attacks and behavior of abnormal activities at the node, so that it can detect abnormal activity with minimum amount of time. The other modules Expert System, Knowledge Expresser, Data Miner and Interference Engine are used are to expert the system with proper information to classify the abnormal and normal activities with minimum amount of time. The graphical representation is given in Fig. 3.

The advantage of this approach is that it has expert system, Knowledge Expresser, Data Miner and Interference Engine. This means that the chances of one system catching intrusions missed by the other will increase. The below architecture takes care to ensure that the IDS running on each host does not drain resources more than necessary. Here all the modules work collectively at the same time to provide the necessary support for the intrusion detection in the network

We have divided our architecture into following three main modules :

- Audit Data Collection Module
- Intrusion Detection Module
- Response Module

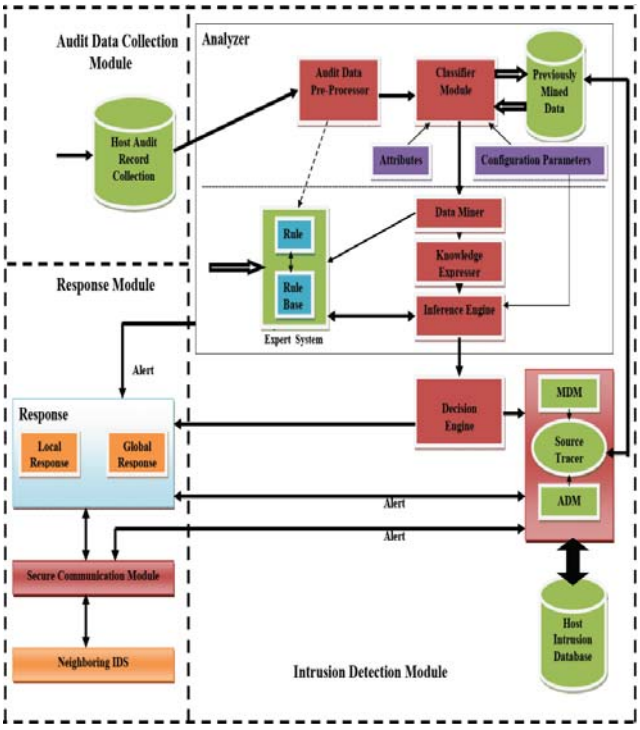


Figure 3: Intrusion Detection System for Mobile Ad Hoc Network: Proposed Architecture

Audit Data Collection Module: This module is to collect the audit records.

i. Host Audit Records Collection (HARC) [28]: - In this sub module, each operation of a host should be recorded to check that whether an intrusion is taking place. HARC is responsible for collecting useful information to minimize the volume of the audit data, responsible for gathering and storing not to processing it. This module usually passes the audit records to the Audit Data Preprocessor.

Intrusion Detection Module: This module play very important role to analyze the activities to determine that any of the activity is violating the security rules on the host after taking the audit data record from Audit Data Collection Module. Once an IDS determines that an unusual activity or an activity that is known to be an attack occurs, it then generates an alert to the response module.

i. Audit Data Preprocessor (ADP)[33]: - This refers to one or more individual preprocessors used by IDS to the collect audit data and to quantify and transform into the appropriate data format for the subsequent module. Some records of the ongoing activity of the users must be maintained to provide as input to the intrusion detection system. In this detection module specific audit Data are created from the host audit records. Each record contains the fields like subject, actions, object, exception condition, resource-usage, time-stamp etc.

ii. Classifier Module (CM):- In this module we can employ classification algorithms in order to perform intrusion detection in MANETs by mapping a data item into one of several pre-defined categories. Compared to other methods, classification algorithms have the advantage that they are largely automated and that they can be quite accurate and gather sufficient ``normal" and ``abnormal" audit data for a user or a program, then apply a classification algorithm to learn a classifier that will determine (future) audit data as belonging to the normal class or the abnormal class [34].

iii. Data Miner(DM) [29, 30]: - Data Mining is a branch of computer science, is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. DM module is used to automatically discover new patters from a large amount of the audit data. Association rules have been successfully used to mine audit data to find normal patterns for anomaly intrusion detection.

iv. Knowledge Expresser (KE)[31]:- Knowledge Expresser from Classsifier Module, in its most fundamental form, is to extract interesting, nontrivial, implicit, previously unknown and potentially useful information from Classifier Module.

v. Expert Systems (ES): These systems are modeled in such a way as to separate the rule matching phase from the action phase.

vi. Inference Engine (IE) [36]: Inference engine is a computer program that attempt to infer or derive a deep insights or answers from the knowledge base. The inference engine is used to evaluate each short sequence to find anomalies by combining multiples sequence parameters by taking configurations parameters as input and applying the rules from the expert system. Moreover, the inference engine is considered as the brain of the expert systems which has the ability of reasoning depending on the methods applied or used [35]. This uses a set of parallel soft computing based classifiers for detecting abnormal behaviors of network data.

vii. Decision Engine (DE): Decision engines refer to software that can be programmed to make decisions based on a number of identifiable factors.

viii. Misuse Detection Module (MDM) [37]: - Detection is performed by looking for the exploitation of known weak points in the system, which can be described by a specific pattern or sequences of events or the data (the "signature" of the intrusion). Here the collections of signatures (representative patterns) define the known attacks. The primary purpose of MDM is only to identify the known patterns of attacks that are specified in the local intrusion database.

ix. Anomaly Detection Module (ADM) [37]: -Each ADM is responsible for detecting a different type of anomaly. There can be many ADM modules based on the complexity of the IDS architecture. In this architecture ADM will analyze data, compare with known profile which already defined, run the statistical analysis to determine if any deviation is significant, and flag the events as a true attack state, false attack state, or normal state. If it finds a false positive, then profile must be updated to reflect the results.

x. Host Intrusion Database (HID) [37]: - is a database maintaining in the nodes that warehouses all the information necessary for the IDS, such as the signatures of known attacks, the established patterns of users and resource usage and the normal volume of data flow in the network.

Response Module: If an intrusion is detected by the Detection Engine then the Response Engine is activated by issuing the alerts from Intrusion Detection Module[38]. The Response Engine is responsible for sending a local and a global alarm in order to notify the nodes of the mobile ad hoc network about the incident of intrusion.

i. Secure Communication Module (SCM): It is necessary to enable IDS to communicate with other IDS on other nodes with effective data encryption and decryption technique. It will allow the MDM s and ADM s to use cooperative algorithms to detect intrusions. It may also be used to initiate a global response when an IDS or a group of IDS detects an intrusion on the basis of aggregation of local response from different nodes. Basically, any communication that needs to occur from one IDS to another will use the Secure Communication Module.

This architecture is concerned with what activity is happening on each host. This architecture is ideal if it is able to detect actions and other activities with high confidence. In order to function properly, IDS has to be installed on every node in the network to processes and perform analysis on the audit data gathered locally, at the expense of the already limited resources on the hosts.

Conclusion and Future Work

In this paper we have discussed several new issues and ideas that must be addressed when designing intrusion detection systems for mobile ad hoc networks. There are still internal and insider attacks that utilize software vulnerability, even if the prevention schemes are perfect and implemented correctly. The intrusion detection system should be designed in such a way that it can work independently without minimum human supervision and provide a necessary level of protection to the node and network. Furthermore, the main characteristics of the architecture is modular Architecture that allows it to be properly configured, easily extended and modified, either by adding new components, or by replacing some of existing components when they need to be updated. Future work includes implementation of such IDS architecture and testing its effectiveness in mobile ad hoc networks environments. Further to increase the effectiveness and scalability of this proposed architecture, provide useful information to the each module of the architecture. Through continuing investigation, it can be shown that this architecture is well suited for better intrusion detection in wireless ad hoc network that are distributed and co-operative in nature.

References

[1] C. Perkins, “Ad hoc Networks, ” Addison-Wesley, 2001.

[2] M. Ilyas, “The Handbook of Ad Hoc Wireless Networks, ” CRC Press, 2003.

[3] C.K.Toh, “Ad Hoc Mobile Wireless Networks: Protocols and Systems, ” Prentice Hall Englewood Cliff, NJ 07632, 2002

[4] C. Murthy and B.Manoj, “Ad hoc Wireless Networks: Architectures and Protocols, ” Prentic Hall PTR, 2005.

[5] IETF MANET Working Group. Mobile Ad Hoc Networks (MANET). Working Group, Charter available at http://www.ietf.org/html.charters/ manet-charter.html.

[6] Zhang Y, Lee W, Huang Y., “Intrusion Detection Techniques for Mobile Wireless Networks”, ACM/Kluwer Wireless Networks Journal (ACM WINET), 2003, 9, pp 545–56.

[7] Mishra A, Nadkarni K, Patcha A., “Intrusion Detection in Wireless Ad Hoc Networks”, in Proceeding IEEE Conference on Wireless Communications, 2004, 11, pp 48-60.

[8] Chaki R, Chaki N., “IDSX: A Cluster Based Collaborative Intrusion Detection Algorithm for Mobile Ad-Hoc Network”, in Proceeding IEEE 6th International Conference on Computer Information System and Industrial Management Application (CISIM'07), 2007, pp 179-84.

[9] Amitabh Mishra, “Security and Quality of Service in Wireless Ad hoc Networks”, Cambridge University Press, February 2008.

[10] Anantvalee T, Wu J. A survey on intrusion detection in mobile ad hoc networks, wireless/mobile network security. Springer;2006. Chapter 7, p. 170e196.

[11] Jacoby GA, Davis NJ. Mobile host-based intrusion

detection and attack identification. IEEE Wireless Communications August 2007; 14(4): 53e 60.

[12] Lauf A, Peters RA, Robinson WH. A distributed intrusion detection system for resource-constrained devices in ad hoc networks Ad Hoc Networks May 2010;8(3): 253e66.

[13] H.–Y. Chang, S.F. Wu and Y.F. Jou, "Real-Time Protocol Analysis for Detecting Link-State Routing Protocol Attacks", ACM Tran. Inf. Sys.Sec., 1, Pp. 1-36, 2001.

[14] Farooq Anjum and Dhanant Subhadrabandhu and Saswati Sarkar "Signature based Intrusion Detection for Wireless Ad-Hoc Networks: A Comparative study of various routing protocols" Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th, Oct. 2003.

[15] D. Sterne, P. Balasubramanyam, D. Carman, B. Wilson, R. Talpade, C. Ko, R. Balupari, C.Y. Tseng, T. Bowen, A general cooperative intrusion detection architecture for MANETs, , in: 3rd IEEE Int'l Workshop on Information Assurance, College Park, MD, March 2005, pp. 57–70.

[16] Ioanna Stamouli, Patroklos G. Argyroudis, Hitesh Tewari, "Real-Time Intrusion Detection for Ad Hoc Networks", Proceedings of the Sixth IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks, 2005, pp. 374-380.

[17] P. Yi, Y. Jiang, Y. Zhong, and S. Zhang, Distributed Intrusion Detection for Mobile Ad hoc Networks, Proceedings of the 2005 Symposium on Applications and the Internet Workshops (SAINTW' 05), pp.94-97.

[18] B. Wang, S. Soltani, J. Shapiro, and P. Tan, Local Detection of Selfish Routing Behavior in Ad Hoc Networks, ispan, pp. 392-399, 8th International Symposium on Parallel Architectures, Algorithms and Networks (ISPAN'05), 2005.

[19] A. Patcha and J. Park, A game theoretic approach to modeling intrusion detection in mobile ad hoc networks, 2004 IEEE Workshop on Information Assurance and Security, June 2004.

[20] Bo Sun, Kui Wu, Udo W. Pooch, Alert aggregation in mobile ad hoc networks, Workshop on Wireless Security 2003: pp. 69-78.

[21] A. Cardenas, , V. Ramezani, J.Baras, HMM Sequential Hypothesis Tests for Intrusion Detection in MANETs Extended Abstract, Tech rept. 2003.

[22] H. Hassan, M. Mahmoud, and S. El-Kassas, "Securing the AODV Protocol Using Specification- Based Intrusion Detection, " the 2nd ACM International Workshop on QoS and Security for Wireless and Mobile Networks, October 2—6, 2006, Torremolinos, (Malaga), Spain, p.p. 33-36, ISBN: 1-59593-486-3.

[23] W. Lee, S.J. Stolfo, K.W. Mok. "A Data Mining Framework for Building Intrusion Detection Models". IEEE Symposium on Security and Privacy (Oakland, California), 1999

[24] G. Florez, S.M. Bridges, and R.B. Vaughn, "An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection". The North American Fuzzy Information Processing Society Conference, New Orleans, LA, 2002.

[25] Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data, " Communications of the ACM 39 (11), November 1996, 2734.

[26] Ghosh, A. K., A. Schwartzbard, and M. Schatz, "Learning program behavior profiles for intrusion detection", In Proc. 1st USENIX, 9-12 April, 1999

[27] Kumar, S., "Classification and Detection of Computer Intrusion", PhD. thesis, 1995, Purdue Univ., West Lafayette, IN.

[28] F. H. Wai, Y. N. Aye, and N. H. James, "Intrusion Detection in Wireless Ad-Hoc Networks, " Term Paper, School of Computing, National University of Singapore, 2003

[29] W. Lee, S.J. Stolfo, K.W. Mok. "A Data Mining Framework for Building Intrusion DetectionModels". IEEE Symposium on Security and Privacy (Oakland, California), 1999.

[30] G. Florez, S.M. Bridges, and R.B. Vaughn, "An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection". The North American Fuzzy Information Processing Society Conference, New Orleans, LA, 2002.

[31] D. Wilson and D. Kaur. Knowledge extraction from kdd'99 intrusion data using grammatical evolution.WSEAS Transactions on Information Science and Applications, 4:237–244, February 2007.

[32] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Elsevier, 2006.

[33] R. Nakkeeran, T. Aruldoss Albert and R. Ezumalai, "Agent Based Efficient Anomaly Intrusion Detection System in Ad hoc Networks", IACSIT International Journal of Engineering and Technology Vol. 2, No.1, February, 2010.

[34] A. Mitrokotsa, M. Tsagkaris and C. Douligeris, "Intrusion Detection in Mobile Ad Hoc Networks Using Classifica-tion Algorithms, " IFIP International Federation for Infor-mation Processing, Palma de Mallorca, 2008, pp. 133-144.

[35] N.L. Griffin and F.D. Lewis, "A rule-based inference engine which is optimal and VLSI implementable", IEEE Int'l Workshop on Tools for AI Architectures, Languages and Algorithms, Oct. 23-25, 1998, pp. 246-251.

[36] Mahmoud Jazzar and Aman Jantan. A Novel Soft Computing Inference EngineModel for Intrusion Detection. IJCSNS International Journal of Computer Science and Network Security, Vol. 8, No. 4, April 2008.

[37] T. Anantvalee and J. Wu. "A Survey on Intrusion Detection in Mobile Ad Hoc Networks", Book Series Wireless Network Security, Springer, pp. 170 – 196, ISBN: 978-0-387-28040-0 (2007).

[38] Y. Zhang, W. Lee, and Y. Huang, "Intrusion Detection Techniques for Mobile Wireless Networks, " *Wireless Networks*, vol.9, no. 5, 2003, pp.545–56.

Cooperative Caching Strategies for MANETs and IMANETs

Atul Rao, Prashant Kumar and Naveen Chauhan

Department of Computer Science and Engineering
National Institute of Technology, Hamirpur, India
E-mail: raonithmr@gmail.com, prashantkumar32@gmail.com, naveenchauhan.nith@gmail.com

Abstract
Caching in mobile computing environment is a capable technique that can improve data access performance and reduce the heavy communication between client and server. Cooperative caching allows sharing and coordination of cached data between mobile hosts. In this paper all the caching schemes for wired, Ad hoc and internet mobile Ad hoc network are discussed and compared in terms of cache resolution and cache replacement policy.

Keywords: Cooperative caching, Internet-based mobile adhoc network, cache

Introduction
The term Mobile Adhoc Networks (MANETs) refers to a multi hop packet based wireless network composed of a set of mobile nodes that can communicate and move at the same time, without using any kind of fixed wired infrastructure. MANETs are actually self-organizing and adaptive networks that can be formed and deformed on-the-fly without the need of any centralized administration. The main two characteristics of MANETs are mobility and multi hop communication. MANETs usage areas:

- Military scenarios
- Sensor networks
- Rescue operations
- Students on campus
- Free Internet connection sharing
- Conferences

With the recent advent in wireless technologies and mobile devices, wireless networks have become a ubiquitous communication infrastructure. In addition, growing interest in accessing the wired network or Internet has fueled the development of mobile wireless networks, which can be used in many realistic applications. It is envisaged that in the near future, users will be able to access the Internet services and information anytime and anywhere. Internet-based mobile Ad hoc networks (IMANETs) [3, 8] are an emerging technique that combines a wired network (e.g. Internet) and a mobile Ad hoc network (MANETs) for developing a ubiquitous communication infrastructure. Internet-based Mobile Ad hoc Networking is a technology that supports self-organizing, mobile networking infrastructures, and is one which appears well-suited for use in future commercial and military applications. Thus, to put the MANET technology into the context of real life, we consider an Internet-based MANET (IMANETs), which is an evolving communication infrastructure that combines the wired Internet and wireless mobile Ad hoc networks. IMANET is getting more attention and is applied to realistic Internet applications because of its flexible accessibility and information availability. The followings are some of the applicable uses for IMANETs:

Case1: During events such as Olympic Games the demand from users to access the Internet and communicate among themselves are very high. While a fixed infrastructure may be in place, it is challenging to accommodate all the users due to limited wireless bandwidth. With an IMANET, users can either access the required information directly or indirectly (through relays).

Case2: In a battle field or emergency site, one MT may be connected to the Internet by a satellite and serve as a proxy for other MTs. The accessed information and services can be shared by the other MTs via local Ad hoc communication.

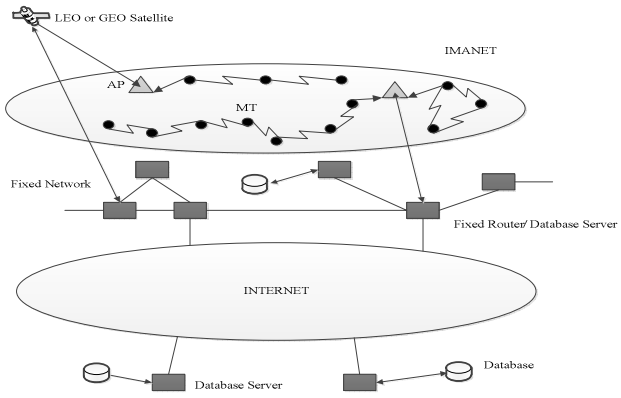


Figure 1: System Model for IMANET.

Caching is an important technique to improve the performance of wireless or wired network. Aim of various caching schemes is to reduce the web traffic, congestion control, bandwidth management and data accessibility in internet and in IMANETs. The rest of paper is organized as follows. Section II describes the different cooperative schemes in wired and adhoc networks. Section III describes the comparison of various caching scheme and section IV gives

overview of IMANETs and available caching schemes. Section V concludes the paper.

Cooperative Caching

A number of data replication schemes and caching schemes have been proposed in order to facilitate data access in mobile Ad hoc networks (MANETs). Data replication studies the issue of allocating replicas of data items to meet access demands. These techniques normally require a priori knowledge of the network topology. Caching schemes however do not facilitate data access based on the knowledge of distributed data items. In Simple Cache the requested data item has always been cached by the requester node. The node uses the cached copy in order to serve subsequent requests when they arrive. The requester node has to get the data from the data center in case of cache miss. However increasing the hop distance between the requester node and caching node will increase the response time for the request. In the research area of mobile Ad hoc networks have been developed a number of caching protocols.

Why Cooperative Caching?

Caching of frequently accessed data in multi-hop adhoc environment is a potential technique that can improve the data access performance and availability. Cooperative caching allows the sharing and co-ordination of cached data among different clients and groups. Due to mobility and constraints on resources like bandwidth, computational resources and limited battery power in mobile adhoc networks, some cooperative cache management schemes need to be designed.

Cooperative Caching in Wired Network:

Squirrel Cache

Squirrel is a decentralized peer-to-peer web cache [15]. It is scalable, self-organizing and resilient to peer failures. Without the need for additional hardware or administration it is able to achieve the functionality and the performance of a traditional centralized web cache. It is proposed to run in a corporate LAN type environment, located e.g. in a building, a single geographical region. Squirrel is build up on Pastry, an object location and routing protocol for large scale peer-to-peer systems, which provides the mentioned features. The goals and the motivation for web caching are decline of load on external web servers, corporate routers and of course external traffic (traffic between the corporate LAN and the internet) which is expensive, especially for large organizations. Squirrel is a possibility to achieve these goals without the use of a centralized web cache or even clusters of web caches.

Squirrel uses Pastry as a location and routing protocol. When a client1 requests an object it first sends a request to the Squirrel proxy running on the client's machine. If the object is unreachable then the proxy forwards the request directly to the origin Web server. Otherwise it checks the local cache. If a fresh copy of the object is not found in this cache, then Squirrel tries to locate one on some other node. To do so, it uses the distributed hash-table and the routing functionalities provided by Pastry. First, the URL of the object is hashed to give a 128-bit Object Id from a circular list. Then the routing procedure of Pastry forwards the request to the node with the Node Id (assigned randomly by Pastry to a participating node) numerically closest to the Object Id. This node then becomes the home node for this object. Squirrel then proposes two schemes named:

- Home-store
- Directory schemes

The home-store scheme is the one to be used in reality and approaches the performance of a centralized web cache with infinite storage while using e.g. 100 MB cache size per node. Disadvantages of Squirrel Cache:

- In decentralized caching churn arises from continued and rapid arrival and failure (or departure) of a large number of participants in a peer-to-peer system. This will increase host loads and block a large fraction of normal insert and lookup operations in the system.
- There is no cooperation between the mobile hosts or peers.

Summary Cache

In the summary cache scheme [5], each proxy stores a summary of its directory of cached document in every other proxy. When a user request misses in the local cache, the local proxy checks the stored summaries to see if the requested document might be stored in other proxies. If it appears so, the proxy sends out requests to the relevant proxies to fetch the document. Otherwise, the proxy sends the request directly to the Web server. The key to the scalability of the scheme is that summaries do not have to be up-to-date or accurate. A summary does not have to be updated every time the cache directory is changed; rather, the update can occur upon regular time intervals or when a certain percentage of the cached documents are not reflected in the summary. The sharing of caches among Web proxies is an important technique to reduce Web traffic and alleviate network bottlenecks A summary only needs to be inclusive (that is, depicting a superset of the documents stored in the cache) to avoid affecting the total cache hit ratio. That is, two kinds of errors are that may occur:

False misses: The document requested is cached at some other proxy but its summary does not reflect the fact. In this case, a remote cache hit is not taken advantage of, and the total hit ratio within the collection of caches is reduced.

False hits: The document requested is not cached at some other proxy but its summary indicates that it is. The proxy will send a query message to the other proxy, only to be notified that the document is not cached there. In this case,a query message is wasted.

The errors affect the total cache hit ratio or the inter proxy traffic, but do not affect the correctness of the caching scheme. For example, a false hit does not result in the wrong document being served. In general we strive for low false misses, because false misses increase traffic to the Internet and the goal of cache sharing is to reduce traffic to the Internet.

Two factors limit the scalability of summary cache: the network overhead (the interproxy traffic), and the memory required storing the summaries (for performance reasons, the summaries should be stored in DRAM, not on disk). The network overhead is determined by the frequency of summary

updates and by the number of false hits and remote hits. The memory requirement is determined by the size of individual summaries and the number of cooperating proxies. Since the memory grows linearly with the number of proxies, it is important to keep the individual summaries small.

In summary cache, cache sharing under finite cache sizes, a number of schemes are explored for the evaluation of summery cache. Different scheme are as under:

- No Cache Sharing.
- Simple Cache Sharing
- Single-Copy Cache Sharing
- Global Cache

Two questions are answered whether simple cache sharing significantly reduces traffic to Web servers, and whether the more tightly coordinating schemes lead to a significantly higher hit ratio. Here the hit ratio includes both local hits and remote hits. Local hits are those requested documents found in the proxy's cache; remote hits are those documents found in the neighboring proxies 'cache. Both kinds of hit avoid traffic to web servers. Impact of update delays is also studied in the summary cache.

Cooperative Proxy Caching: Mainly a cache is defined as a fast, temporary store for commonly used items. High speed memory units on microprocessor chips cache data from main memory; main memory units cache sections of disk files; and local disks cache documents from the network file server. Cooperating proxy caches [4] are groups of HTTP proxy servers that share cached objects. Client caching on the Web succeeds at the browser level (L1) and the local proxy level (L2) for the same reasons that memory and disk caches succeed; specifically

- Local disk and network transfers are faster than remote network transfers from the Web server
- Web documents are often re-requested by the same user or by other users on the same local network .

At the next level of Web caching (L3), proxy servers cooperate to share their cached documents. If a cache miss occurs at the local proxy server, that proxy may forward the request to a remote proxy instead of to the origin Web server. However, the viability of cooperative proxy caching is unclear because, unlike browser and local proxy caches, a remote proxy is not inherently faster than the origin Web server. Retrieving objects from either a remote cache or a remote server involves wide-area network transfers, and there is no guarantee that the remote cache will have faster server hardware, higher bandwidth, or better routes to the client. When cache and server have similar resources and communication costs, cooperative proxy caching actually increases response time due to cache overhead. In cache hierarchies, L3 proxy caches are separate physical sites such as national caches. In cache meshes, the same site serves as a local L2 proxy cache to its clients and as a remote L3 proxy cache to its cooperative peers. Cooperating caches have three separate functions:

- Discovery
- Dissemination
- Delivery of cache objects

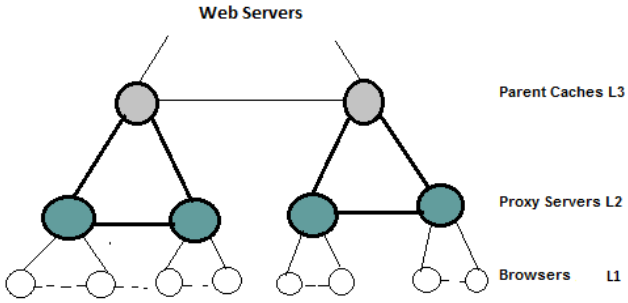


Figure 3: Cooperating proxy caches organized as a mesh a hierarchy

Discovery refers to how a proxy locates cached objects. Dissemination is the process of selecting and storing objects in the caches. Delivery defines how objects make their way from the Web server or a remote proxy cache to the requesting proxy.

Cooperative Caching in MANETS

Push and Pull Based Approach

These two are the basic cache sharing techniques in MANET [1, 4]. With push-based cache sharing, when a node acquires and caches a new data, it actively advertises the caching event to the nodes in the neighborhood. Mobile nodes in the vicinity will record the caching information upon receiving such an advertisement and use it to direct subsequent requests for the same item. This scheme enhances the usefulness of the cached contents. The cost we have to pay is the communication overhead for the advertisement; an advertisement is useless if no demands for the cached item arise in the neighborhood. In the push-based scheme, the caching information known to a node may become obsolete due to node mobility or cache replacement. The pull-based approach may overcome this problem. Withpull based cache sharing, when a mobile node wants toaccess a data item that is not cached locally it will broadcast a request to the nodes in its vicinity. A nearby node that has cached the data will send a copy of the data to the request originator (a pull operation) unlike pushing; pulling allows the node to utilize the latest cache contents.

Zone Based Cooperative Caching Scheme

In Zone Cooperative caching scheme[1, 12], one hop neighbor of a mobile client form a cooperative cache zone since the cost for communication with them is low both in terms of energy consumption and message exchange. Zone cooperative caching scheme is for data retrieval in mobile Ad hoc networks. The design rationale of ZC caching is that it is considered advantageous for a client to share cache with its neighbors lying in the zone (i.e., mobile hosts that are accessible in one hop) Mobile hosts belonging to the zone of a given host then form a cooperative cache system for this host . In ZC caching Cache discovery is a problem. During Cache Discovery Process When a data request is initiated in an MH, it first looks for the data item in its own cache. If there is a local cache miss, the MH checks if the data item is cached in other MHs within its home zone. When an MH receives the

request and has the data item in its local cache (i.e., a zone cache hit), it will send a reply to the requester to acknowledge that it has the data item. In case of a zone cache miss, the request is forwarded to the neighbor along the routing path. If the data item is not found on the zones along the routing path (i.e., a remote cache miss), the request finally reaches the data source and the data source sends back the requested data.

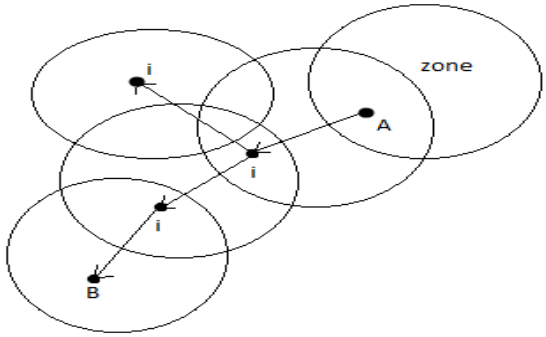


Figure 2: Cache Discovery Process in ZC caching.

COOP-A Cooperative Caching service in MANET

It is a novel cooperative caching scheme for on-demand data access applications in MANETs. The objective is to improve data availability and access efficiency by getting together local resources of mobile nodes. The cooperation of caching nodes is twofold. First, a caching node can answer the data requests from other nodes. Second, a caching node stores the data not only on behalf of its own needs, but also based on other nodes' needs. COOP addresses two basic problems for cooperative caching in MANETs:

Cache resolution – Cache Resolution addresses how to restore a data request with minimal cost of time, energy and bandwidth. COOP'S cache resolution is a cocktail scheme, which consists of three basic schemes:

Adaptive Flooding: It calculates proper flooding range based on the cost to fetch the requested data. Limited flooding is used for cache resolution, not only because it has potential to discover the closest cache around the requested , but also because flooding can serve as an announcement in the neighbourhood and effectively segment the whole network into the cluster, with in which they can share and manage cached contents.

Profile Based Resolution: It maintains a historical profile of previously received data request and determines a closer data sources for user's request based on the profile.

Roadside Resolution: If a data request cannot get resolved using these two schemes the data.

Request is forwarded to the original data source. The Roadside Resolution is used to resolve the data request along the forwarding path.

Cache management –In COOP cache management scheme we have to decide which data item to keep in a node's local cache.The aim is to increase the cache hit ratio, which largely depends on the capacity of the cache. To enhance the capacity of cooperative caches, COOP tries to reduce duplicated caching within short distance neighborhood, such that the cache space can be used to accommodate more distinct data items. Authors categories cached data copies based on whether they are already available in the neighborhood or not. A data copy is primary if it is not available within the neighborhood. Otherwise the data copy is secondary. Cache misses the main factor for differentiating between primary and secondary data.Cache miss cost is directly proportional to the travel distance of a data request. The inter-category and intra-category rules are used to decide caching priorities of primary and secondary data.

Limitations of Coop

To improve data availability and access performance, COOP addresses two basic problems of cooperative caching. For cache resolution, COOP uses the cocktail approach which consists of two basic schemes: hop-by-hop resolution and zone-based resolution. By using this approach, COOP discovers data sources which have less communication cost. For cache management, COOP uses the inter- and intra-category rules to minimize caching duplications between the nodes within a same cooperation zone and this improves the overall capacity of cooperated caches. The disadvantage of the scheme is that flooding incurs high discovery overhead and it does not consider factors such as size and consistency during replacement.

Semantic Caching: Semantic caching [6] is used to manage the location dependent data in mobile computing environment. Location dependent data (LDD) is the data whose value is determined by the location to which it is related. Examples include local yellow pages, traffic reports, weather information, and maps and so on. A location dependent query is a query that is processed on location dependent data, and whose result depends on the location criteria explicitly or implicitly specified. The idea of semantic caching is that the mobile client maintains both the semantic descriptions and associated answers of previous queries in the cache. If a new query is totally answerable from the cache, no communication with the server is necessary; if it can only be partially answered, the original query is trimmed and the trimmed part is sent to the server to be processed. Semantic caching is by nature an ideal cache scheme for location dependent applications due to the following reasons:

Semantic caching is built on the semantic locality among queries, which just fits the LDD applications where much semantic rather than temporal or spatial locality is exhibited.

Continuous LDD queries can be incrementally processed by semantic caching. With each successive request, a much smaller trimmed LDD query is processed at the server side and only the differences are transmitted over thewireless link.

Semantic caching makes cache management more flexible. The cache can be managed based on temporal or location information.

Semantic caching also facilitates disconnections even though the data at current location cannot be obtained, the mobile user might still be able to learn the information for other neighbour locations from the local cache.

Semantic caching has been widely used in centralized systems, client-server environment, OLAF systems mobile

computing and heterogeneous systems. The cache is composed of a set of items attached with the related Semantic descriptions, which are called semantic regions in, semantic segments in and so on. While logically the cache is always organized using an index which maintains the semantic as well as physical storage information for every cached item, there are various ways to physically store the data. The commonly used cache replacement strategies are built on temporal locality, such as LRU, MRU and CLOCK. Semantic cache organizes data by semantic metrics; it makes cache management more flexible.

Other Caching Techniques

Yin and Cao [13] propose three schemes: CachePath, CacheData and HybridCache. In CacheData, intermediate nodes cache the data to serve future requests instead of fetching data from the data center. In CachePath, mobile nodes cache the data path and use it to redirect future requests to the nearby node which has the data instead of the faraway data center. To further improve the performance [10, 11] a hybrid approach (HybridCache), is given by taking advantage of CacheData and CachePath while avoiding their weaknesses.

A cooperative caching scheme, called CoCa, was proposed by C.Y.Chow [17]. The CoCa framework facilitate mobile nodes to share their cached contents with each other in order to reduce the number of server requests and the number of access misses in a single hop wireless mobile network. The authors extended CoCa with a group-based cooperative caching scheme, called GroCoCa [18]. According to GroCoCa, the decision of whether a data item should be cached depends on two factors:

- Access affinity on the data items and
- The mobility of each node.

Papadopouli suggested the 7DS architecture [16] in which a couple of protocols are defined to share and disseminate information among users. It operates either on a prefetches mode, based on the information and user's future needs or on an on-demand mode, which searches for data items in a single-hop multicast basis. Hassan Artail proposed a caching scheme COACS [11] which stands for Cooperative and Adaptive Caching System. The idea is to create a cooperative caching system that minimizes delay and maximizes the likelihood of finding data that is cached in the Ad hoc network, all without inducing excessively large traffic at the nodes. COACS is a distributed caching scheme that relies on the indexing of cached queries to make the task of locating the desired database data more efficient and reliable.

Comparison of Cooperative Caching Schemes

Table 1

Schemes	Cache Resolution	Cache Replacement
Summary	Directory-Based	LRU
Squirrel	Hash-Based	LRU
COOP	Cock Tail Scheme	Inter-category and intra category Rules
Yin	CacheData, CachePath, HybridCache	LRU
N.Chand	ZC	LUV
Semantic	No Specification	Temporal locality such as LRU, MRU
Cooperative Proxy Caching	Directory-Based	No Specification
S.Lin	Aggregate	Two factors: 1.Distance 2.Access frequency

Caching In IMANETs

To combine MANETs with the wired network or Internet, we consider the Imanet infrastructure and come across the problem of information search and access under this environment. In IMANETs the routing protocols such as Destination Sequenced Distance Vector (DSDV), Dynamic Source Routing (DSR), Ad hoc On Demand Distance Vector (AODV), Zone Routing Protocol (ZRP), and Temporally Ordered Routing Algorithm (TORA) are explored. They are similar as in MANETs. These protocols based on the model that a sender MT knows the location of receiver MT based on the route information, which is accumulated and analyzed by a route discovery or route maintenance algorithms. A Route discovery operation captures the current network topology and related information. It has to be executed when an MT needs to transmit a data item. To avoid repetitive route discovery, the MTs can cache the old route information.

Why Caching in IMANETs?

IMANETs [3, 8] has several constraints. First, all the MTs cannot access the Internet. Second, due to mobility, a set of MTs can be separated from the rest of the MTs and get disconnected from the Internet. Finally, an MT requiring multi-hop relay to access the Internet may incur longer access latency than those which have direct access to the Internet.

To address these constraints, an aggregate caching mechanism for IMANETs is proposed. The basic idea is that by storing data items in the local cache of the MTs, members of the IMANETs can efficiently access the required information. Thus, the aggregated local cache of the MTs can be considered as a unified large cache for the Imanet. The proposed aggregate cache can alleviate the constraints of IMANETs discussed above. When an MT is blocked from direct access to the Internet, it may access the requested data items from the local cache of nearby MTs or via relays. If an MT is isolated from the Internet, it can search other reachable MTs for the requested data item. Finally, if an MT is located further from the Internet, it may request the data items from

other close by MTs to reduce access latency. Here, two issues are addressed for implementation of an aggregate caching mechanism in IMANETs:

Efficient search: An efficient information search algorithm is fundamental for locating the requested data in IMANETs.

Cache management: To reduce the average access latency as well as enhance the data accessibility, efficient cache admission control and replacement policies are critical. The cache admission control policy determines whether a data item should be cached, while the cache replacement policy intelligently selects a victim data item to be replaced when a cache becomes full.

Information search in IMANETs is different from the search engine based approach used in the wired Internet. An MT needs to broadcast its request to the possible data sources (including the Internet and other MTs within the Imanet) in order to retrieve the requested data efficiently. An aggregate cache for IMANETs is proposed to address the issues of accessibility and latency.

Conclusion

In this paper we have discussed cache sharing issues, schemes related to mobile Ad hoc network environment and give analysis of some popular cooperative caching schemes. These caching schemes are useful in MANETs environment. Here we present how these schemes are advantageous in order to find a data item in a MANETs by using less resources (e.g. network bandwidth, energy etc.) and improves the performance(data availability and latency time). We also discussed the limitations of these techniques. As the cooperative caching is a useful technique to improve the data availability in the MANETs so these analyses will be helpful for the future research.

References

[1] N.Chand, R.C.Joshi and Manoj Misra, “Cooperative caching in mobile Ad hoc networks based on data utility”,Mobile Information Systems pp. 19–37,2007.

[2] Yu Du and S.Gupta, “COOP-A Cooperative Caching service in MANETs”, Proceedings of the IEEE ICAS/ICNS (2005), pp.58-63, 2006.

[3] S. Lim, W. Lee, G. Cao and C. Das, “A Novel Caching Scheme for Internet based Mobile Ad hoc Network Performance,” Ad hoc Networks, vol.4, no.2 pp.225-239,2006.

[4] S.G.Dykes and K.A.Robbines, “A Viability of Cooperative Proxy Caching,” IEEE INFOCOM,pp 1205-1214,2001.

[5] L.Fan, P. Cao, J. Almeida, and A. Z. Broder,“Summary Cache: A Scalable Wide Area Web Cache Sharing Protocol,” IEEE/ACM Transaction on Networking,vol 8,no3,June 2000.

[6] Qun Renet,“Using Semantic Caching to Manage Location Dependent Data in Mobile Computing,” IEEE MOBOCOM, pp.210-220,2000.

[7] T.Hara “Data Replication for Improving Data Accessibility in Ad hoc Networks,” IEEE Transaction on Mobile Computing,vol 5 no11,pp 1515-1532,Nov 2006.

[8] D. Barbara and T. Imielinksi “Sleepers and Workaholics: Caching Strategies for Mobile Environments. In Proc. ACM SIGMOD, pp1–12, 1994.

[9] S. Lim, W. Lee, G. Cao, and C. R. Das. “A Novel Caching Scheme for Internet based Mobile Ad hoc Networks”.In Proc. 12th International Conference on Computer Communications and Networks (ICCCN), pp 38–43, 2003.

[10] F.Sailhan and V. Issarny. “Cooperative Caching in Ad hoc Networks”,In Proc. 4th International Conference on Mobile Data Management (MDM), pp 13–28, 2003.

[11] H.Artail, H.Safa, "COACS: A Cooperative and Adaptive Caching System for MANETs," IEEE Transactions On Mobile Computing, Vol. 7, no. 8, pp 961-977 Aug 2008.

[12] N.Chand, R.C.Joshi and M.Misra.“Cooperative caching strategy in mobile Ad hoc networks based on clusters,” Wireless Personal Communications, 43(1):41–63, 2007.

[13] L. Yin and G. Cao “Supporting cooperative caching in adhoc networks”,IEEE Transactions on Mobile Computing, 5(1):77–89, 2006.

[14] B.Tang, "Benefit-Based Data Caching in Ad hoc Networks" IEEE Transaction on Mobile Computing, vol 7, no3, pp289-304, March2008.

[15] SitaramIyer, Antony Rowstron,and Peter Druschel. “Squirrel: a decentralized peer-to-peer web cache”,In PODC’02:Proceedings of the twenty-first annual symposium on Principles of distributed computing, pages 213–222. ACM Press, 2002.

[16] M.Papadopouli and H. Schulzrinne “Effects of Power Conservation,WirelessConverage and Cooperation on Data Dissemination among Mobile Devices”,In Proceedings of MobiHoc, pages 117–127, 2001.

[17] C.Y. Chow, H.V. Leong and A Chan, “Peer-to-Peer Cooperative Caching in Mobile Environments” Proceedings of 24th International Conference on Distributed Computing Systems Workshop pp528-533,2004

[18] Chi-Yin Chow, Hong Va Leong "GroCoca: Group-based Peer-to-Peer Cooperative Caching in Mobile Environment", IEEE Journal On Selected Areas In Communications, Vol. 25, no. 1, January 2007.

Design of a Compact U-shape Planar Antenna with Multiple Branches

Naresh Kumar[1], Davinder Parkash[2], Sandeep Panwar[1] and Rajesh Khanna[3]

[1]Assistant Professor, ECE Deptt., [2]Associate Professor, ECE Deptt.,
Haryana College of Technology & Management, Kaithal, India
E-mail: devnitk1@gmail.com
[3]Professor, ECE Deptt., Thapar University, Patiala (Punjab), India

Abstract
A compact multiband planar monopole antenna composed of U-shape strip which contains multiple branches is proposed in this paper. The proposed antenna is particularly attractive for WLAN/WiMAX devices that integrate multiple systems. Prototype of the proposed antenna has been designed. The overall size of the antenna is 29.54 mm×17.37 mm×1.6 mm including the CPW feeding mechanism. The greatest feature of a proposed design is a very compact size i.e. volumetric size of 0.82 cm^3. The proposed designed antenna covers the frequency bands from 3.41 GHz to 3.7 GHz (lower-band) and from 4.1 GHz to 6.34 GHz (upper-band) such that total bandwidth of the proposed antenna is 2.51 GHz. The parametric study is performed to understand the effect of multiple strips on the proposed antenna. The maximum simulated gain of the proposed antenna is 5.15 dBi at 6.9 GHz frequency.

Keywords: U-shape, Monopole Antenna, CPW feeding, WLAN and WiMAX

Introduction

Recently, there has been a growing demand of microwave, and wireless communication systems in various applications resulting in an interest to improve antenna performances [1]. Modern communication systems and instruments such as Wireless local area networks (WLAN), mobile handsets require lightweight, small size and low cost. Two commonly used protocols [2] for Wireless Local Area Networks (WLANs) based on access points to relay data, are Wi-Fi and WiMAX, which promise higher data rates and increased reliability. A challenge in designing such multiple wireless communication protocol systems is to design compact, low cost, multiband and broadband antennas. Planar microstrip patch antennas are popular candidates for modern wireless communication systems due to their simple feeding, low profile, low manufacturing cost and easy-to-integrate features [4]. Some popular antenna designs suitable for WLAN and WiMAX operation for 2.4 GHz band (2.4–2.484 GHz), 5.2/5.8 GHz bands (5.15–5.35 GHz/5.725–5.825 GHz) and 2.5/3.5/5.5 GHz (2500–2690/3400–3690/5250–5850 MHz) bands has been reported in [1-14].

In this paper, a single layer substrate antenna design with CPW-feed technology has been used to achieve multi-band operation for WLAN/WiMAX applications. The proposed antenna design is basically composed of U-shape strips with multiple branch strips. The parametric study is performed to understand the characteristics of the proposed antenna. The antenna geometry and design methods are described in Section 2. The simulated results and parametric study are presented in Section 3, followed by a conclusion of this work.

Antenna Structure

The geometry of the proposed dual-band antenna for WLAN/WiMAX applications is shown in Figure 1. The patch is fed by a CPW structure and is printed on a 1.6 mm-thick FR4 substrate, with relative permittivity 4.4 and with overall dimensions of 29.54 mm×17.37 mm×1.6 mm. The proposed antenna has a single layer metallic structure on one side of substrate layer whereas the other side is without any metallization. The basis of the antenna structure is a U-shape strip monopole, which has dimensions of length L and width W, and connected at the end of the CPW feed-line. The designed structure was simulated using IE3D simulation software.

The optimized geometric parameters of the proposed antenna are: length of the U-shape strip L=17.67 mm, width of the U-shape strip W_1=2.35 mm, length of the ground plane L_g=10.10 mm, width of the ground plane W_g=7.5 mm, length of the branch strip L_1=4.67 mm, length of branch strip L_2=4.25 mm, width of the branch strip W_2=1.90 mm, length of the branch strip L_3=8.27 mm, width W_3=1.90 mm, length of the branch strip L_4=4 mm, width W_4=1.90 mm, length of one arm of L shape branch strip L_5=6.8 mm, 2nd arm length of L shape branch strip L_6=5.11 mm and width of the L shape strip W_5=1.45 mm. To give feeding to this geometry a feed line of having length L_f =13.37 mm and width W_f =1.92 mm is used. The distance between the ground plane and rectangular patch is denoted by S= 1.77 mm and the distance between the feed line and ground plane is denoted by ‘d’ is equal to 0.45 mm.

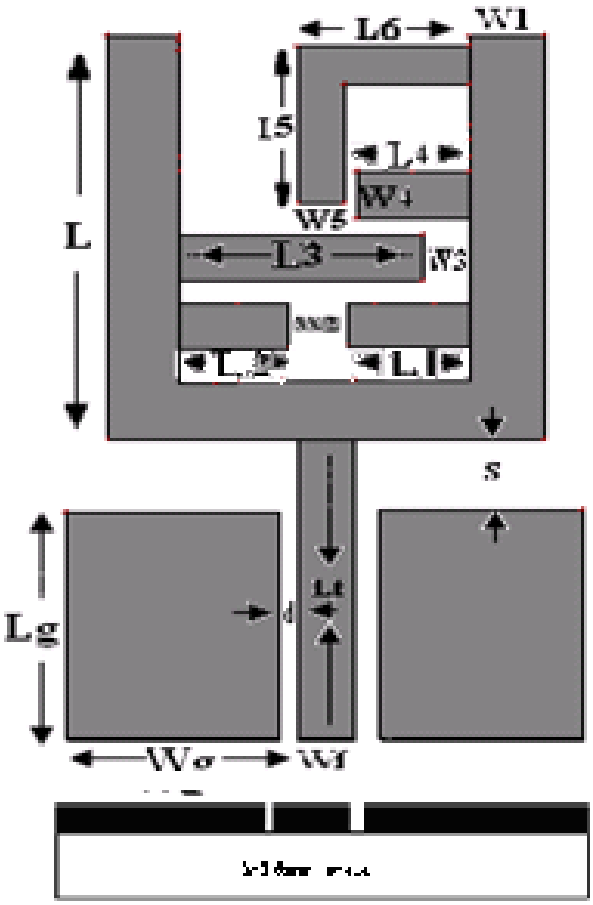


Figure 1: Geometry of the proposed CPW-fed monopole antenna.

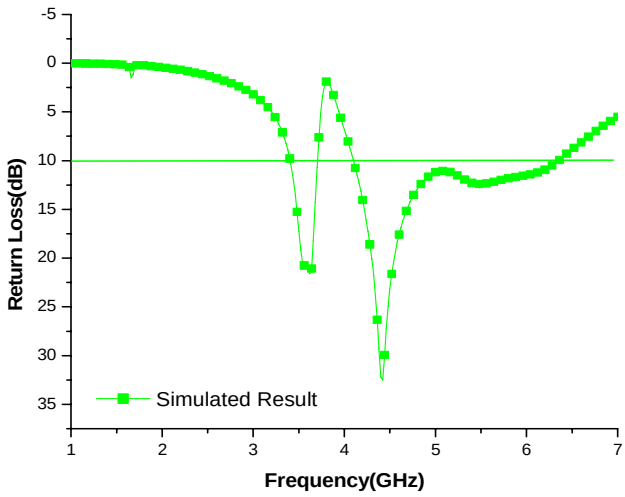


Figure 2: Simulated return-loss of proposed planar antenna.

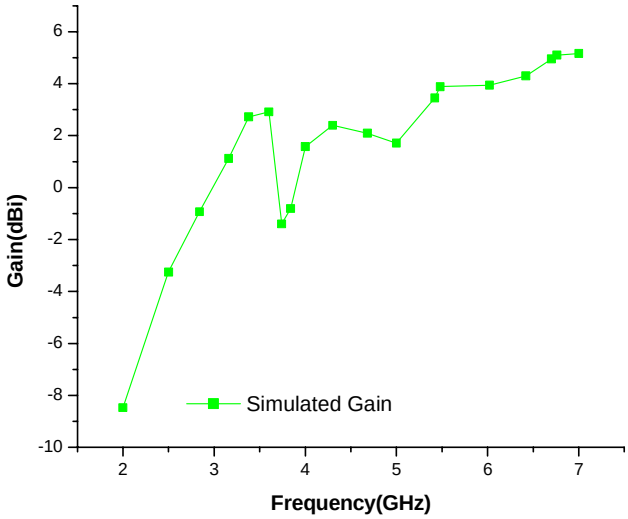


Figure 3: Simulated Gain of the proposed antenna.

Simulation Result and Discussions

The simulated return loss and parametric study results for the proposed monopole antenna are obtained. The simulated return loss and gain are presented for the optimized set of antenna parameters. Simulated return loss of the optimized proposed antenna is shown in Figure 2. From the simulated results, it is clear that dualband operating bandwidths are obtained. The simulated results has a 10 dB impedance bandwidth ranging from 3.41 GHz to 3.7 GHz and from 4.1 GHz to 6.34 GHz such that total bandwidth of the proposed antenna is 2.51 GHz with respect to the central frequency. Obviously, the proposed antenna has very broader bandwidth which covers the required bandwidths of the IEEE 802.11 WLAN standards in the bands at 5.2 GHz (5150–5350 MHz) and 5.8 GHz (5725–5825 MHz) and WiMAX standards in the bands at 3.5 GHz (3.4-3.690 GHz) and 5.5 GHz (5.250-5.850 GHz). The simulated gain of the proposed antenna is shown in Figure 3. The antenna has a maximum gain of about 5.15 dBi at 6.9 GHz frequency with small gain variations in the operating bandwidth. Simulation studies indicate that the maximum antenna radiation efficiency is approximately 90%.

A parametric study is investigated and it demonstrates that the following parameters influence on the performance of the proposed antenna in terms of bandwidth. The parametric study is carried out by simulating the antenna with one geometry parameter slightly changed from the reference design while all the other parameters are fixed. Figure 4 shows the simulated return loss of the proposed antenna as a function of frequency for different shapes. It is observed from the simulation results study that by using only U-shape strip and U-shape with branch strip L_1, dual band is merged into a single band with decrease in the bandwidth as compared to the optimum value of the bandwidth. Figure 5; compare the simulated return loss of the proposed antenna with different configurations of the proposed antenna geometry. It can be seen from the figure that dual band is merged into single band in the case of U-shape with L_3 & L_4 branch strips. It can be seen from the figure that by adding L shape strip, it is converted into dual band with small decrease in the bandwidth as compared to the optimum value of the bandwidth. All the

comparisons show that strips L_1 & L_2 has little effect on the bandwidth of the proposed antenna design.

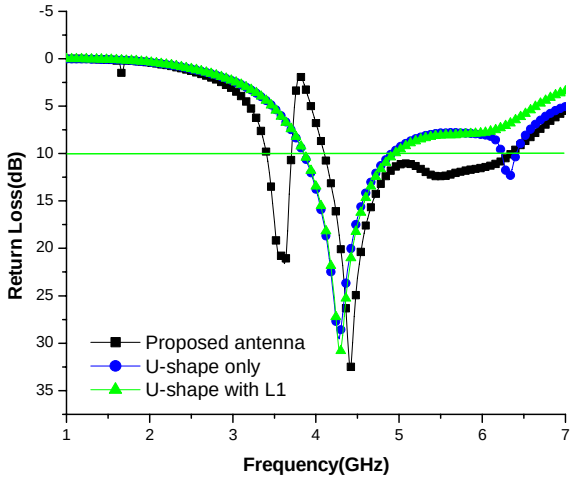


Figure 4 Effect of L_1 & U shape on the proposed antenna

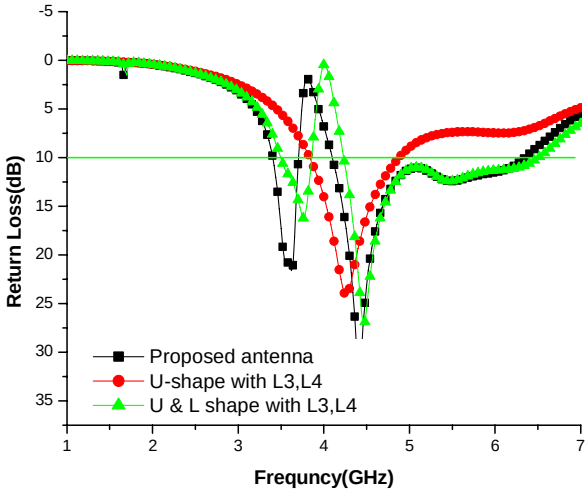
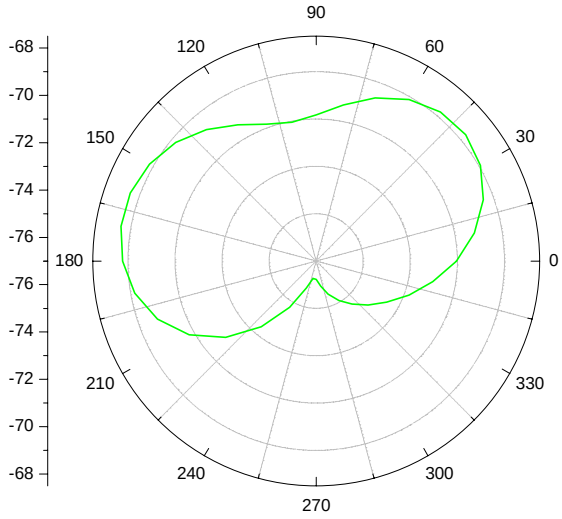


Figure 5 Effect of L_3, L_4 U-shape on the proposed antenna

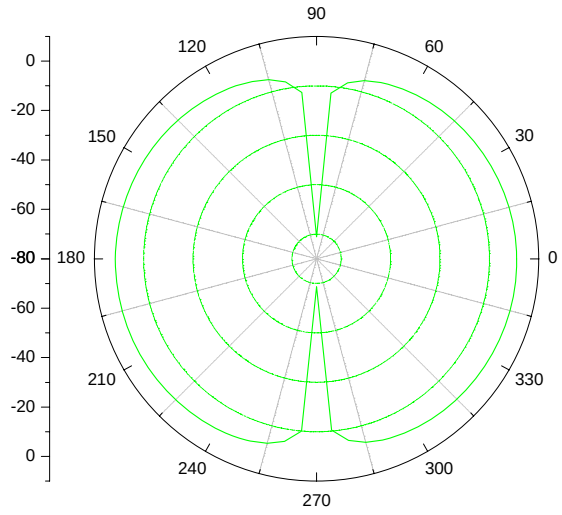
Figure 6 shows the simulated radiation patterns of the proposed dualband CPW-fed monopole antenna at frequency 5 GHz. The simulated radiation patterns cut in the Azimuthal (x-y) plane and cut in the elevation plane (y-z) for the proposed antenna is presented in the figure. Similar to monopole kind of antenna, the radiation pattern obtained in the x-y plane are similar to omni-directional. At the above mentioned frequency, nearly figure of eight radiation pattern is obtained in the y-z plane.

Conclusions

An optimum CPW-fed planar U-shape monopole antenna with multiple branch strips has been proposed and simulated for WLAN/WiMAX operation. The simulated result shows a good impedance bandwidth ranging from 3.41GHz to 3.7 GHz and from 4.1GHz to 6.34 GHz, which shows that the designed antenna is suitable for both WiMAX and WLAN applications with respect to the central frequency. The parametric study shows that the branch strip lengths L_1, L_2, L_3, L_4, and L-shape strip has significant effects on the impedance bandwidth of the proposed antenna. Besides its dual-band characteristics, the proposed antenna remains compact with a volumetric size of 0.82 cm^3 and it is suitable candidate for wireless communication system.



(a) Azimuthal Plane



(b) Elevation Plane

Figure 6 Simulated radiation patterns at 5 GHz for proposed antenna

References

[1] Mouloud Challal,, Arab Azrar and Mokrane Dehmas, "Rectangular Patch Antenna Performances Improvement Employing Slotted Rectangular shaped for WLAN Applications," IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011.

[2] Jiang Zhu, Student Member, IEEE, Marco A. Antoniades, Member, IEEE, and George V. Eleftheriades, Fellow, IEEE, "A Compact Tri-Band Monopole Antenna With Single-Cell Meta-material Loading," IEEE Transactions on Antennas and Propagation, Vol. 58, NO. 4, April 2010.

[3] Indra Surjati, Yuli KN and Yuliastuti, "Increasing Bandwidth Dual Frequency Triangular Microstrip Antenna for WiMAX Application," International Journal of Electrical & Computer Sciences IJECS-IJENS Vol: 10 No: 06.

[4] Davinder Parkash and Rajesh Khanna, "Design and Development of CPW-Fed Microstrip Antenna for WLAN/WiMAX Applications," Progress in Electromagnetics Research C, Vol. 17, pp.-17-27, 2010.

[5] Yen-Liang Kuo, Kin-Lu Wong, "Printed double-T monopole antenna for 2.4/5.2 GHz dual-band WLAN operations," IEEE Trans. Antennas Propagation, vol. 51, pp. 2187–2192, September, 2003.

[6] M. K. Mandal, P. Mondal, S. Sanyal, and A. Chakrabarty, "An Improved Design Of Harmonic Suppression For Microstrip Patch Antennas", Microwave and Optical Technology Letters, pp. 103-105 Vol. 49, No. 1, January 2007.

[7] Haiwen Liu, Zhengfan Li, Xiaowei Sun, and Junfa, "Harmonic Suppression With Photonic Bandgap and Defected Ground Structure for a Microstrip Patch Antenna", IEEE Microwave and Wireless Components Letters, VOL. 15, NO. 2, Feb. 2005.

[8] S. - B. Chen, Y. - C. Jiao, F.-S. Zhang, and Q.- Z. Liu, "Modified T shaped planar monopole antenna for multiband operation," IEEE Transactions on Microwave Theory and Techniques, Vol. 54, No. 8, 3267–3270, August 2006.

[9] W. C. Liu, "Broadband dual-frequency meandered CPW-fed monopole antenna," Electron. Letters, Vol. 40, 1319–1320, 2004.

[10] C.–Y. Wu, S.– H. Yeh and T.-H. Lu, "Planar High Gain Antenna for 5.8-GHz WiMAX Operation," Antennas and Propagation Society International Symposium, 2007 IEEE, pp: 3488 – 3491.

[11] Lin Y.-D. and P.-L. Chi, "Tapered Bent Folded Monopole for Dual-Band Wireless Local Area Network Systems," IEEE Antenna Wireless Propagation Letters, vol. 4, pp. 355–357, 2005.

[12] Wen-Chung Liu, "Broadband Dual-Frequency CPW-Fed Antenna with A Cross-Shaped Feeding Line for WLAN Application," Microwave and Optical Technology Letters, Vol. 49, No. 7, pp. 1739-44, July 2007.

[13] Chien-Yuan Pan, Tzyy-Sheng Horng, Wen-Shan Chen, and Chien-Hsiang Huang, "Dual Wideband Printed Monopole Antenna for WLAN/WiMAX Applications," IEEE Antennas and Wireless Propagation Letters, Vol. 6, pp. 149-151, 2007.

[14] Siddik Cumhur Basaran and Yunus E. Erdemli, "A Dual-Band Split-Ring Monopole Antenna for WLAN Applications," Microwave and Optical Technology Letters, Vol. 51, No. 11, pp. 2685-2688, November 2009.

Electronic Order of Battle Records of Unfriendly Radar Systems using Certain Advanced Techniques as Electronic Support Measures

[1]Ch. Raja, [2]D. Anand and [3]E.G. Rajan

[1]Associate Professor, Electronics & Communication Dept.
Associate Professor, ECE Department, MGIT, Hyderabad, India
E-mail: rajachaluvadi@yahoo.com
[2]Scientific officer-D, Tata Institute of Fundamental Research
Balloon facility, ECIL post, Hyderabad-500062, India
E-mail: anandd24@yahoo.co.in
[3]Founder President, Pentagram Research Centre Pvt. Ltd.
#201, Venkat Homes, MIGH-59, Mehdipatnam, Hyderabad, India
E-mail: rajaneg@yahoo.co.in

Abstract
Radar is the key sensor of any modern weapon system. Its capability to function in all weather environments at long ranges is unmatched with any other available sensor. Land based radars are used for variety of tasks ranging upward in size and complexity from man portable radars for detection of vehicle and personnel to ballistic missile tracking phased arrays. Increased reliance on radars, communication systems, speed of missile and weapon system and high speed detection and tracking has increased the importance of Electronic Warfare (EW). Electronic warfare is subdivided into Electronic Support Measure (ESM), Electronic Counter Measure (ECM) and Electronic Counter Counter Measure (ECCM) systems. This paper deals with the ESM which is the division of electronic warfare involving action taken to search for intercept, identify and locate sources of radiated electromagnetic energy from radar for the purpose of threat recognition.

Introduction
Electronic Support Systems fingerprints the electromagnetic sources with the help of sensitive receivers and direction finding (DF) systems. This provides information in the timely fashion that are readily usable by the activities in support. The prime consumers of ESM information are the operations personnel engaged in other EW activities that is ECM and ECCM. ESM provides essential to the success of operations in the air combat environment. ESM includes signal interception and detection, measurement of signal parameters, signal identification and threat signal evaluation. Signal parameters include frequency, emission type, PRF, Pulse-width, antenna scan and characteristics, polarization and angle of arrival. Special mission flights are generally taken to intercept EM radiations for ON line and OFF line signal analysis. These recorded signals are processed for estimating radar signal parameters (offline radar signal processing) to prepare Electronic Order of Battle (EOB). This paper discusses the state-of-the-art Electronic Support Measures undertaken by almost all developed countries.

Electronic Support Measures (ESM)
The prime objectives of ESM are (i) To acquire technical information about various radars deployed in the enemy territory, (ii) To monitor enemy radar stations round the clock and supply information to defense forces for strategic planning. The objects that are incidental to the main objectives are (i) To undertake special mission flights and special operations for intercepting Electro Magnetic (EM) radiations for ON line and OFF line signal analysis, (ii) To process recorded signals for estimating signal parameters(Off Line Radar Signal Processing), (iii) To prepare Electronic Order of Battle (EOB) records of various enemy radar regiments for measuring war time, (iv) T o have joint action with Air Force, Army and Navy in the process of acquiring technical information about radars, and (v) To support Defense Laboratories by extending first hand technical information about special equipment. This can be achieved by either air surveillance or monitoring station situated at suitable geographic heights. Special flights are arranged to operate along the border in order to intercept signals from hostile radars. Tuners and wide band receivers are used for this purpose. Intercepted signals are recorded on magnetic media for OFF line analysis. Detected signals with down translated IF are stored by high speed recorders. Post detected signals are stored as stretched pulses by low speed recorders. Signal bearing is determined for each emitter from the aircraft itself using direction finding systems like rotating dipole or Bellini Tosi Antenna. Monitoring stations situated at suitable geographic height are ideal sources of acquiring radar signals. Round the clock watch is generally carried out. Very useful information about ON time, OFF time, Mean Time Between Failure (MTBF), types of ECMs and ECCMs incorporated in radar under watch could be obtained. Tuners and wide band receivers are used for the intercept.

Radar Intercept System and Recording

A typical airborne intercept system is shown in figure 1.

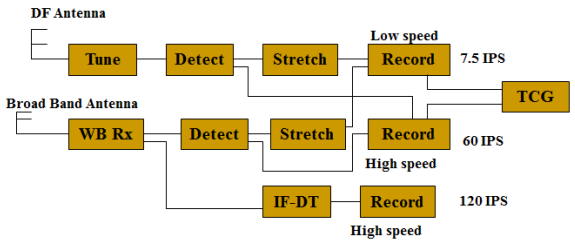


Figure 1: Airborne intercept system.

A typical ground based intercept system is shown in figure 2.

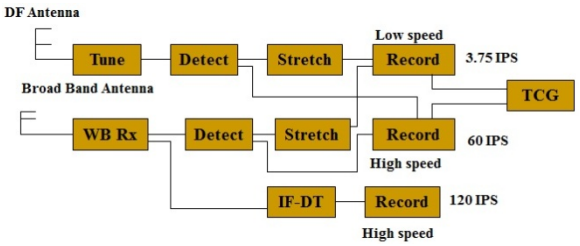


Figure 2: Ground based intercept system.

Recorded signals are sent to BASE station for further analysis. A traditional technique which is used to fix the position of an emitter is shown in figure 3.

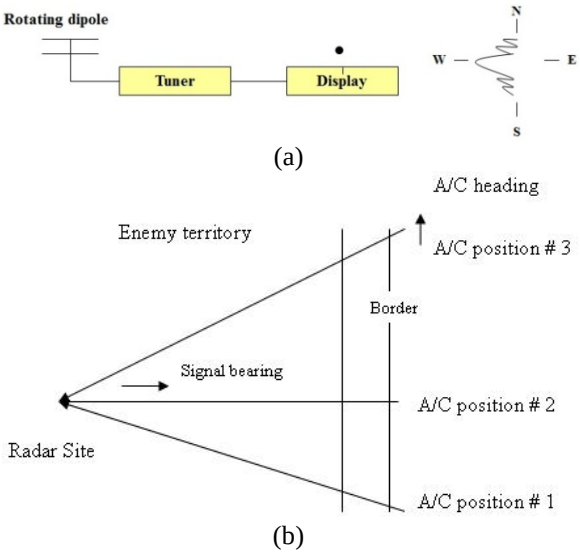


Figure 3: Position fixing of an emitter using direction finding (a) Block diagram of receiver and display (b) Typical mission flight

Signal bearings are obtained at various positions of a surveillance aircraft and the emitter position is estimated at the meeting point of the signal bearings. In practice, three signal bearings are taken and the plot drawn to estimate the position of the emitter. Three bearings would generally yield a triangle as shown in figure 4 and in such a case the centroid of the triangle is considered to be the estimate position of the emitter.

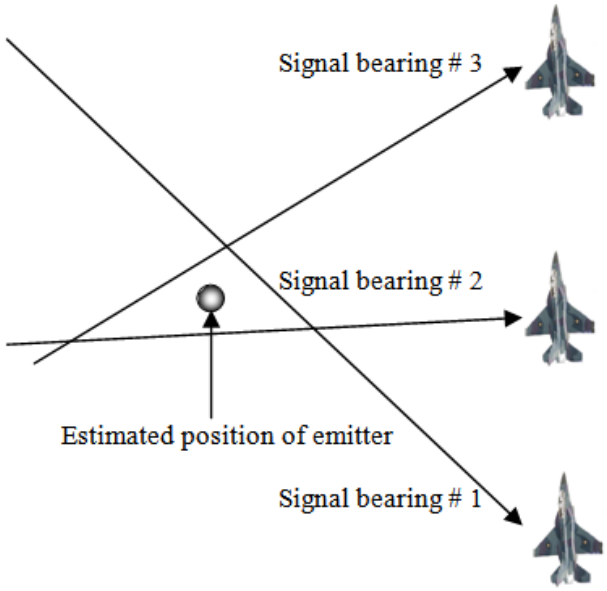


Figure 4: Position fixing of an emitter

Signal Intelligence (SIGINT)

Standard Radar Frequency Letter-Band Nomenclature

Table 1 shows the standard radar frequency bands and their nomenclature. RAVEN 1 and RAVEN 2 operators tune the frequencies band by band as per supervisor's instructions and record the pulse stretched signals.

Note that the wide band receivers of all the bands are kept on throughout the mission flight and the pulse stretched signals recorded continuously.

Table 1: Band of radar frequencies

Band Designation	Frequency Range	Specific Frequency allocation Bands
HF	3 – 30 MHz	
VHF	30 – 300 MHz	138 – 144 MHz 216 – 225 MHz
UHF	300 – 1000 MHz	420 – 450 MHz 890 – 942 MHz
L	1000 – 2000MHz	1215 – 1400 MHz
S	2000 – 4000 MHz	2300 – 2500 MHz 2700 – 3700 MHz
C	4000 – 8000 MHz	5250 – 5925 MHz
X	8000 – 12000 MHz	8500 – 10680 MHz
Ku	12 – 18 GHz	13.4 – 14 GHz 15.7 – 17.7 GHz
K	18 – 27 GHz	24.05 – 24.25 GHz
Ka	27 – 40 GHz	33.4 – 36.00 GHz
mm	40 – 300 GHz	

Recording of Intercepted Signals on a Magnetic Tape

Figure 5 shows an ideal pulse train output of the receiver. But the actual pulse train recorded on the tape would not be ideal. In fact, the pulse is stretched and low pass filtered. Because of the pulse stretch, the bandwidth would be considerably reduced.

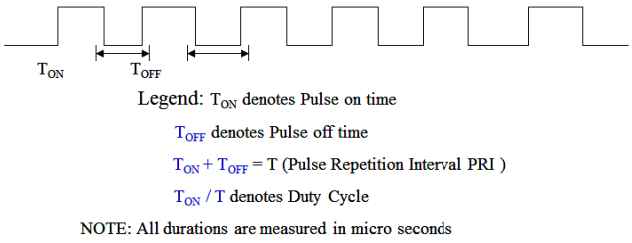


Figure 5: Ideal shape of a pulse train.

A 1400 feet long tape of width of one inch having 14 tracks is used for recording purposes. Recording is done using a compact airborne recording system. Band 1, 2, 3, 4 and 6 receiver outputs are recorded in the tracks 1, 2, 3, 4 and 5 respectively. The tuner output which covers all the five bands is recorded in the track 6. RAVEN #1 operator comments are recorded in track 7. Band 7, 8 and 9 receiver outputs are recorded in the tracks 8, 9 and 10 respectively. The time code signal with 8421 BCD modulated on 1 KHz carrier is recorded in track 11. A 100 KHz reference signal is recorded in the 12th track. RAVEN #2 tuner output which covers all the three bands is recorded in the track 13. RAVEN #2 operator comments are recorded in track 14. A typical ground based 14 channel recording system is shown in figure 6.

http://en.wikipedia.org/wiki/File:Ampex_FR-900_at_LOIRP.jpg (Courtesy)

Figure 6: One inch 14-channel Instrumentation recorder reproducer system used for signal analysis

The recording format of various bands of intercepted and pulse stretched radar signals on a one inch 14 track tape is shown in figure 7.

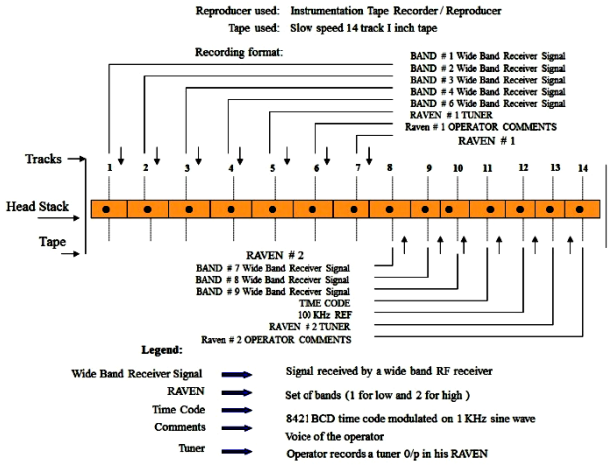


Figure 7: Recording format on a multi track tape

Signal Intelligence From Low Speed Tape (Tape Speed: 3-3/4 and 7-1/2 Inches Per Second)

Pulse Repetition Frequency (PRF)

The recorded tape is run at the appropriate speed and the tuner signal record is fed to the first vertical input of a dual beam oscilloscope. The corresponding wide band signal output is fed to the second vertical input of the scope. Internal signal is used to stabilize both the waveforms. When tuner signal is stabilized the corresponding wide band signal will also be locked with it. The corresponding ramp output of the oscilloscope is fed to the input of a frequency counter. The frequency counter shows the PRF of the tuned signal.

Scan Type and Scan Time

Run the tape at the appropriate speed. Lock the tuner signal with the corresponding wide band signal. Use a stop watch. Put on the watch when the wide band signal strength is maximum and put off the watch when the next immediate maximally strong signal is heard. Note the time duration. Repeat this many times. Take the average of these trials. This is the scan period of the radar. Similar procedure can be used to find the vertical scan of the height finding radar. The beam is intercepted four times, twice for each nod. The period between two successive nods is the scan period of height finder. Refer to figure 8 which is self explanatory.

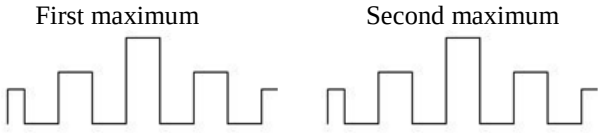


Figure 8: Successive maxima in one nod

Time Code

Sinusoidal oscillations at a predetermined constant frequency are rectangularly amplitude modulated according to a binary

code (IRIG Standard Time Code). The modulated signal is recorded on the magnetic tape. This code provides the actual time information like time of intercept. Figure 9 shows a typical time code generator / reader used for this purpose. Figure 10 shows the IRIG serial time code format.

Figure 9: Front end panel of a TCG

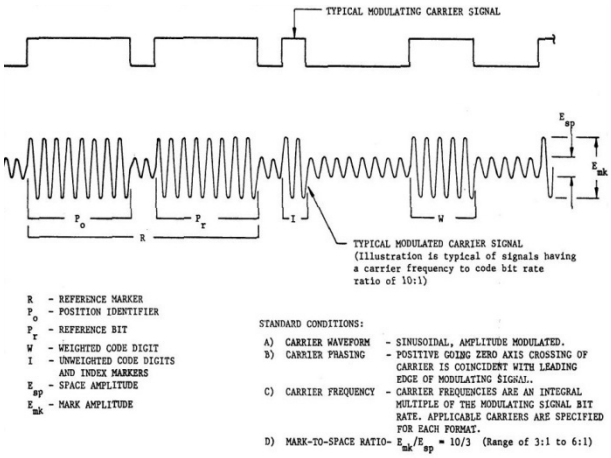


Figure 10: IRIG serial time code format

Reference signal

A reference signal of 100 KHz is recorded along with other signals in track #12. This is done with the idea of running the tape with the same speed at which the recording was done, during off-line signal analysis. The recorded reference signal of 100 KHz is fed to the digital phase lock loop circuit of the recorder / reproducer used for off-line signal analysis and the speed of the take up reel motor is controlled in order to match with the speed with which the tape was run during recording. Figure 11 shows a part of the DPLL circuit of the recorder / reproducer system.

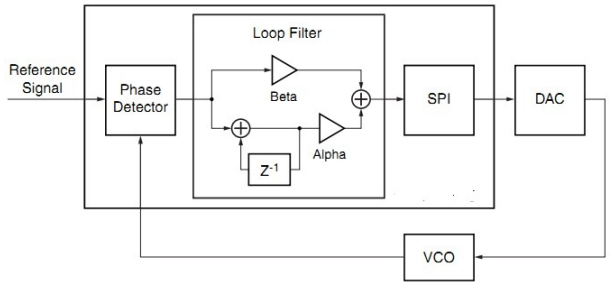


Figure 11: DPLL circuit of the recorder / reproducer system

Signal Intelligence From High Speed Tape
(Tape Speed: 60 and 120 Inches Per Second)

Pulse Width

Tektronix 565 dual beam dual trace 10 MHz oscilloscope is traditionally used to measure pulse width of a radar signal recorded in the high speed tape. Generally the tape is run at 60 or 120 inches per second speed during recording and reproducing. Figure 12 shows the front panel view of the Tektronix 565 oscilloscope.

Figure 12: Tektronix 565 oscilloscope

The high speed recorder output is fed to one of the channels of the oscilloscope. The time base vernier control knob is kept at 'calibrated' position. Internal trigger is used to display a single pulse. The measurement technique is shown in figure 13.

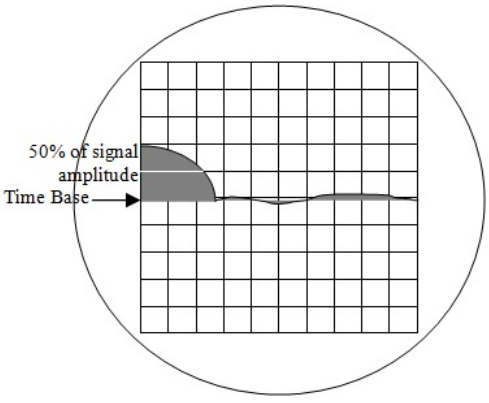


Figure 13: Pulse width measurement

Pulse Jitter or Stagger Ratio

Staggering of PRIs is used to avoid blind speed effect in moving target indicator (MTI) radars. A suitable delay line is used to delay the PRI and delayed signal is added to the original signal to produce staggered variations. Some times two different PRIs are added to form staggered waveforms. Stagger is the delay time between two leading edges and it is visualized and measured in an oscilloscope as shown in figure 14.

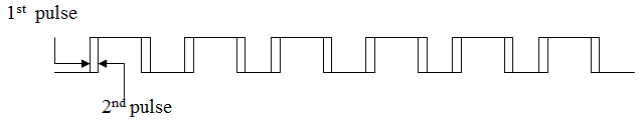


Figure 14: Staggering of pulses

Intra-Pulse Modulation

Parameters such as frequency profile within pulse, amplitude profile within pulse, phase profile within pulse, rise time, fall time, pulse width, amplitude variations over time and frequency and signal modulations are determined.

In modern radar systems FM chirp, Barker codes, pseudorandom codes and poly-phase codes are used as an ECCM measure. A pulse to pulse frequency agile radar operates with a set of discrete RF switched in a pseudorandom fashion. To extract intelligence from intra-pulse modulation, lock wide band receiver signal with tuner signal. Trigger wide band receiver signal at its leading edge in the oscilloscope. Feed the signal to the RF section of the spectrum analyzer and tune it to obtain the spectrum of the signal. From the spectrum, one can verify intra-pulse FM, chirp and Barker codes.

For example, AN/TPS-43 is a light weight transportable radar designed for use in a wide variety of tactical environments. The radar works in 16 discrete frequencies ranging from 2800 MHz to 3000 MHz. It radiates six beams stacked one over the other to form a cosec^2 radiation pattern. Each beam is switched with one of the 16 discrete frequencies following a pseudo random bit sequence. The system provides 3-D cover to 447 km on a fighter or fighter-bomber aircraft and measures heights over the full range by signal amplitude comparisons in six channels. Clutter rejection and electronic counter-countermeasures features are incorporated in the design. Figure 15 shows the spectrum of all 16 discrete frequencies.

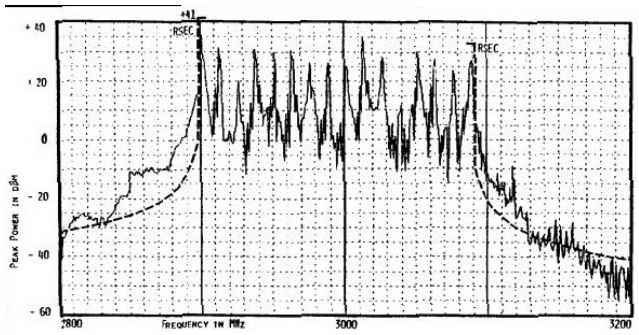


Figure 15: Spectrum of 16 frequencies of TPS 43

One may observe that the received signal by the tuner is not continuous where as the one received by the wide band receiver would present a scenario which is shown in figure 15. Hence, one should take care in measuring pulse width information from the recorded signal. The radio frequency of this radar is phase code modulated inside the pulse. Appropriate techniques are used to decode.

Signal Intelligence From Operator's Comments (RAVEN #1 and RAVEN #2)

RAVEN #1 operator's comments are recorded in track #7 and RAVEN #2 operator's comments are recorded in track # 14 of the tape. The operators provide information about the RF of the radar with the help of tuner and Angle Of Arrival (AOA) of the signal. They also provide very valuable information about the signal strength, locking of the height finding system with early warning system, locking of the fire control radar with a target and so on.

Signal Intelligence From Collateral Sources

Apart from the techniques outlined in earlier sub sections, one can collect intelligence from literature and contacts. The literatures available are Janes Weapon Systems, ECM and ECCM Journals, text books, research articles, conference proceedings etc. Foreign intelligence agencies, national intelligence agencies, professionals and informers are also the sources of information through contacts. From the analysis report, it is usual practice to match a set of parameters with existing literature and estimate the presence and type of radar at a particular site.

With these details, most of the intelligence is collected about various radar systems in the enemy territory. Now, one goes in for creating the Order Of Battle (EOB) record as explained below.

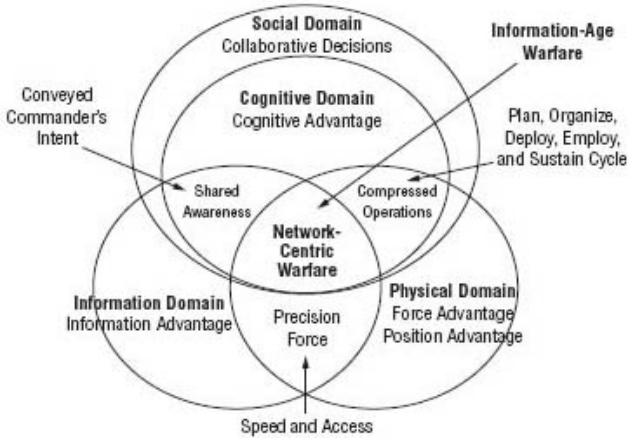
Electronic Order of Battle

Generating an **Electronic Order of Battle** (EOB) requires identifying SIGINT emitters in an area of interest, determining their geographic location or range of mobility, characterizing their signals, and, where ever possible, determining their role in the broader organizational order of battle. EOB covers both Communication Intelligence (COMINT) and Electronic Intelligence (ELINT), The National Defense Intelligence Agency maintains an EOB by location. Army, Nay and Air Force Intelligence Agencies involve in SIGINT and COMINT activities and the Joint Intelligence Committee (JIC) of the government consolidates the EOB reports of all such agencies.

Modern ESM

Network Centric Warfare (NCW)

Unlike traditional ESM and warfare activities, the Network Centric Warfare (NCW) involves various domains in spite of the fact that domains do clash with one another.



http://www.airpower.au.af.mil/airchronicles/apj/apj08/spr08/pattee.html

Figure 16: Domain conflicts in NCW.

Figure 16 depicts one such possible domain conflicts scenario which clearly indicates that the nations which have spectrum superiority and technical superiority blended with national integrity would always win the future battles,

Conclusions

In this paper, we have just outlined a broad based and sensitive concept of Electronic Support Measures and their significance in the light of security.

The material covered in this paper is due to certain basic research carried out by the first two authors and a vast practical experience of the third author.

Acknowledgements

The authors express their sincere thanks to the administration of Pentagram Research Centre Pvt Limited, Hyderabad, India for the technical and logistic support given to them in carrying out advanced research in allied fields.

References

[1] August Golden Jr., Radar Electronic Warfare AIAA Education series 1987.

[2] Fred E. Nathanson, Radar Design Principle, Second Edition.

[3] Dr. V K Atre, Electronic Warfare- A perspective, IETE Technical review Vol 17, No6, Nov-Dec-2000.

[4] G Nagendra Rao, CVS Sastry, N Diwakar, Trends in Electronic Warfare, IETE Technical review Vol20, N02, and March-April 2003.

[5] T.D. Bhatt, E.G. Rajan, P.V.D. Somasekhar Rao, "Design of frequency-coded waveforms for target detection", IET Radar, Sonar and Navigation, March 2008, Vol. 2, No. 5, pp. 388–394

[6] T.D. Bhatt, E.G. Rajan, P.V.D. Somasekhar Rao, "Design of High-Resolution Radar Waveforms for Multi-radar and Dense Target Environments", IET Radar, Sonar and Navigation, 2011.

[7] Rajan E. G., Symbolic Computing-Signal and Image Processing, Anshan Publications, United Kingdom, 2003

Quality of Service (QoS) Based Sheduling Environment

Arun Kumar[1] and Dr. A.K. Garg[2]

[1]Electronics & Communication Department, M.M. University, India
E-mail: ranaarun1@gmail.com
[2]Electronics& Communication Department, M.M. University, India
E-mail: garg_amit03@yahoo.co.in

Abstract
In the field of computer networking and other packet-switched telecommunication networks, the traffic engineering term quality of service (QoS) refers to resource reservation control mechanisms rather than the achieved service quality. Quality of service is the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow. For example, a required bit rate, delay, jitter, packet dropping probability and/or bit error rate may be guaranteed. Quality of service guarantees are important if the network capacity is insufficient, especially for real-time streaming multimedia applications such as voice over IP, online games and IP-TV, since these often require fixed bit rate and are delay sensitive, and in networks where the capacity is a limited resource, for example in cellular data communication. The IEEE 802.16 is a standard for broadband wireless communication in Metropolitan Area Networks (MAN). To meet the QoS requirements of multimedia applications, the IEEE 802.16 standard provides four different scheduling services: Unsolicited Grant Service (UGS), real-time Polling Service (rtPS), non-real-time Polling Service (nrtPS), and Best Effort (BE). The paper aim is to verifying, via simulation, the effectiveness of rtPS, nrtPS, and BE in managing traffic generated by data and multimedia sources and Simulation results show that our scheme is capable to provide QoS.

Keywords: QoS, WiMAX, IEEE 802.16 etc.

Introduction

IEEE 802.16 [1] is a very promising system enabling broadband wireless access (BWA). IEEE 802.16 standard also known as worldwide interoperability for microwave access (WiMAX) defines two modes to share wireless medium: point-to-multipoint (PMP) mode and mesh mode. In the PMP mode, a base station (BS) serves several subscriber stations (SSs) registered to the BS. In IEEE 802.16, data are transmitted on the fixed frame based. The frame is partitioned into the downlink subframe and the uplink subframe. Frame duration and the ratio between the downlink subframe and the uplink subframe are determined by the BS. In the PMP mode, the BS allocates bandwidth for uplink and downlink. The BS selects connections to be served on each frame duration[2]. the IEEE 802.16 standard [3] defines four types of service flows, each with its own QoS needs. Each connection between the SS and the BS is coupled with one service flow. The Unsolicited Grant Service (UGS) transmit constant bit rate (CBR) flows of CBR like applications such as Voice over IP. The real-time Polling Service (rtPS) is considered for applications with real time needs which produce variable size data packets regularly, such as MPEG video streams. In this class, QoS guarantees are given in the form of restricted delay with minimum bandwidth guarantees. The nonreal-time Polling Service (nrtPS) is adequate for better than- best-effort services such as FTP services. Similar to rtPS, minimum bandwidth guarantees are also given to nrtPS connections. The Best Effort service (BE) is used for best-effort traffic such as HTTP[4]. For years, the IEEE has devoted continuous efforts to develop the wireless metropolitan area network (MAN) 802.16 standard, streamlined as the Worldwide Interoperability for Microwave Access (WiMAX) by the WiMAX Forum. This standard has since attracted a great deal of attention in both the research and industry communities, and is touted as the next killer technology that promises to offer *multiplay* services in the future wireless multimedia marketplace. The main advantages of WiMAX lie in its cost-competitive deployment and comprehensive quality of service (QoS) support for large numbers of heterogeneous mobile devices with high-datarate wireless access. Since 2004, WiMAX has established its relevance as a wireless extension (or alternative) to conventional wired access technologies, such as T1/E1 lines, cable modems, and digital subscriber line (xDSL), extending the reach to remote areas. Mobile WiMAX, based on the IEEE 802.16-2004 and IEEE 802.16e amendment [5], fills the gap between the wireless local area network (WLAN) and third-generation (3G) cellular systems with respect to their data rate and coverage trade-offs, and acts as a strong competitor to the current 3G Partnership Project (3GPP) long-term evolution (LTE) on the road to 4G wireless broadband markets[6]. there are huge and different kinds of videos streaming from different users which may influence each other and thus, it is essential to enforce a scheduling policy designed for suitable video metrics and efficient network utilization, preferably in a distributed manner[7].

Background of Scheduling Environment

There are many papers that propose new packet scheduler environment for 802.16 network, in order to provide different levels of QoS guarantees for various applications. This is driven by the lack of standardisation for the Admission

Control and Uplink Scheduling algorithm for rtPS, nrtPS and BE service flows in the 802.16 standard. [8] Proposes an architecture that introduces a framework for the scheduling algorithm and admission control policy for 802.16. They also suggest system parameters that may be used, and define traffic characteristics for which the network can provide QoS .[9] provides a detailed description of the proposed architecture and more background on the 802.16 standard. Authors in [10] Presents a scheduler where the priority is based on the channel and service quality. Huei-Wen Ferng and Han-Yu Liau[11] has proposed how to simultaneously achieve fairness and quality-of-service (QoS) guarantee in QoS-oriented wireless local area networks (LANs) is an important and challenging issue. Targeting at this goal and jointly taking priority setting, fairness, and cross-layer design into account, four scheduling schemes designed for the QoS-oriented wireless LAN mainly based on concepts of deficit count and allowance are proposed in this paper to provide better QoS and fairness. Bader Al-Manthari, et al.[12] has proposed a novel downlink packet scheduling scheme for QoS provisioning in BWASs. The proposed scheme employs practical economic models through the use of novel utility and opportunity cost functions to simultaneously satisfy the diverse QoS requirements of mobile users and maximize the revenues of network operators. Liang Zhou, et al.[7] has proposed important issue of supporting multi-user video streaming over wireless networks is how to optimize the systematic scheduling by intelligently utilizing the available network resources while, at the same time, to meet each video's Quality of Service (QoS) requirement. In this work, they proposed the problem of video streaming over multi-channel multi-radio multihop wireless networks, and developed fully distributed scheduling schemes with the goals of minimizing the video distortion and achieved certain fairness. HONGFEI DU, et al.[6] has proposed the design issues and the state of the art of multimedia downlink scheduling in the multicast/broadcast-based WiMAX system. This proposed a viable end-to-end framework, connection-oriented multistate adaptation, by considering cross-layer adaptations in source coding, queue prioritization, flow queuing, and scheduling. Its performance is confirmed by simulations on important metrics, showing that the framework can effectively accommodate heterogeneity in link variations, queue fluctuations, and reception diversities.

System Overview

IEEE 802.16 defines two types of operating modes: PMP mode and mesh mode. In the PMP mode SSs are geographically scattered around the BS. The performance of IEEE 802.16 in the PMP mode is verified in [8]. Our system model is based on a time-division-duplex (TDD) mode. The frame of IEEE 802.16 is divided into the downlink subframe and the uplink subframe. The downlink subframe starts with preamble followed by frame control header (FCH), downlink map (DL-MAP), uplink map (UL-MAP) messages and downlink burst data. The IEEE 802.16 frame structure is illustrated in Fig. 1.

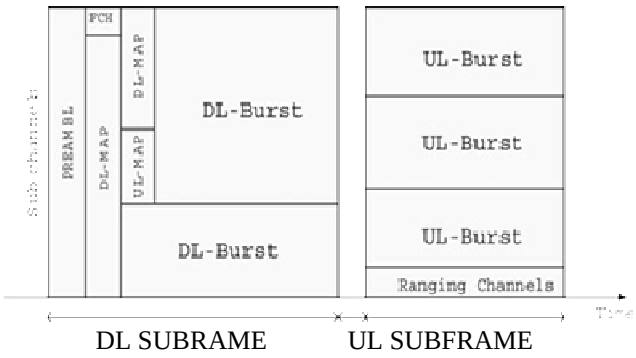


Figure 1: IEEE 802.16 frame structure[8]

The DLMAP message defines the start time, location, size and encoding type of the downlink burst data which will be transmitted to the SSs. Since the BS broadcasts the DLMAP message, every SS located within the service area decodes the DL-MAP message and searches the DL-MAP information elements (IEs) indicating the data burst directed to that SS in the downlink subframe. After the transmit/receive transition gap (TTG), the uplink subframe follows the downlink subframe. IEEE 802.16 provides many advanced features like adaptive modulation coding (AMC), frame fragmentation and frame packing.

System Model and Packet Scheduling Model

We consider a BWAS consisting of a downlink time-slotted channel, as shown in Fig. 2. Transmission is done in time frames of fixed or variable size duration, where each frame consists of a number of time slots[12]. We assume that the base station serves N users. We also assume that there are K classes of traffic, where class i has higher priority than class i+1. Let Ni denote the number of class i users and N= $\sum_{i=1}^{k} Ni$. We allow users within the same class to have different QoS requirements depending on the type of applications they are running. Packet scheduling in next generation BWASs works as follows. Each user regularly informs the base station of his channel quality condition by sending a report in the uplink to the base station. The report contains information about the instantaneous channel quality condition of the user. The base station then would use this information to select the appropriate user(s) according to the adopted scheduling scheme. For example, in HSDPA, users are able to measure their current channel quality conditions by measuring the power of the received signal from the base station

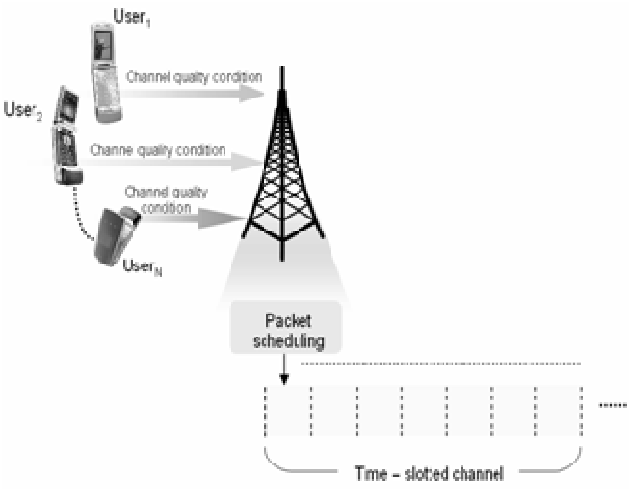


Figure 2: Packet scheduling[12]

Simulation Background

A simple WiMAX network simulated in OPNET modeler version 14.0-PL0[13] as illustrated in Figure 3. The network consists of several cells containing one BS and five SSs in each cell.

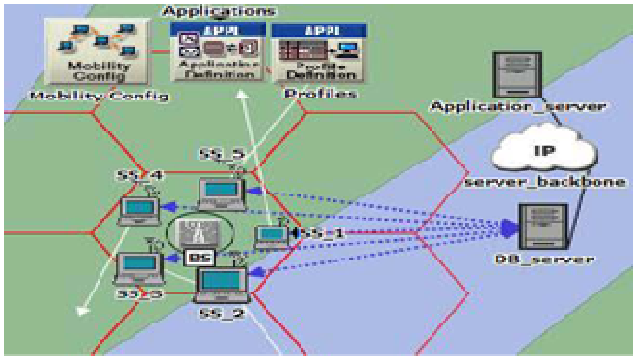


Figure 3: OPNET simulation of a small WiMAX network including one cell, one BS and 5 SSs[13]

The source of traffic is an application server that could provide five types of application, one for each type of traffic. The assumption is that each SS carries the traffic from only one user, and each user is using only one type of application at-a-time. Applications, the type of traffic they represent, and Subscribers that request them are listed in Table 1.

Table 1: Applications and type of traffic used in OPNET

Application	voice	video	remote log	http	e-mail
Traffic type	UGS	ertPS	rtPS	nrtPS	BE
Requester	SS_1	SS_2	SS_3	SS_4	SS_5

The simulation started by downloading files from the application server, one by each SS. First scenario is tested with the network topology in Figure 2 using the basic scheduler provided by OPNET. A simple priority queue discipline is implemented, which prioritizes different types of services based on strict priority rules. This scheme schedules traffic from different classes based on the requirements set by the standard in a priority order with UGS as the highest priority and BE with lowest priority.

Figure 4 shows the results for throughput achieved by each service class in the basic scenario. As the traffic starts, the throughput achieved by UGS application increases rapidly to over 2.5Mbps, and then, it slows down and gradually increases to a level slightly above 3Mbps. The next lines below UGS are ertPS which reaches a level close to 1.5Mbps, and the rtPS, which settles down at around 783Kbps. Then there is a large drop to 83Kbps and 36Kbps respectively for nrtPS and BE services.

These results indicate that although the low priority flows are not completely starved, however, the higher priority RT applications are consuming a considerable portion of the bandwidth. In this scenario, a link with high data rate is employed, which resembles a typical WiMAX over-provisioning case, in which the total bandwidth is not completely utilized.

Figure 4: Throughput achieved by each type of service using basic OPNET scenario[13].

Conclusion & Future Scope

In future, it is plan to further develop this scheme by introducing two new parameters called Fairness and Utilization based on bandwidth distribution among all service classes. QoS differentiation is characterized by the ability of the scheduling scheme in delivering fairness to all service flows simultaneously, and improving utilization of the system. QoS differentiation for WiMAX has been studied in the literature recently, but there is neither clear definition nor a set of parameters to quantify the differentiation. The main motivation for this study is to open a direction in the study of QoS support which yields to further analyze and quantify QoS differentiation using the proposed parameters. In this, via simulation, the effectiveness of scheduling services: Unsolicited Grant Service (UGS), real-time Polling Service (rtPS), non-real-time Polling Service (nrtPS), and Best Effort (BE) in managing traffic generated by data and multimedia

sources and to achieve optimum result in QoS using software like NS2, Qualnet, Opnet etc.

References

[1] IEEE 802.16a-2003: 'IEEE Standard for local and metropolitan access network part 16: air interface for fixed broadband wireless access systems'. Amendment 2: Medium Access Control Modifications and Additional Physical Layer Specifications for 2-11 GHz, 1 April 2003.

[2] T. Kim J.T. Lim, "Quality of service supporting downlink scheduling scheme in worldwide interoperability for microwave access wireless access systems", IET Commun., 2010, Vol. 4, Iss. 1, pp. 32–38.

[3] IEEE 802.16-2004 (802.16REVd) Specification,.802.16-2004 IEEE Standard for Local and Metropolitan Area Networks, Part 16, Air Interface for Fixed Broadband Wireless Access Systems,. June 24, 2004, p.31, p.207.

[4] Elmabruk.Laias, Irfan Awa, Pauline.ML.Chan, "An Integrated Uplink Scheduler In IEEE 802.16" ,2008 IEEE pp 518-523.

[5] IEEE 802.16e-2006, "IEEE Standard for Local and Metropolitan Area Networks — Part 16: Air Interface for Fixed Broadband Wireless Access Systems," Feb. 2006.

[6] Hongfei du, Jiangchuanliu, and Jieliang, "Downlink scheduling for multimedia multicast/broadcast over mobile wimax: connection-oriented multistate adaptation", 2009 IEEE pp 72-79.

[7] Liang Zhou, Xinbing Wang, Wei Tu, Gabriel-MiroMuntean, and Benoit Geller, "Distributed Scheduling Scheme for Video Streaming over Multi-Channel Multi-Radio Multi-Hop Wireless Networks", VOL. 28, NO. 3, APRIL 2010 pp409-419.

[8] CICCONETTI C., ERTA A., LENZINI L., MINGOZZI E.: 'Performance evaluation of the IEEE 802.16 MAC for QoS support', IEEE Trans. Mob. Comput., 2007, 6, (1), pp. 26–38

[9] KittiWongthavarawat, Aura Ganz, Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems, Wiley, c2003.

[10] SupriyaMaheshwari, An Efficient QoS Scheduling Architecture for IEEE 802.16 Wireless MANs, Master's thesis, Indian Institute of Technology, Bombay, c2005.

[11] Qingwen Liu, Xin Wang, Georgios B. Giannakis, "Cross- Layer Scheduler Design with QoS Support forWireless AccessNetworks", Second International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks, Aug. 22-24, 2005.

[12] Huei-Wen Ferng and Han-Yu Liau, "Design of Fair Scheduling Schemes for the QoS-Oriented Wireless LAN", VOL. 8, JULY 2009 pp 880-894.

[13] Bader Al-Manthar, HossamHassanein, Najah Abu Ali,andNidal Nasser, "Fair Class-Based Downlink Scheduling with Revenue Considerations in Next Generation Broadband Wireless Access Systems", VOL. 8, NO. 6, JUNE 2009 pp721-734.

[14] Amir Esmailpour and Nidal Nasser, "Packet Scheduling Scheme with Quality of Service Support for Mobile WiMAX Networks" 20-23 October 2009 pp,1040-1045.

Performance Analysis of Total Inter-Carrier Interference for MC-CDMA System in Mobile Environment

Ravinder S. Bisht[1] and Dr. A.K. Garg[2]

[1]Electronics & Communication Department, M M University, India
E-mail: ravibisht48@gmail.com
[2]Electronics& Communication Department, M M University, India
E-mail: garg_amit03@yahoo.co.in

Abstract
Multi-carrier code division multiple access (MCCDMA) has been considered as a strong candidate for next generation wireless communication system due to its excellent performance in multi-path fading channel and simple receiver structure. Recent advances in wireless communications have made use of MC-CDMA and OFDM techniques to allow for high data rate transmission. Rapid time variations of the wireless communication channel have a effect on the performance of multicarrier modulation.Many ICI cancellation methods such as windowing and frequency domain coding have been proposed in the literature to cancel ICI and improve the BER performance for multi-carrier transmission technologies. Other frequency-domain coding methods do not reduce the data rate, but produce less reduction in ICI as well. In this thesis, my main objective is to evaluate the Inter channel Interference which include the Signal to Interference Ratio (SIR) and Inter Carrier Interference (ICI) in a MC DS CDMA wireless system. I find out the analytical results of interference. Simulations are given to support the system and receiver design. All the simulation is carried out on MATLAB tool.

Keywords: Doppler effect, fading channels, intercarrier interference, multicarrier code division multiple access (MC-CDMA), multicarrier modulation, orthogonal frequency division multiplexing (OFDM).

Introduction

Wireless communication is the transfer of information over a distance without the use of electrical conductors or "wires". The distances involved may be short (a few meters as in television remote control) or long (thousands or millions of kilometers for radio communications). When the context is clear, the term is often shortened to "wireless". Wireless communication is generally considered to be a branch of telecommunications. It encompasses various types of fixed, mobile, and portable two-way radios, cellular telephones, personal digital assistants (PDAs), and wireless networking. Other examples of include GPS units, garage door openers and or garage doors, wireless computer mice, keyboards and headsets, television and cordless telephones.

The demand for wireless communications services has grown tremendously. Although the deployment of 3rd generation cellular systems has been slower than was first anticipated, researchers are already investigating 4th generation (4G) systems. These systems will transmit at much higher rates than the actual 2G systems, and even 3G systems, in an ever crowded frequency spectrum. Signals in wireless communication environments are impaired by fading and multipath delay spread [23]. This leads to a degradation of the overall performance of the systems. Hence, several avenues are available to mitigate these impairments and fulfill the increasing demands.

Next-Generation Mobile Broadband Technologies

The next-generation of IMT systems based on CDMA and OFDM, as well as broadcast technologies, will be key enablers of the transition to the next dimension of wireless broadband capabilities and services. In particular, mobile broadband technologies such as CDMA2000 EV-DO Revision B (Rev. B), HSPA+, Long Term Evolution (LTE), and Mobile WiMAX (802.16m) will support multi-megabit-per-second data delivery to users, carrier-grade VoIP and other real-time and broadband intensive applications. For specific bandwidth-intensive applications such as multicasting and broadcasting, OFDM-based technologies such as DVB-H, FLO, ISDB-T, S-DMB and T-DBM have been commercialized since 2006 [30].

CDMA System

In CDMA systems, the narrowband message signal is multiplied by a very large bandwidth signal is a pseudo-noise code sequence that has a chip rate which is orders of magnitudes greater than the data rate of the message. All users in a CDMA system, use the same carrier frequency and may transmit simultaneously. Each user has its own pseudorandom codeword which is approximately orthogonal to all other code words. The receiver performs a time correlation operation to detect only the specific desired codeword. All other code words appear as noise due to decorrelation. For detection of the message signal, the receiver needs to know the codeword used by the transmitter. Each user operates independently with no knowledge of the other users.

CDMA is achieved by modulating the data signal by pseudo-random noise sequence (PN Code), which has a chip rate higher than the bit rate of the data as shown in Fig.1 The PN code sequence is a sequence of ones and zeros (called chips), which alternate in a random fashion. Modulating the data with this PN sequence generates the CDMA signal. The

modulation is performed by multiplying the data (XOR operator for binary signals) with the PN sequence.

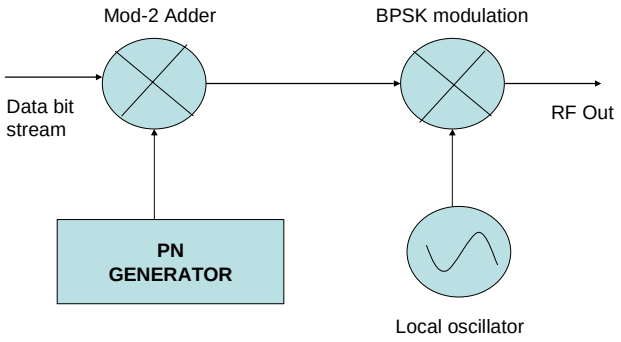


Figure 1: Basic CDMA Transmitter.

MC-CDMA System

The previous chapter presented an overview of OFDM systems, the importance of cyclic prefix and the analysis of Inter Carrier Interference in OFDM. OFDM is an effective technique to combat the frequency selectivity of the channel. Code Division Multiple Access (CDMA) has been a strong candidate to support multimedia mobile services because it has the ability to cope up with the asynchronous nature of the multimedia traffic and can provide higher capacity as opposed to the conventional access schemes such as TDMA or FDMA. By employing Rake receivers CDMA systems can coherently combine the multipath components due to the hostile frequency selective channel. The processing gain due to spreading provides robustness to the multi-user interference. The use of conventional CDMA does not seem to be realistic when the data rates go up to a hundred megabits per second due to severe ISI and the difficulty in synchronizing a fast sequence. Techniques for reducing the symbol and chip rate are essential in this case [9].

Simulation of Gaussian Noise

This channel is affected by Gaussian noise. It is very easy to simulate Gaussian noise in Matlab. The "randn" command was used to generate normally distributed noise as shown in the Fig.2 below [28].

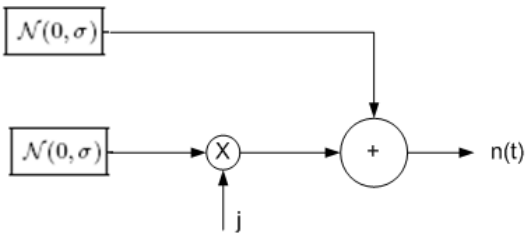


Figure 5.2: Gaussian noise.

In the Fig.5.3, there are two zero mean Gaussian generators. The non-complex part refers to the in-phase component of the noise and complex part refers to the quadrature component. Both components are at right angle to each other, therefore both of them will be normally distributed at right angles. Gaussian noise in time domain looks like random fluctuations with amplitude depending upon the power of noise. In the following Fig.5.4, the complex Gaussian noise can be seen for 200 input samples. The input samples were taken as zeros to observe noise. If we consider the pdf of this noise, it follows Gaussian distribution. So, it can be seen that even if there is nothing transmitted the receiver might detect something which can cause errors. Modulation is used in order to handle this problem.

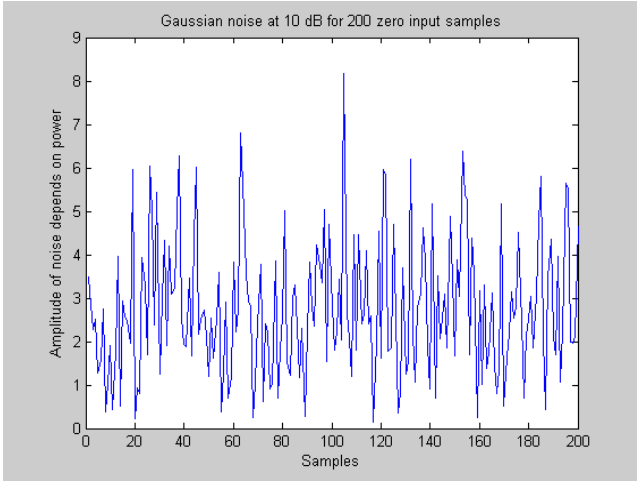


Figure 5.3: Noise simulation flow chart.

Conclusion

This review paper consists analyze the future generation wireless communication system.This paper go through the concept of multicarrier modulation and CDMA concept and analyse the Multicarrier CDMA system. This topic is basically focused on the MC DS CDMA system and its analysis to evaluate the Inter channel Interference which include the Signal to Interference Ratio (SIR) and Inter Carrier Interference (ICI) in a MC DS CDMA wireless system.On Complete analysis of interchannel interference conclusion can be made that when the interchannel interference is low, i.e., small filter overlap, the results obtained with the Chi-Square Approximation are valid.

References

[1] Xue Li, Ruolin Zhou, Steven Hong, and Zhiqiang Wu , "Total Inter-Carrier Interference Cancellation for MC-CDMA System in Mobile Environment" , IEEE Communications Society subject matter experts for publication in the IEEE Globecom 2010 proceedings.

[2] V.Jagan naveen, K.Murali Krishna, K.RajaRajeswari "Performance Analysis Of MC-CDMA And OFDM In Wireless Rayleigh Channel" International Journal of Advanced Science and TechnologInternational Technology Vol. 25, December, 2010

[3] Tao LUO, Gang WU, and Shaoqian LI, Yong Liang GUAN, Choi Look LAW, "DS-CDMA and MC-

CDMA with Per-User MMSE Frequency Domain Equalization", International journal of hybrid information technology, vol 1.1,No. 3, july2008

[4] V.Nagarajan and P.Dananjayan, "Performance Enhancement Of MC-DS/CDMA System Using Chaotic Spreading Sequence" , Journal of Theoretical and Applied Information Technology, 2005 - 2008 JATIT.

[5] Arvind Kumar, and Rajoo Pandey "An Improved ICI Self-Cancellation Scheme for Multi-Carrier Communication Systems" World Academy of Science, Engineering and Technology 32 2007

[6] Salih M. Salih, N. Uzunoglu, A. A. Ali and Waleed A. Mahmud "A Proposed Improvement Model of MC-CDMA Based FFT in AWGN Channel" Pg.517-520, IEEE 2007.

[7] Hen-Geul Yeh, Yuan-Kwei Chang, and Babak Hassibi "A Scheme for Cancelling Intercarrier Interference using Conjugate Transmission in Multicarrier Communication Systems" IEEE Transactions On Wireless Communications, VOL. 6, NO. 1, January 2007

[8] Mohd. Hasan, Tughrul Arslan, and John S. Thompson, "Low-Power-Adaptive MC-CDMA Receiver Architecture" , ETRI Journal, Volume 29, Number 1, February 2007

[9] Ranjith Padmasiri Takeo Fujii Yukihiro Kamiya Yasuo Suzuki , "Variable Spreading Method for MC/DS-CDMA Road to Vehicle Communication System" , 2006 IEEE Ninth International Symposium on Spread Spectrum Techniques and Applications

[10] Vasu Chakravarthy, Abel S. Nunez, And James P. Stephens, "TDCS, OFDM, AND MC-CDMA: A Brief Tutorial", IEEE Radio Communications, September 2005

[11] Xiaodong Cai , Georgios B. Giannakis, "Bounding Performance and Suppressing Intercarrier Interference in Wireless Mobile OFDM" IEEE Transactions On Communications, VOL. 51, NO. 12, December 2003

[12] A. C. Mccormick and E.A. Al-susa, "Multicarrier CDMA for Future generation mobile communication" , Electronics & Communication Engineering Journal, april 2002

[13] Jean-Paul Linnartz and Nathan Yee, "Multi-Carrier CDMA in Rayleigh Fading Channel", IEEE 2002

Low Power Strategies for Network Processors: A Survey

Roopa Kulkarni[1] and Dr. S.Y. Kulkarni[2]

[1]Department of Electronics and Communication Engg., Gogte Institute of Technology, Belgaum, India
[2]Principal, N.M.A.M.I.T., Nitte, India
E-mail: [1]roopa.patavardhan@gmail.com, [2]sy_kul@yahoo.com

Abstract
Network processors have emerged over the years which incorporate multiprocessing and multithreading. These are used in the networking and routers, hence leads to a lot of switching activity. The switching activity leads to power consumption. Thus, in this paper a survey of Network Processors (NP) and the different low power strategies are done. The design for a modular NP architecture is also presented. Power is the key factor in the design of VLSI. Hence power dissipation, its estimation, analysis, optimization and management is the major concern for NP. Hence, this paper gives insight into the different NPs that are available, their architectural and functional parameter through a comparative study. A brief overview of the low power techniques at different design levels and the strategies are also being discussed. Thus, this paper is an effort put to understand the different NPs and low power strategies.

Keywords: Low Power, Network Processors, RTL, ASIP and ATGP

Introduction

In the past years most networking functions above the physical layer have been implemented by software running on general purpose processors. With the growth in the internet and networking, new solutions have come up to cope with the traffic explosion. Though earlier, hardwired solutions appeared [2], but were not scalable. Hence, the need for customizability, in the field programmability and shrinking the time to market had focussed most activity on Network Processors (NP). The network implementations are based on board categories namely: ASIC, ASIP, Co-processors which are possibly configurable solutions with limited programming interface, FPGA and GPP. Analysing the space of system implementation using different categories, it is observed that ASIC is less flexible with high performance while GPP are more flexible with less performance. The rest have some percentage of flexibility and performance. Thus network processors are part of broader movement from ASICs to programmable system implementations. A network processor is an ASIP for the networking application domain – a software programmable device with architectural features and/or special circuitry for packet processing.

There are a number of challenges for the implementation of NPs and power consumption is one of them. As the NP's incorporate multiprocessing and multi threading concept, and used in routers, there involves a lot of switching activity and scheduling taking place and hence leading to power consumption. Thus, there is a need to design NPs that are power efficient.

This paper is an overview of different network processors and low techniques that can be applied to NPs. As NPs is application specific, the low power techniques can be applied based on the application. In section II, a brief description about NPs is done. Section III gives an insight about different low power techniques and strategies. The section IV deals with applying low power techniques to NPs is discussed. Finally the concluding remarks are noted in section V.

Network Processors

Now-a-days network implementation is based on FPGA for low level processing and General Purpose Processors for higher level processing. Independently these will not meet the processing demands. Hence, an alternative solution for system implementation can be either with ASIC, ASIP, Co-processors, FPGA and GPP individually or with the combination of these.

Network processors perform central functions in network elements (NE) [5]. The requirement of an NE both in functionality and performance are dependent on the target market and application area. The basic performance requirement of NEs is scalability, throughput and low power consumption. Therefore a network processor unit is a SoC that includes a highly integrated set of programmable or hardwired accelerated engines, a memory subsystem, high speed interconnect, and media interfaces to handle packet processing at wire speed.

The different network processors available are Agere (payload Plus), Alchemy (Au100), BRECIS Communication, Bay Microsystems, Applied Micro circuits (mP7xxx), Broadcom (Mercurian SB-1250), Cisco (PXF/Toaster2), Clearspeed, Clearwater Networks (CNP810SP), Cognigine, EZChip (NP-1), IBM Power NP, Intel IXP1200, Motorola C-5 DCP, PMC-Sierra RM7000, PRISM IQ 2000, Xelerated and Packet Devices (X40 & T40). The study of these architectures is made and the comparisons of these are tabulated in table 1.

Once the network processor is designed, there is a need of a development platform which helps in the system-level exploration of the processors. One such is the stepNP [3]. The StepNP platform has 3 main frameworks: the application software development platform, the Network Processor unit architecture simulation platform, and the SoC tools platform. Click modular router framework is chosen for its modularity,

flexibility and ease of reconfiguration. The architecture simulation is done using SystemC 2.0 modelling language. The StepNP architecture platform includes models of ARM V4 and PowerPC & Standford DLX processor model. New co-processor models can be integrated into step NP using SystemC open core protocol (SOCP) communication channel interface. Performance of functional SOCP channel on solaris sun U80 and LINUX PC is tabulated and observation is that, the transaction time made a difference in simulation speed for Solaris platform.

Table 1: Architectural and Software Comparisons Of NPs [1].

NP	Central control	Multi-PE	Interface	Compilers	Operating Systems	Libraries
Cognigine	None	16 RCUs	SPI -4, PCI	C/C++ compiler		application library for common L2-7 functions
EZChip (NP-1)	None	64 (TOPs)		C compiler and assembler		
IBM Power NP	On chip power PC core	16 programmable protocol processors	40 Fast Ethernet/ 4Gb MACs with SMII and GMII, POS	Assembler only		None
Intel IXP1200	on-chip 200MHz Strong ARM coordinates system activities	6 programmable microengines	4.2Gb/s 66MHz IX bus, PCI	C compiler and Assembler		None
Motorola C-5 DCP	1 executive processor	16-channel processors	33/66MHz PCI, UTOPIA (Level 2 and 3)	C/C++ compiler		CPI that abstracts common networking tasks
PRISM IQ 2000	None	4 CPUs for route processing and system management		C/C++ compiler	Windows CE, Linux, VxWorks	None
Applied Micro circuits (mP7xxx)	None	Yes	Fast ethernet	C/C++ compiler	Wind River	Network software reference library

.

Low Power Strategies

Requirement for low power consumption continue to increase significantly as components become battery-powered, smaller and require more functionality. The level-by-level power analysis and estimation tools are available that leads to fast and accurate results. The different tools available are as shown in figure 1 [4]. As seen from the flow there are broadly three levels of power estimation namely [4]: Software-level, behavioural level and RT level Estimation.

In the software level approach initially the architectural simulation is performed, then the characteristic profile is extracted next, the mixed integer linear programming and heuristic rules are used to obtain a fully functional program. Finally, the RT-level simulation is performed of the new synthesized program. In behavioural power estimation the physical capacitance and the switching activity is considered. In RT-level estimation regression-based, switched-capacitance models techniques for circuit modules are used. The basic low power design techniques, such as clock gating for reducing the dynamic power, or multiple voltage thresholds to decrease leakage current, are established and supported existing tool.

The different low power strategies[6] available at different levels of VLSI design process for optimizing power is as: at the technology level, strategies are threshold reduction and multi threshold devices; at circuit/logic level, strategies are logic styles, transistor sizing and energy recovery; at architecture level, strategies are pipelining, redundancy and data encoding; at software level, strategies are regularity, locality and concurrency, and at operating system level, strategies are portioning and power down.

Apart from designing low power circuits it is equally important to know different low power testing techniques for SOCs or even the system level design. In [6], the author describes the reasons and effects of high power consumption during test. The low power testing schemes can be divided into two categories: one for the external circuits and the other for internal circuits. Apart from the basic testing techniques for the circuits, there needs to be modification for the SOCs. The list of power reduction techniques are: modification in LFSR, partitioning the circuits, separate testing strategy for memory, improved ATGP algorithm ordering technique and exploring don’t care bits.

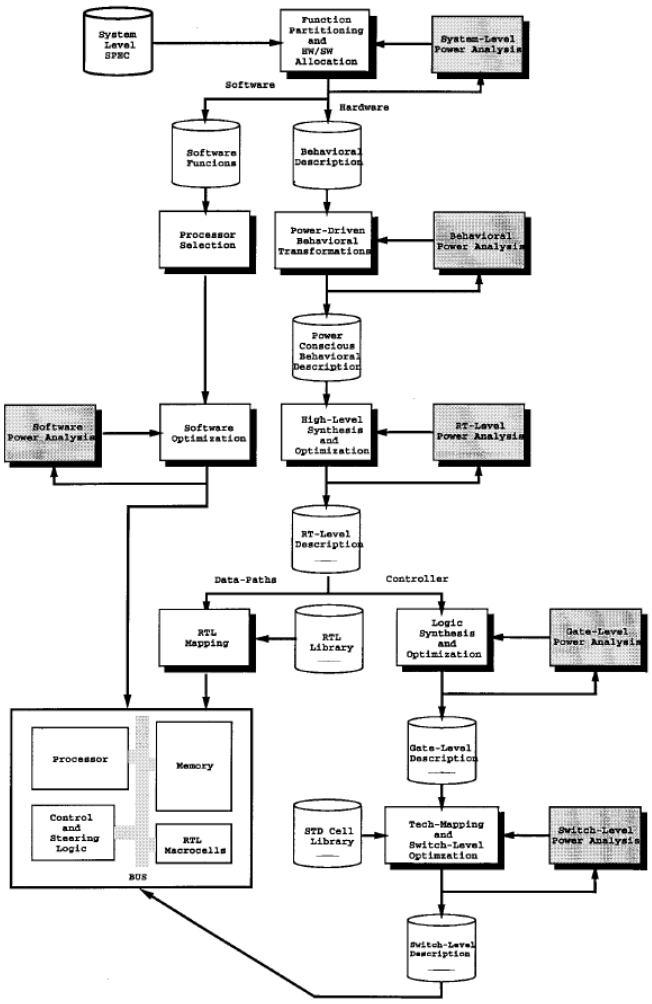


Figure 1: Low Power Design Flow [4].

Low power techniques for network processors

In this section a few examples are coated to explain how the analysis varies at different level of design. For the security system [6] there needs to be a variation at the architectural level. The architecture template for security is mainly assembled with a main controller, a DMA module, internal buses, and crypto modules. Here, the scheduling algorithm and the DMA module is redesigned and implemented for power optimization. As discussed in the section III, low power strategies are applied at different levels of design.

System level modeling with executable languages such as C/C++ or other modeling frameworks have been crucial in designing a large electronic system. One essential approach is to develop a cycle-level accurate simulator. Most cycle-level accurate simulators only report power and performance data for worst and/or average cases, which limits the capability of power and performance analysis. The assertion-based [b] analysis methodology is very suitable for transaction-level or cycle-level power and performance analysis for NPU designs. In [7], the author through the experimental results shows that assertion-based methodology is an efficient tool to help designers analyze power and performance characteristic of design and choose the configuration.

New generation of NPs offer high performance through parallel processing architecture, this incorporates multiple processing elements (PEs) configured as either independent or pipelined units. Processing elements (PE) are a part of the network processors. Network processors when used in router, it is observed that, during low traffic many of the PEs are idle but still consumed some amount of dynamic power. This PEs is to be put OFF when not in use. This means that a thread is to be killed, which again depends on the status [8] of the thread. The author explains the challenges involved in gating OFF these PEs.

From these surveys, it is analyzed that the power optimization, estimation and reduction can be done based on the application and the design level. The power analysis or the performance evaluation of NPs for power depends on the application that is developed and at what level the optimization is required.

Conclusion

This paper gives an insight into the different Network Processors available, giving in detail the features, architecture, compiler used and software required for the usage. Later, an exhaustive section is on power modeling. In this paper an effort is put to study the different network processors and different low power techniques, and power modeling and estimation. From the survey of applications it is analyzed that low power strategies and techniques can be applied at different level of design flow which depends on the application developed.

References

[1] N.Shah, "Understanding Network Processors," Internal Report, Dept. of Electrical Engg. and Computer Science, University of California, Berkeley.2001.

[2] Agere, Inc. "The Challenge for Next Generation Network Processors." White Paper. Agere, Inc. September 10, 1999.

[3] Paulin P.G., Pilkington C., and Bensoudane E. "*StepNP: a system-level exploration platform for network processors*," IEEE Design & Test of Computers, Vol. 19, Issue 6, pp 17-26.

[4] Enrico Macii, Massoud Pedram, and Fabio SomenziX ,"High Level Power Modelling Estimation and Optimization", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Syatems, Vol. 17, NO. 11, Nov 1998

[5] MariaGabrani, Gero Dittmann, Andreas Doring, Andreas Herkersdorf, Patricia Sagmeister, Jan van Lunteren,"Design Methodology for a Modular Service-Driven Network Processor Architecture", IBM Zurich Research Laboratory, Oct 2002.

[6] Kanika Kaur and Arti Noor, "*Strategies and Methodologies for Low Power VLSI Designs: A Review*," International Journal of Advances in Engineering and Technology, Vol. 1, Issue 2, pp 159-165, May 2011.

[7] Yi-Ping You, Chun-Yen Tseng, Yu-Hui Huang,Po-Chiun Huang, TingTing Hwang, and Sheng-Yu Hsu,"Low Power Techniques for Network Security Processors", Proceedings of the ASP-DAC 2005. Asia and South Pacific, Design Automation Conference, 2005, pp 355-360.

[8] Jia Yu, Wei Wu, Xi Chen, Harry Hsieh, Jun Yang and Felice Balarin," Assertion-Based Power/Performance Analysis of Network Processor Architectures" Ninth IEEE International High-Level Design Validation and Test Workshop, 2004, pp 155-160.

[9] Yan Luo, Jia Yu ,Jun Yang and Laxmi Bhuyan," Low Power Network Processor Design Using Clock Gating", Proceedings of 42nd Design Automation Conference, 2005, pp 712-715.

An Empirical Evaluation of Fuzzy and Counter based Handoff Systems for the avoidance of Ping-Pong Effect

Randheer Singh[#], Surender Singh Dahiya[@] and Amit Doegar[#]

[#]National Institute of Technical Teachers' Training & Research, Chandigarh, India
[@]Shivalik Institute of Engineering & Technology, Aliyaspur, Ambala, India
E-mail: chauhan.randheer@yahoo.co.in,surendahiya@gmail.com,amit@nitttr.ac.in

Abstract
Handoff in cellular systems is a complex problem. Handoff is to be initiated and executed at an appropriate time so as to ensure the quality of service (QoS). While a handoff is initiated to ensure QoS, the rate of handoff has to be kept low to save cost. A hasty handoff can cause ping pong effect whereas a delay can cause QoS to degrade or even cause the call to drop. Under such conflicting requirements, an intelligent handoff system is the need that reduces the handoff rate to minimum while maintaining QoS. Fuzzy logic is the one of the intelligent control strategies, which is used for such conflicting control requirements. In this paper, a fuzzy system is presented which decide about the handoff requirement based on the signal strength of approaching tower and receding tower. Handover is assumed to switch between global system for mobile communication (GSM) and wideband code division multiple access (WCDMA) networks. The proposed inter system fuzzy based handoff system is compared with the counter based traditional system with a view of alleviating the effect of ping pong phenomenon and also optimizing handoff cost without degrading the quality of service. .

Introduction
The future communication depends upon many types of co-existing mobile communication Systems/Networks normally GSM Networks, WCDMA Networks, Satellite communication based networks, intelligent transportation systems, high altitude stratospheric platform systems (HAPS). An individual have to move from one area covered by one type system to another area covered by different kind of Technology/Network. Under the given scenario when a mobile system moves from one system to the area covered by another system, the handover of the ongoing call should be carried out at inter-system level. This inter-system handoff assumes significance for seamless communication without call being dropped.

A handoff algorithm must be fast, reliable and must maintain QoS and minimal Congestion, Call Blocking. The more the attempts for handoff, the more chances that a call will be denied access to a channel, resulting in a higher handoff call dropping probability which is not desirable. In order to deal with such complexities associated with the handover, there exists a need for some new systems, which are intelligent enough to predict the timely need for the handover and are capable of executing the same swiftly once initiated. This intelligence can be imparted to the system using one of the emerging technologies such as fuzzy logic, neural networks, particle swarm optimization etc. In this paper, a fuzzy handoff system based on signal strengths of competing networks, trigger threshold and, degree of stability trigger threshold has been proposed, which has a far reaching potential as building blocks in tomorrow's computational world. In this, a comparative analysis of the proposed system with traditional algorithms is made with a view of alleviating the effect of ping pong effect and also optimizing handoff cost without degrading the quality of service.

Ping-Pong Effect
According to Hierarchical Optimization Handover Algorithm (HOHA) proposed by Lim and Wong [7] handoff is initiated on the basis of the values of three thresholds as shown in figure 1 [7]; registration threshold R_t, is the threshold value which is sufficient for continuation of a call handover threshold H_t, is the threshold value when there is need of the handoff from one base station (BS) to another BS critical threshold C_t, comes into existence when the strength of signal reaches at very poor level. At C_t, call dropped automatically. Sometimes when the user moves from one base station to another, then signal strength of first BS goes on decreasing while the signal strength of another goes on increasing. This variation in signal strength causes an effect known as ping-pong effect. This causes to drop in call. When the user moving across cell boundary, ping-pong affects the call at very high rate and call drop rate reaches very high. This increases the cost as well as number of handoff. A scenario of GSM and WCDMA system is shown in figure 2 generated from MATLAB. In this, ping pong effect is shown from 1000 meters to 1110 meters. This effect should be minimized to increase the QoS. In this paper, a fuzzy based ping pong affect avoidance system is proposed and the result is compared with other traditional systems of ping-pong avoidance. Section 2 of this paper presents literature survey on the selected topic. The proposed methodology and system model is discusses in section 3.

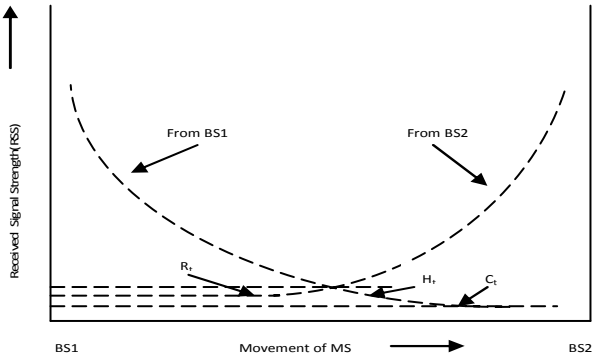


Figure 1: Handover Scenario

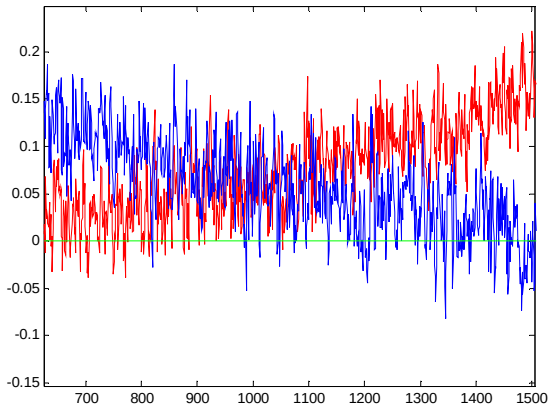


Figure 2: Ping-Pong effect

Literature Survey

Tripathi has categorized and compared several conventional and emerging Handoff Algorithms on the basis of several techniques [14]. He listed out the two pre conditions for handover initiation. The first condition necessitates that the average signal strength of the serving BS falls below an absolute threshold and second condition specifies that the average signal strength of the candidate BS should exceed the average signal strength of the current BS by an amount of hysteresis. The first condition prevents the occurrence of handoff when the current BS can provide sufficient signal quality. An optimum threshold achieves the narrowed handoff area (and hence reduced interference) and a low expected number of handoffs. Tripathi *et al.* [8] also discussed that the relative distance measurement is obtained by comparing propagation delay times. This criterion allows handoff at the planned cell boundaries, giving better spectrum efficiency compared to the signal strength criterion. However, it is difficult to plan cell boundaries in a microcellular system due to complex propagation characteristics. Thus, the advantage of distance criterion over signal strength criterion begins to disappear for smaller cells due to inaccuracies in distance measurements. Singh *et al.* [6], discussed several variations of signal strength based algorithms, including relative signal strength algorithms, absolute signal strength algorithms, and combined absolute and relative signal strength algorithms. Pollini [10] discussed the approaches avoidance ping pong effect based on the hysteresis margin and received signal strength (RSS) value. Handoff will be initiated only if the RSS of new BS is higher than serving BS by a certain margin of hysteresis. This hysteresis value is designed to reduce the ping pong effect in handoff procedure and handoff performance optimization. Roy [4] has done the evaluation of various handover algorithms. In this, handover delay, shadow fading and effects of averaging are evaluated. Absolute signal strength based algorithm has to be considered for intersystem handoff, as relative measurements are not possible for different cellular systems because of their different power requirement and other criteria.

Lee [16] provided a two-level algorithm in which there is a variation of the threshold, which provides more opportunity for a successful handoff. It suggested that for good quality voice, signal interference ratio (SIR) at the cell boundary should be relatively high (e.g., 18 dB for WCDMA and 12 dB for GSM). However, a lower SIR may be used for capacity reasons since co channel distance and cluster size (i.e., the number of cells per cluster) are small for lower SIR and channels can be reused more frequently in a given geographical region. This algorithm makes a handoff when the current BS's SIR drops below a threshold and another BS can provide sufficient SIR. Hysteresis can be incorporated in the algorithm. The lower SIR may be due to high interference or low carrier power. In either case, handoff is desirable when SIR is low. However, SIR-based handoff algorithms prevent handoffs near nominal cell boundaries and cause cell dragging and high transmit power requirements. Edwards *et al.* [9], gave the concept of adaptive velocity handoff algorithm that can serve as an alternative to the umbrella cell approach to tackle high speed users if low network delay can be achieved, which can lead to savings in the infrastructure. In [15], one of the velocity estimation techniques uses level crossing rate (LCR) of the RSS in which the threshold level should be set as the average value of the Rayleigh distribution of the RSS [15], requiring special equipment to detect the propagation dependent average receiver power. In [8], a direction biased handoff algorithm represents such an alternative solution. Direction biasing improves cell membership properties and handoff performance in line of sight (LOS) and non LOS (NLOS) scenarios in a multi-cell environment. In [9, 13, 23], RSS and bit error rate (BER) based algorithms are described.

Another category of algorithms is emerging. Pattern recognition (PR) techniques can help reduce this uncertainty by efficiently processing the (RSS) measurements [14]. It is better than the relative signal strength algorithm and the combined absolute and relative signal strength algorithm via simulations. The final handoff decision is made based on the calculated handoff priority. Edwards [9], presented a signal strength based handoff initiation algorithm using a binary hypothesis test implemented as a neural network. Kwong *et al.* [2], proposed a newer approach using Adaptive Network Fuzzy Inference System (ANFIS) where the training element is incorporated into the existing fuzzy handoff algorithm. Sheng Jie *et al.* [11], proposed a triangle module operator and fuzzy logic based handoff algorithm for heterogeneous wireless networks. It adapts fuzzy logic algorithm to fulfill the fuzzy decision values of RSS based algorithm and QoS

parameters based algorithm respectively. This method can not only make a more reasonable choice of network, but also ensure the QoS of the network, and reduce the switching frequency and the ping-pong effect effectively. Israt *et al.* [3], have developed a new fuzzy logic-based adaptive handoff (FLAH) management protocol for next generation wireless system for seamless communication. This estimates the precise time for initiation of handoff in intra-system and inter-system.

Sharma *et al.* [1], have proposed intelligent approach for handover decision in heterogeneous wireless environment using sugeno type fuzzy system. In this, RSS, network availability and bandwidth are taken as fuzzy input variables and handoff decision is considered as output variable of fuzzy system. Barolli *et al.* [12], proposed a Fuzzy-based handover system for avoiding Ping-Pong effect in cellular networks using Monte Carlo technique. Kassar *et al.* [5], proposed an intelligent approach for handover in intersystem environment for future generation wireless networks with the concept of always best connected (ABC) for seamless connectivity through the different available networks. Singh *et al.* [13], proposed a fuzzy based multicriteria handoff algorithm. The fuzzy handoff algorithm has been shown to possess enhanced stability (i.e., less frequent handoffs). In [9], a fuzzy classifier is used to process the signal strength measurements to select a BS to serve a call. The performance of this algorithm in a microcellular environment is evaluated.

Proposed Methodology

Simulation Environment

To better understand the concept, a basic system consisting of two BSs separated by a distance of *D* is considered in this paper as shown in the figure 3. Both of the BS(s) are assumed to be located in the center of the respective cell and operating at the equal transmitting power. Hexagonal geometry of the cell has been considered. The user equipment (UE) moves from one cell to another along a straight line trajectory with constant speed.

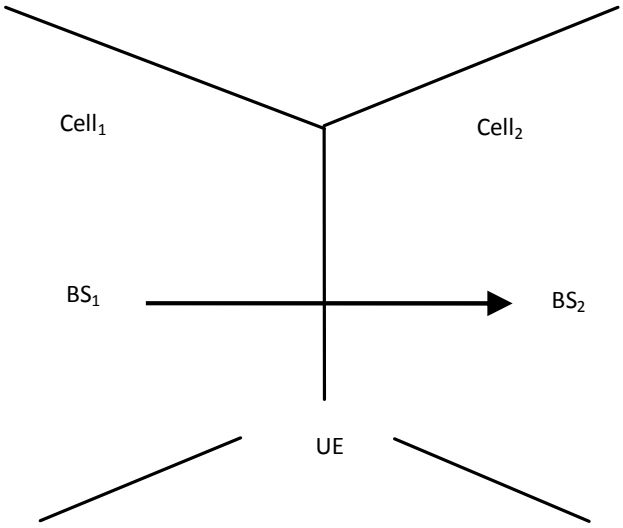


Figure 3: System Model

RSS can be evaluated by outdoor propagation path loss models [17-18]. Shadowing is caused due to the obstruction of the line of sight path between transmitter and receiver by buildings, hills, trees and foliage. Multipath fading is due to multipath reflection of a transmitted wave by objects such as houses, buildings, other man-made structures, or natural objects such as forests surrounding the UE. The UE measures RSS from each BS. The measured value of RSS (in dB) is the sum of two terms, one due to path loss and the other due to lognormal shadow fading. The propagation attenuation is generally modeled as the product of the η^{th} power of distance and a log normal component representing shadow fading losses [10]. These represent slowly varying variations even for users in motion and apply to both reverse and forward links. For UE at a distance'd' from BSi, attenuation is proportional to

$$\alpha\,(d,\zeta) = d^n\,10^{\frac{\zeta}{10}} \qquad (1)$$

where ζ is the dB attenuation due to shadowing, with zero mean and standard deviation σ. Alternatively, the losses in dB are

$$\alpha(d,\zeta)[dB] = 10\eta \log d + \zeta \qquad (2)$$

It is neglected for handover initiation trigger due to its short correlation distance relative to that of shadow fading. The autocorrelation function between two adjacent shadow fading samples is described by a negative exponential function as given in [10]. Let d_i denote the distance between the UE and BS_i, i=1, 2. Therefore, if the transmitted power of BS is P_t, the signal strength from BS_i, denoted S_i,(d) i=1, 2, can be written as

$$S_i(d) = P_t - \alpha(d_i,\zeta) \qquad (3)$$

The measurements are averaged using a rectangular averaging window to alleviate the effect of shadow fading according to the following formula.

$$\hat{S}_i(k) = \frac{1}{N_w}\sum_{n=0}^{N-1} S_i\,(k-n)W_n, \qquad i=1,2 \qquad (4)$$

where $\hat{S}_i$ is the averaged signal strength and S_i is the signal strength before averaging process. W_n is the weight assigned to the sample taken at the end of *(k - n)* th interval. *N* is the number of samples in the averaging window

$$N_w = \sum_{n=0}^{N-1} w_n$$

In the case of rectangular window W_n = 1 for all *n*.

Using above equations, signal strengths of approaching (WCDMA) and receding towers (GSM) can be generated which can simulate the scenario of shadowing which is responsible for ping-pong effect. In our experimentations, we have generated 2000 data points. In fuzzy whole of the data is generated as a test data and in neural network based system 400 points are used for training purpose and rest is being used as a test data set.

Handover Initiation Algorithms

Counter Based Traditional based Algorithm

A handover is performed if the following conditions are

simultaneously fulfilled: the averaged signal strength from the serving base station falls below a threshold value "*And*" the averaged signal strength from the target base station becomes greater than a preset threshold. This type of preconditions is responsible for creating ping-pong effect if the region in which Mobile station (MS) is travelling has shadowing effect in prominence. To avoid ping-pong effect counter based system can be used in traditional system. In these algorithms, MS can wait for the stabilization of threshold before initiating handover. A counter value is generated based on the stability of threshold value for every data point in data set. If threshold is persisting in last four data points, counter value is assumed as four. If the counter value is increases up to five then this makes the control switches to another network means from GSM to WCDMA or vice-versa. This makes less number of connections released because this creates the persistency in the connection to until the counter value strikes. But when the fuzzy system is used to make this decision for switching, then number of connections released is much less or we can say that it is equivalent to zero. This saves the cost in terms of bandwidth, as it saves to break the connection and making call again and again. The result section represents it effect much clearly.

A Fuzzy Model for Intersystem Handover

It is a proven fact now that fuzzy logic is a powerful problem-solving methodology with wide range of applications in industrial control, consume electronics, management, medicine, expert systems and information technology [19]. It provides a simple way to draw definite conclusions from ambiguous or imprecise and incomplete information. Any fuzzy system consists of four major modules of the system; fuzzification, inference engine, knowledge base and defuzzification module. Fuzzification is a process of mapping input values in crisp sets to values in fuzzy sets. The knowledge base module contains the knowledge of the application domain and the procedural knowledge such (as attendant control goals in the case of a fuzzy controller).it consists of a data base and linguistic (fuzzy) control rule base or production rules. The engine inference simulates the decision making capabilities of human brain. Based on input from fuzzifier, domain knowledge and set of control rules the output decisions or the necessary control actions are evaluated in fuzzy domain. Since usually more than one rules fire there are more than one outputs at a given instant. Inference engine also takes into account this fact by combining these fuzzy outputs into a single fuzzy set. This process of combining a number of fuzzy sets into a single fuzzy set is called aggregation. Finally defuzzification module converts the range of values of output variables into corresponding universe of discourse.

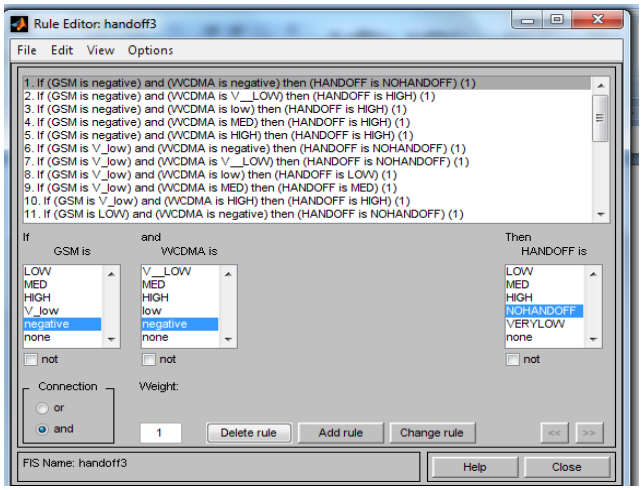


Figure 4(a): Rules Generation

For implementing fuzzy system in our problem, two input variables i.e. GSM and WCDMA have been selected. The signal strength of GSM system ranging from -20 to 250 dB has been partitioned into 5 fuzzy sets namely negative, V_Low, LOW, MED and HIGH. The signal strength of WCDMA system ranging from -20 to 250 dB is also partitioned into five fuzzy sets (membership functions) namely negative, V_Low, LOW, MED and HIGH. The output of the system is chosen as probability of handoff and it is partitioned in five i.e. HIGH, MED, LOW, VERYLOW, and NOHANDOFF singleton values of 0.7, 0.4, 0.2, 0.1 and 0 respectively.

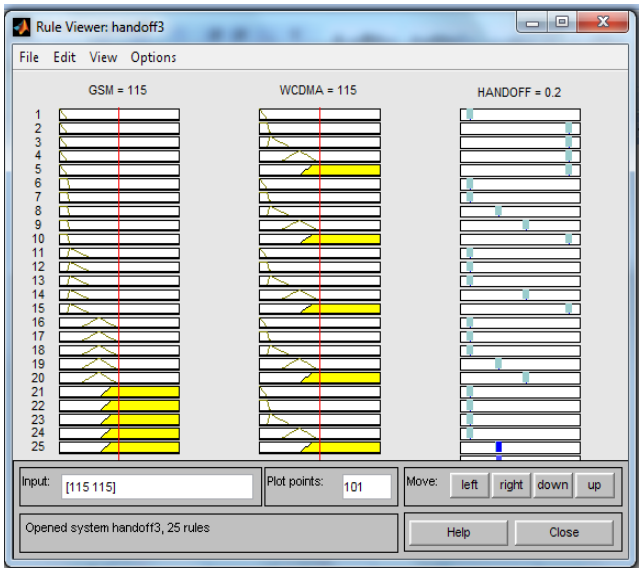


Figure 4(b): Rule Viewer.

After output specification, the next step is rulebase generation. We have created nine rules as shown in figure 4(d). The rulebase generation is the last step of the creation of a fuzzy systemThe system can be checked with the help of rule viewer, which is available in fuzzy toolbox of Matlab.The

rule viewer for the above system is shown in the figure 4(e).in the figure 4(d), the signal strength for gsm is 115 and the signal strength for wcdma is 115. According to these inputs the handoff probability will be 0.2. Hence we can check the output of the fuzzy system by applying any input in the fuzzy rule viewer. This complete procedure will produce a .fis file. This file contains the complete information of the designed system.

Results & Discussion

Table 1 presents the result obtained by the experimentation. The different handoff methods are compared on the basis of two parameters. One parameter representing the total number of handoff and another parameter shows the number of connection releases in the data series of 2000 points simulating ping-pong effects of the signal strength of GSM and WCDMA towers. Both parameters are required to be minimized to effectively decrease the cost of handoff but exhibits opposite characteristics. If we go on minimizing the number of handoff, the total number of connection releases increases. On the other side if connection releases is minimized the handoff frequency is increased. Therefore ideal case for a handoff system will be to make a trade-off between two extreme. In the results, it can be clearly seen that number of handoff is minimum for counter based system with threshold values between 10 to 20 but in these cases number of connection releases are high clearly making it a bad choice for handover. In the case of fuzzy based handoff system there is no connection releases while number of handoff switches are quite low so making it the favorite handoff system. Figure 5 is the graphical representation of the tabular values.

Table 1: Fuzzy Vs Traditional Handoff.

Method Type	Parameters		No of Handoff	No of Connection Release
	MAX Threshold	MAX Counter		
Counter Based Traditional handoff method	20	5	1	17
	10	5	1	10
	5	5	7	9
	20	3	1	7
	10	3	5	6
	5	3	7	5
Threshold based Handoff (Plane method)	0	0	109	0
	5	0	55	0
	10	0	37	0
	20	0	17	4
Fuzzy Based Method	Handoff probability >0.4		5	0

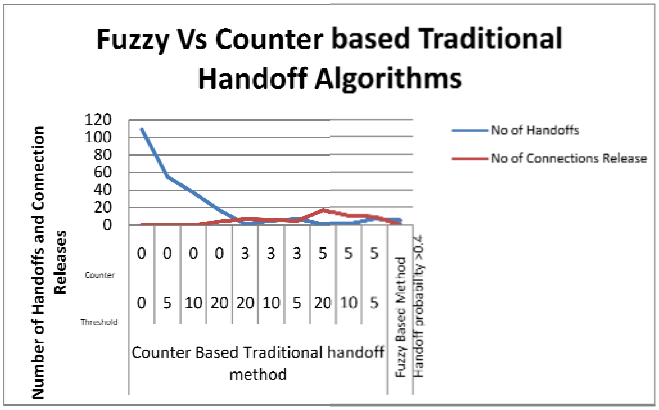


Figure 5. Fuzzy Vs Traditional Handoff.

Conclusion

In this paper a fuzzy based intersystem handoff system is proposed. The handoff is a complex process and in which two conflicting requirements of reduction of handoff rate with maintaining quality of service have to be fulfilled. The fuzzy based handoff system is simulated for an environment containing shadowing effects and is compared with traditional counter based switching systems for their efficiencies of avoidance of ping-pong effect in terms of number of handoffs with number of call drops. The results obtained are very encouraging and clearly validate the supremacy of the approach of fuzzy based handoff system over traditional system.

References

[1] Manoj Sharma & Dr. R. K. Khola, "An Intelligent Approach for Handover Decision in Heterogeneous Wireless Environment", *International Journal of Engineering*, Vol. 4, Dec. 2010, pp. 452-462.

[2] Chiew Foong Kwong, Teong Chee Chuah and Sze Wei Lee, "Adaptive Network Fuzzy Inference System (ANFIS) Handoff Algorithm", *International Journal of Network and Mobile Technologies,* Electronic Version Vol. 1, Nov. 2010, pp. 54-59.

[3] Presila Israt, Namvi Chakma, and M. M. A. Hashem, "A Fuzzy Logic-Based Adaptive Handoff Management Protocol for Next-Generation Wireless Systems", *Journal of Networks*, Vol. 4, Dec. 2009, pp. 931-940.

[4] Sanjay Dhar ROY, "Performance Evaluation of Signal Strength Based Handover Algorithms", *International Journal Communications, Network and System Sciences*, Vol. 2, Oct. 2009, pp. 657-663.

[5] Meriem Kassar, Brigitte Kervella, and Guy Pujolle, "An Intelligent Handover Management System for Future Generation Wireless Networks", *EURASIP Journal on Wireless Communications and Networking*, Vol. 2008, Aug. 2008, pp. 1-12.

[6] B. J. Singh, S. Kumar, K. K. Aggarwal, "Handover Initiation Control Techniques in Mobile Cellular Systems" *IETE Technical Review*, Vol. 20, Jan-Feb

2003, pp. 13-21.

[7] Lim B. L. and Wong Lawrence W. C., "Hierarchical optimization of microcellular call handoffs", *IEEE Transactions on Vehicular Technology*, Vol. 48, Mar. 1999, pp. 459-466.

[8] N. D. Tripathi, N. Jeffrey, H. Reed, and H. F. Vanlandingham, "Handoff in Cellular Systems", IEEE Personal Communication, Vol. 5, Dec. 1998, pp. 26-37.

[9] G. Edwards and R. Sankar, "Microcellular handoff using neuro-fuzzy Techniques", *Wireless Networks*, Vol. 4 , Sep. 1998, pp. 401–409.

[10] G. P. Pollini, "Trends in handover design", *IEEE Communications Magazine*, Vol. 34, Mar. 1996, pp. 82-90.

[11] Sheng Jie and Tang Liangrui, "A Triangle Module Operator and Fuzzy Logic Based Handoff Algorithm for Heterogeneous Wireless Networks", *IEEE International Conference on Communication Technology*, 11-14 Nov. 2010, North China Electric Power University, China, pp. 488 – 491.

[12] Leonard Barolli, Fatos Xhafa, Arjan Durresi and Akio Koyama, "A Fuzzy-based Handover System for Avoiding Ping-Pong Effect in Wireless Cellular Networks", *International Conference on Parallel Processing – Workshops*, 8-12 Sept. 2008, Portland State University, USA, pp. 135-142.

[13] B. J. Singh, S. Kumar and K. K. Aggarwal, "A Fuzzy Based Multicriteria Handover Algorithm for Cellular Systems", *in Proceeding of National Conference on Computer devices for Communication*, 21-23 Feb. 2001, REC, Jalandhar, pp. 8-15.

[14] N. D. Tripathi, "Generic Adaptive Handoff Algorithms using Fuzzy Logic and Neural Networks", Ph.D. Thesis, Aug. 1997, Blackburgs, Virginia.

[15] T. S. Rappaport, "Principles of Mobile Communications", *Pearson Education Asia*, Second Edition, 2002.

[16] W. C. Y. Lee, "Mobile Communications Design Fundamentals", *John Wiley & Sons Incorporation*, Second Edition, 1993.

[17] M. Gudmundson, "Correlation models for Shadow fading in Mobile Radio Systems," *Electronics Letters* Vol. 27, No. 23. November 1991.

[18] G.E. Corazza, D. Giancristofaro, and F. Santucci, "Characterization of Handover Initialization in Cellular Mobile Radio Networks," *Vehicular Technology Conference*, pp. 1869 – 1872, June 1994.

[19] Nasri Sulaiman, Zeyad Assi Obaid, M. H. Marhaban and M. N. Hamidon, "FPGA-Based Fuzzy Logic: Design and Applications – a Review," *IACSIT International Journal of Engineering and Technology,* Vol. 1, No.5, Dec. 2009, pp. 491-503.

Design & Simulation of a Planar Monopole Antenna based on Double E &T Shape Slots

Sandeep Panwar[1], Davinder Parkash[2], Naresh Kumar[1] and Rajesh Khanna[3]

[1]Assistant Professor, ECE Deptt., [2]Associate Professor, ECE Deptt.,
Haryana College of Technology & Management, Kaithal, India
E-mail: devnitk1@gmail.com
[3]Professor, ECE Deptt., Thapar University, Patiala (Punjab), India

Abstract
This paper presents the design of the multiband microstrip antenna for wireless communication system. The proposed antenna is particularly attractive for WLAN/WiMAX devices that integrate multiple systems. The overall size of the design is 48 mm x 35.2 mm x 1.6 mm with a volumetric size of 2.7 cm^3. The proposed designed antenna covers the three frequency bands from 2.3 GHz to 2.8 GHz and from 3.1 GHz to 3.16 GHz (lower-frequency band) and 4.71 GHz to 6.4 GHz (upper-frequency band) such that total bandwidth of the proposed antenna is 2.3 GHz. The parametric study is performed to understand the characteristics of the proposed antenna. Also, good antenna performances such as radiation patterns and antenna gains over the operating bands have been observed. The maximum simulated gain of the proposed antenna is 5.73 dBi at 5.58 GHz band.

Keywords: Monopole Antenna, CPW feeding, WLAN and WiMAX

Introduction

The new technological trend has focused much effort into the design of a Microstrip patch antenna. Advances in wireless communication technologies are placing greater demands on higher antenna impedance bandwidth and smaller antenna size. The microstrip patch antenna is simply a patch which radiates from one face [1, 2]. Bandwidth and efficiency of a Microstrip antenna depends upon patch size, shape, substrate thickness, dielectric constant of substrate, feed point type and its location, etc.. For good antenna performance, a thick dielectric substrate having a low dielectric constant is desirable for higher bandwidth, better efficiency and better radiation, leading to a larger antenna size [3, 4]. These patch antennas are used as simple and for the widest and most demanding applications. Dual characteristics, circular polarizations, dual frequency operation, frequency agility, broad band width, feed line flexibility, beam scanning can be easily obtained from these patch antennas [5]. Some popular antenna designs suitable for WLAN and WiMAX operation for 2.4 GHz band (2.4–2.484 GHz), 5.2/5.8 GHz bands (5.15–5.35 GHz/5.725–5.825 GHz) and 2.5/3.5/5.5 GHz (2500–2690/3400–3690/5250–5850 MHz) bands has been reported in [1-11].

In this paper, a novel approach to achieve a multiband antenna is introduced. The geometry of the proposed antenna is composed of the rectangular patch with a double merged E shape & T shape slots are used to cut the rectangular patch and a small I-shape strip is placed on the radiating sides of the antenna. The impedance bandwidth, gain and radiation characteristics of the proposed planar antennas are examined. The paper is organized as follows: the section 2 presents the antenna design parameters of the printed monopole antenna for multiband operation. After that, results and parameter study of the proposed antennas with a double merged E & T-shape slot structure are described in Section 3. Finally, the paper is summarized in Section 4.

Antenna Structure

The geometry of the proposed triple-band antenna for WLAN/WiMAX applications is shown in Figure 1. The proposed antenna was designed on a low-cost FR4 substrate with height of substrate h_{sub}=1.6 mm, dielectrics constant ϵ_r=4.4 and tangent loss tanδ=0.002. A rectangular patch was chosen as the monopole radiation element. The antenna is fed by a CPW transmission-line, which can be easily integrated with other CPW-based microwave printed circuits on the same substrate. The proposed antenna is composed of rectangular patch with double merged E-shape & T-shape slots. The CPW feed was easy to connect to the coaxial cable through a standard 50 ohm SMA connector. The designed structure was designed & simulated using IE3D simulation software based on method of moments (MOM).

The optimized geometric parameters of the proposed antenna are: length of the rectangular patch L=33 mm, width of the rectangular patch W=23 mm, length of the ground plane L_g=11.35 mm, width of the ground plane W_g=14.5 mm, length of the added slot L_1=7 mm, length of slot L_2=22.3 mm, width of the both slits is W_1=11.51 mm, W_2=15.21 mm, L_3=4 mm,W_3=4.2 mm, L_4=4.2 mm,W_4=2.9 mm. To give feeding to this geometry a feed line of having length L_f =10.2 mm and width W_f =3.9 mm is used. The distance between the ground plane and the rectangular patch is denoted by 'S' that this is equal to 1.72 mm and the distance between the feed line and ground plane is denoted by 'd' is equal to 1.4 mm. The designed antenna covers the frequency band from 2.3-2.8 GHz, 3.1-3.16 GHz and 4.71 GHz to 6.4 GHz such that total bandwidth of the proposed antenna is 2.3 GHz.

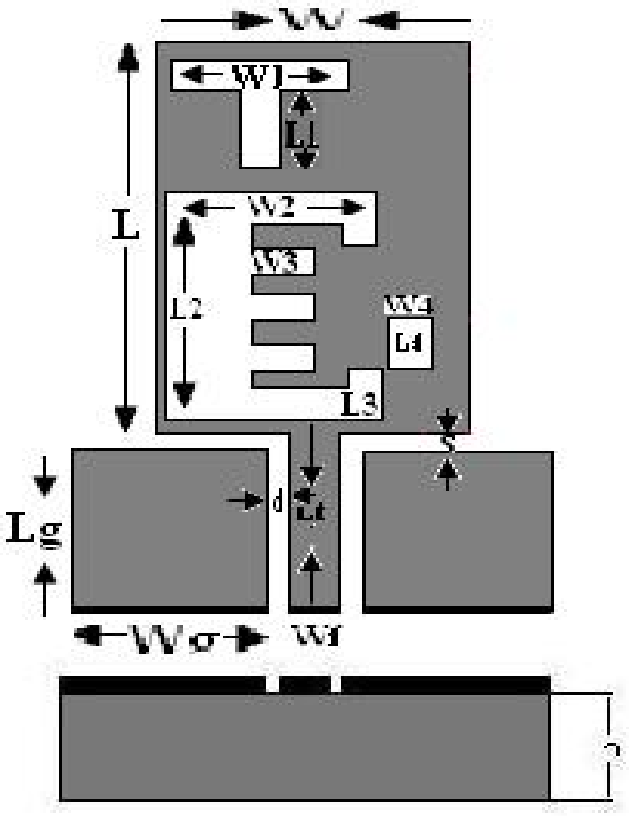


Figure 1 Geometry of the proposed CPW-fed monopole antenna

Simulation Result and Discussions

The simulated return loss and parametric study results for the proposed monopole antenna are obtained. The simulated return loss and gain are presented for the optimized set of antenna parameters. Simulated return loss of the optimized proposed antenna is shown in Figure 2. From the simulated results, it is clear that triple-band operating bandwidths are obtained. The simulated results has a 10 dB impedance bandwidth ranging from 2.3 GHz to 2.8 GHz and from 3.1 GHz to 3.16 GHz and 4.71GHz to 6.4 GHz such that total bandwidth of the proposed antenna is 2.3 GHz with respect to the central frequency. In the first band, the resonant peak of return loss is -20 dBi at 2.5 GHz frequency, in the second band the resonant peak of return loss is -22.4 dBi at 3.12 GHz frequency and in the third band the resonant peak of return loss is -22.3 dBi at 6.32 GHz frequency. Obviously, the proposed antenna has very broader bandwidth which covers the required bandwidths of the IEEE 802.11 WLAN standards in the bands at 2.4 GHz and 5.2 GHz (5150–5350 MHz) /5.8 GHz (5725–5825 MHz) and WiMAX standards in the bands at 2.5 GHz (2.5-2.69 GHz) and 5.5 GHz (5.250-5.850 GHz). The simulated gain of the proposed antenna is shown in Figure 3. The antenna has a maximum gain of about 5.73 dBi at 5.58 GHz frequency with small gain variations in the operating bandwidth. Simulation studies indicate that the maximum antenna radiation efficiency is approximately 85%.

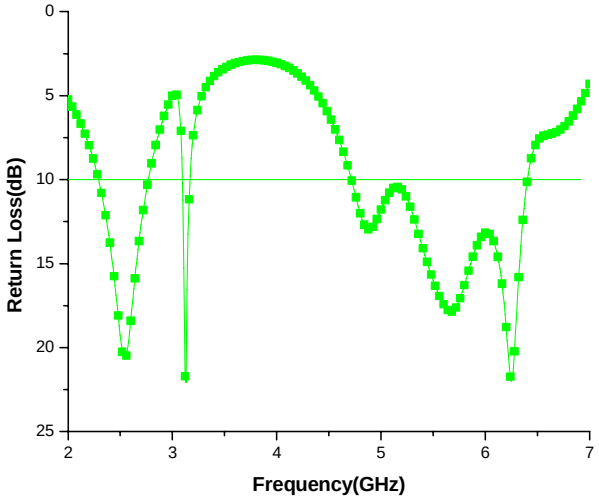


Figure 2 Simulated return-loss of proposed planar antenna

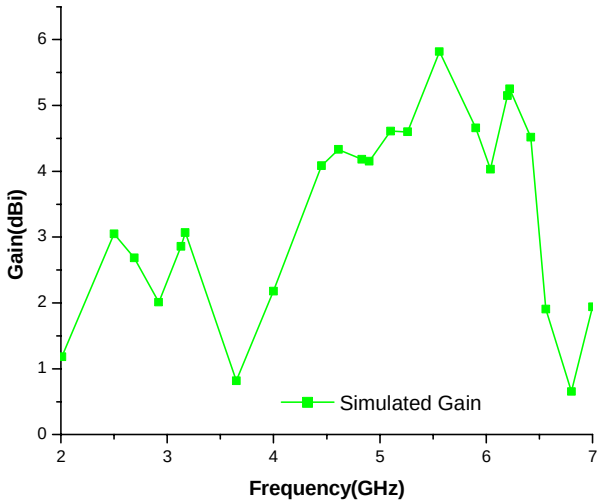


Figure 3 Simulated Gain of the proposed antenna

A parametric study is investigated and it demonstrates that the following parameters influence on the performance of the proposed antenna in terms of bandwidth. The parametric study is carried out by simulating the antenna with one geometry parameter slightly changed from the reference design while all the other parameters are fixed. Figure 4 shows the simulated return loss of the proposed antenna as a function of frequency for different shapes. It is observed from the simulation results study that by using only rectangular shape and rectangular shape with T & I shape slots embedded in the patch, triple band is merged into a single band with huge decrease in the bandwidth as compared to the optimum value of the bandwidth. If we add half E-shape with addition to T & I shape slots then results are very much closed to optimized bandwidth of the proposed antenna with small decrease in the bandwidth.

The current distribution pattern showing that how much of the current is flowing in the proposed structure.The maximum current flowing in the proposed structure is 9.0884

A/m. We can see the average current distribution on the surface of the antenna. It is observed that the current is almost maximum at the lower edge of the patch, along the different slots used in the patch and along the feed line.

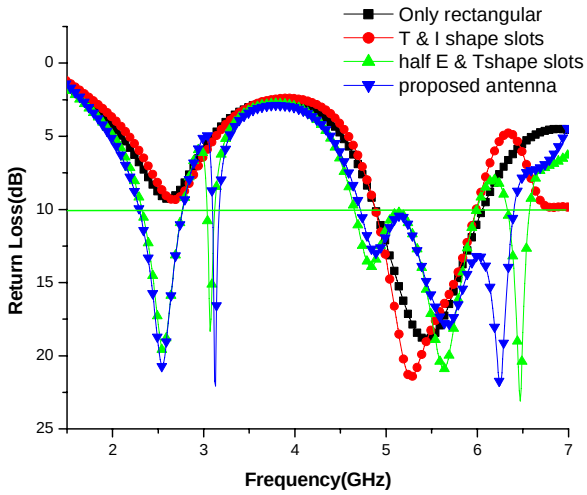
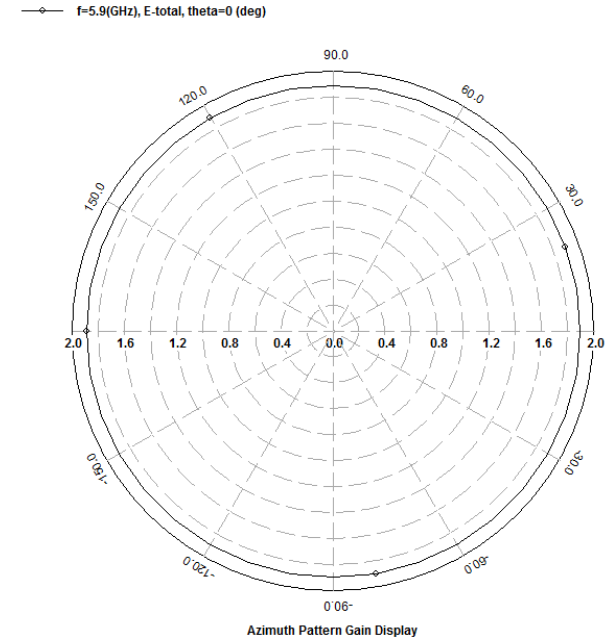


Figure 4 Effect of L_1 & U shape on the proposed antenna

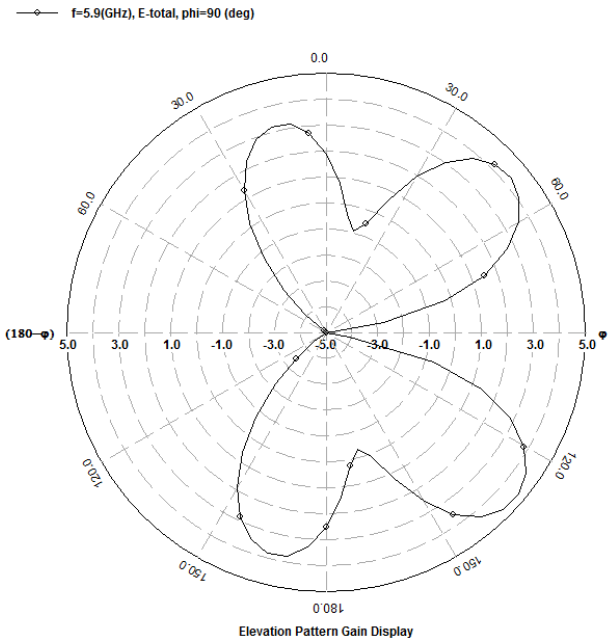
Figure 5 shows the simulated radiation patterns of the proposed dualband CPW-fed monopole antenna at frequency 5.9 GHz. The simulated radiation patterns cut in the Azimuthal (x-y) plane and cut in the elevation plane (y-z) for the proposed antenna is presented in the figure. Similar to monopole kind of antenna, the radiation pattern obtained in the x-y plane are similar to omni-directional. At the above mentioned frequency, nearly figure of eight radiation pattern is obtained in the y-z plane.

Conclusions

A new printed antenna have been designed & simulated to achieve the triple-band operation for wireless communication system. The simulated results show that by using proposed designs and tuning their dimensions, operating bandwidth, measured gain and radiation patterns can be obtained for WLAN/WiMAX applications. The simulated results has a 10 dB impedance bandwidth ranging from 2.3 GHz to 2.8 GHz and from 3.1 GHz to 3.16 GHz and 4.71GHz to 6.4 GHz such that total bandwidth of the proposed antenna is 2.3 GHz with respect to the central frequency. The parametric study shows the significant effects on the impedance bandwidth of the proposed antenna. Besides, its triple-band characteristics, the proposed antenna remains compact with a small volumetric size.



(a) Azimuthal Plane



(b) Elevation Plane

Figure 5 Simulated radiation patterns at 5.9 GHz for proposed antenna

References

[1] Davinder Parkash, Rajesh Khanna, "Design and Development of CPW-Fed Microstrip Antenna for WLAN/WiMAX Applications," International Journal of Progress in Electromagnetics Research C, vol. 17, pp. 17-27, 2010.

[2] D. Sanchez-Hernandez and I. D. Robertson, "Analysis and design of a dualband circularly polarized microstrip patch antenna, " IEEE Transactions on Antennas and Propagation, vol. 43, pp. 201-205,1995.

[3] C. A. Balanies, 'Antenna Theory: Analysis & Design,' John Willey & Sons, Inc., 1997.
[4] Manish Kumar, Manish Kumar Sinha, L. K. Bandyopadhyay, Sudhir Kumar, "Design of a Wideband Reduced Size Microstrip Antenna in VHF/ Lower UHF Range," URSI Proceedings, New Delhi, October,2005.
[5] Wentworth M. Stuart (2005), ''Fundamentals of Electromagnetics with Engineering Applications'', pp 442-445, John Wiley & Sons, NJ, USA.
[6] R. Garg, P. Bhartia, I.Bahl and A. Ittipiboon, "Microstrip Antenna Design Handbook," 1st Edition, Artech House Publisher, Norwood, 2001.
[7] W. S. Chen, C. K. Wu, and K. L. Wong, "Novel Compact Circularly Polarized Square Microstrip Antenna," IEEE Transaction Antennas Propagation, vol. 49, Mar. 2001,pp. 340-342.
[8] F Yang and Y. Rahmat-Samii, "A compact Dual Band circularly Polarized Antenna Design for Mars Rover Mission," IEEE Microwave Wireless Component Letters, vol.-1, June, 2003, pp.858-861.
[9] F. Yang and Y. Rahmat-Samii, "Patch Antenna with Switchable Slots (PASS): Reconfigurable Design for Wireless Communications," IEEE Microwave Wireless Component Letters, vol. 1, June, 2002, pp.462-465.
[10] S. U. Xiang-Fei Peng ,Shun-Shi Zhong; Sai-Qing Xu; Qiang Wu, "Compact dual-band GPS microstrip antenna," Microwave and Optical Technology Letters, vol. 44, pp. 58-61, 2005.
[11] R. B. Waterhouse, S. D. Targonski, and D. M. Kokotoff, "Design and Performance of Small Printed Antennas," IEEE Trans on Antenna and Propagation, Vol. 46, Issue-11, Nov. 1998, pp. 1629 -1633.
[12] Yen-Liang Kuo, Kin-Lu Wong, "Printed double-T monopole antenna for 2.4/5.2 GHz dual-band WLAN operations," IEEE Trans. Antennas Propagation, vol. 51, pp. 2187–2192, September, 2003.

Automatic Localization of Backward Collision of Vehicles Using a Camera

Anitha P., Gajesh K.R., Pruthviraj J., Santosh Kumar S., Nandini. C. and Bhaskara Rao

Department of Computer Science and Engineering,
Dayananda Sagar College of Engineering, Bangalore-78, India.
E-mail: anithacse88@gmail.com, gajeshramachandra@gmail.com, pruviraj@gmail.com,
santosh.subhramani@gmail.com, laasyanandini@yahoo.com, bhaskararao.nadahalli@gmail.com

Abstract
Lots of rear end collisions due to driver inattention have been identified as a major automotive safety issue. A short advance warning can reduce the number and severity of the rear end collisions [4]. This paper describes how to avoid rear end collision when vehicles are moving in the reverse direction. In order to avoid many of the parking lot accidents, our system provides a Backward Collision Warning (BCW) - a new vehicle/obstacle detection method by calculating the area of the obstacle and comparing it to the stored threshold value. The images of the obstacles/objects are captured by monochrome vision camera when a car is moving in the reverse direction. The BCW system uses Digital Image Processing techniques to track the object at the rear end of the vehicle [2], Camera calibration is used to get the distance of the obstacle at the rear end. Kalman filters are used for tracking the obstacles. Secondly bounding box is used to bind(separate each objects by using a rectangular border based on the threshold) the objects. Region properties are used to estimate the area. After determining the area of the obstacle/object, TTC (Time to Collision) is calculated which triggers an alarm system that makes the driver attentive.

The proposed technique is tested on our own generated data sets on parking lot and busy roads. This methodology is found to be efficient and we are planning to test our proposed implementation on road for Real Time application.

Keywords: Backward Collision Warning (BCW), Rear End Collision, Time to Collision (TTC), Camera Calibration.

Introduction

Poor visibility and lack of attention is identified as one of the major reasons for rear end collisions in parking lots and busy roads. Hence the use of Backward Collision Warning (BCW) systems is of great importance. This provides one of the best solutions to avoid such accidents. Nowadays sophisticated cars contain radars and sensors to detect the objects in the rear. Many car manufacturers use the radar technology for finding the accurate range of the objects; however using a lot of radar and sensors is not an economical approach. The radars used do not exhibit better performance due to narrow field of view and poor lateral resolution. This has prevented such systems from entering into today's frequently changing market and demands. Fusion of the radar and vision technology is an attractive approach which can provide better performance, but this approach is costly.

Many of the research done on backward collision detection are not found to be high efficient in detecting and tracking the obstacles, henceforth lot of work can still be done in this area to provide the robust technique in overcoming the backward collision. Hence, we are moving in this direction to come out with a novel methodology which is cost effective and is of high efficiency.

Remarkable development in the field of both video and image processing has inspired many of the experts to work in these fields and to find many novel methods which make life easier and even save them.

Literature Survey

Jianzhu Cui et.al [5] has proposed a technique **"Vehicle Localization using a Single Camera" [2006]** to reduce the number and severity of the rear end collisions. This paper describes a Forward Collision Warning (FCW) system based on monocular vision, and presents a new vehicle detection method: appearance-based hypothesis generation, template tracking-based hypothesis verification which can remove false positive detections and automatic image matting for detection refinement. The FCW system uses time to collision (TTC)[1]to trigger the warning .In order to compute time to collision (TTC), firstly, haar and adaboost algorithm is utilized to detect the vehicle; Secondly, we use simplified Lucas-Kanade algorithm and virtual edge to remove false positive detection and use automatic image matting to do detection refinement; Thirdly, hierarchical tracking system is introduced for vehicle tracking; Camera calibration is utilized to get the headway distance and TTC at last. The use of a single low cost camera results in an affordable system which is simple to install. The FCW system has been tested in outdoor environment, showing robust and accurate performance.

"Mobileye - Advance Warning System" [2006] [1]: Mobileye AWS is an Advanced Warning System that can detect immediate forward collision danger and unintentional lane departure. With this functionality the system provides a timely warning for the most common causes of accidents in nowadays traffic. The 'smart' Mobileye monocular camera[1] analyses the upcoming road while driving the vehicle and registers objects like other vehicles and lane markings. Objects that form a potential danger are transmitted to the warning

panel located inside the vehicle compartment, and warns the driver in advance for the upcoming danger.

Xuezhi Wen.et,.al.[2006] et.al, [8] has proposed a system called **"A Rear-Vehicle Detection System for Static Images Based on Monocular Vision" :**A rear-vehicle detection system of static images based on monocular vision is presented. It does not need the road boundary and lane information. Firstly, it segments the region of interest (ROI) by using the shadow underneath the vehicle and edges. Secondly, it accurately localizes the ROI by vehicle features such as symmetry, edges and the shadow underneath the vehicle, etc. Finally, it completes vehicle detection by combining knowledge-based and statistics-based methods [8]. Under various illuminations and different roads (different day time, different scenes: highway, urban common road, urban narrow road), the system shows good results of recognition and performance.

R. Okada et. al, [2003] [9] has proposed paper on **"Obstacle detection using projective invariant and vanishing lines":** A method for detecting vehicles as obstacles in various road scenes using a single onboard camera. Vehicles are detected by testing whether the motion of a set of three horizontal line segments, which are always on the vehicles, satisfies the motion constraint of the ground plane or that of the surface plane of the vehicles [9]. The motion constraint of each plane is derived from the projective invariant combined with the vanishing line of the plane that is a prior knowledge of road scenes. The proposed method is implemented into a newly developed onboard LSI. Experimental results for real road scenes under various conditions show the effectiveness of the proposed method.

Carlo Tomasi Takeo Kanade [1991] [11] has proposed a White paper on **"Detection and Tracking of Point Features":** The factorization method described in this series of reports requires an algorithm to track the motion of features in an image stream. Given the small inter-frame displacement made possible by the factorization approach, the best tracking method turns out to be the one proposed by Lucas and Kanade in 1981. The method defines the measure of match between fixed-size feature windows in the past and current frame as the sum of squared intensity differences over the windows.

The displacement is then defined as the one that minimizes this sum. For small motions, a linearization of the image intensities leads to a Newton-Raphson style minimization. In this report, after rederiving the method in a physically intuitive way, we answer the crucial question of how to choose the feature windows that are best suited for tracking.

Our selection criterion is based directly on the definition of the tracking algorithm, and expresses how well a feature can be tracked. As a result, the criterion is optimal by construction. We show by experiment that the performance of both the selection and the tracking algorithm are adequate for our factorization method, and we address the issue of how to detect occlusions. In the conclusion, we point out specific open questions for future research.

Peter Hillman et.al [12], Square Eyes Software has proposed a technique to calibrate the camera which is called **"Camera Calibration and Stereo Vision"** This white paper outlines a process for camera calibration: computing the mapping between points in the real world and where they arrive in the image. This allows graphics to be rendered into an image in the correct position. Given this information for a pair of stereo cameras, it is possible to reverse the process to compute the 3D position of a feature given its position in each image - one of the most important tasks in machine vision. The system presented here requires the capture of a calibration chart with known geometry.

LiangLi et.al [3] have proposed a paper on **License Plate Detection Method Using Vertical Boundary Pairs and Geometric Relationship**. License plate detection and recognition is a crucial and difficult issue for an ITS (Intelligent transportation System). This paper proposes a robust license plate detection method using vertical boundary pairs and geometric relationships. A robust and efficient approach to detecting license plates. The main disadvantage is that it is not satisfactory for some specific images due to bad illumination and practical situations.

Proposed Methodology

In this paper we propose Backward Collision Warning (BCW) system based on monocular vision. First we compute the background of the set of images. Based on these results the obstacle/object present at the rear is detected and tracked using the kalman filter [2], a very efficient way to detect objects in video of low resolution. Then the object detected by the kalman filter is fed into the image processing block for drawing a bounding box around the object. Once the bounding box is applied to the object, its area is determined using Regionprops. Based on the area a threshold distance is set which causes an alarm to trigger when the motion of the car is backwards and nearing the object/obstacle at the rear.

Backward Collision Warning (BCW) System is a novel method to detect and avoid backward collisions. It consists of the following steps as shown in Fig.1:

The system is mainly divided into four parts:

- Camera Calibration
- Object detection and Refinement.
- Detecting the Bounding Box with maximum area
- TTC.

Camera calibration

Camera calibration [9] is a crucial phase in most vision systems. We use camera calibration for computing TTC.

The equations used to derive the camera calibration are as follows:

C_h=PCRGw_h [9] represents a perspective transformation [10] involving two co-ordinate systems. We obtain the Cartesian coordinates (x, y) of the imaged point by the first and second components of c_h by the fourth. Converting the above equation to Cartesian coordinates we get:

$$x = \lambda \frac{(X-X0)cos\theta + (Y-Y0)sin\theta - r1}{-(X-X0)sin\theta sin\alpha + (Y-Y0)cos\theta sin\alpha - (Z-Z0)cos\alpha + r3 + \lambda} \quad (1)$$

$$y = \lambda \frac{-(X-X0)sin\theta cos\alpha + (Y-Y0)cos\theta cos\alpha + (Z-Z0)sin\alpha - r2}{-(X-X0)sin\theta sin\alpha + (Y-Y0)cos\theta sin\alpha - (Z-Z0)cos\alpha + r3 + \lambda} \quad (2)$$

With reference to A=PCRG. The elements of A contain all the camera parameters and c_h=Aw_h [9]. Letting k=1 in the homogenous representation yields

$$\begin{pmatrix} c_{h1} \\ c_{h2} \\ c_{h3} \\ c_{h4} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \quad (3)$$

Substituting $c_{h1}=xc_{h4}$ and $c_{h2}=yc_{h4}$ in the above equation and the product yields:

$$\left.\begin{aligned} xc_{h4} &= a_{11}X+a_{12}Y+a_{13}Z+a_{14} \\ yc_{h4} &= a_{21}X+a_{22}Y+a_{23}Z+a_{24} \\ c_{h4} &= a_{41}X+a_{42}Y+a_{43}Z+a_{44} \end{aligned}\right\} \quad (4)$$

c_{h3} was ignored because it is related to z.

$$\left.\begin{aligned} &a_{11}X+a_{12}Y+a_{13}Z-a_{41}xX-a_{42}xY-a_{43}xZ-a_{44}x+a_{14}=0 \\ &a_{21}X+a_{22}Y+a_{23}Z-a_{41}yX-a_{42}yY-a_{43}yZ-a_{44}y+a_{24}=0 \end{aligned}\right\} \quad (5)$$

The calibration procedure then consists of the following steps

Step (a): obtaining m>=6 world points with known coordinates(X_i, Y_i, Z_i),i=1,2,…,m.

Step (b): imaging these points with the camera in a given position to obtain the corresponding image points (x_i,y_i),i=1,2,…,m.

Step (3): using these results in equations (5) we solve for the unknown coefficients.

We have successfully derived the camera calibration with accurate range values based on real time data samples that are collected using the camera which is suitable for system.

We have also taken some of the sample input data and worked towards obstacle tracking and detection. We have tested the proposed approach on many samples of data sets generated.

Kalman filter

Object Detection is done using the Kalman filter [2] which can be defined as the following:

Data fusion using a Kalman filter can assist computers to track objects in videos with low latency. The tracking of objects is a dynamic problem, using data from sensor and camera images that always suffer from noise. This can sometimes be reduced by using higher quality cameras and sensors but can never be eliminated, so it is often desirable to use a noise reduction method.

The iterative predictor-corrector nature of the Kalman filter can be helpful, because at each time instance only one constraint on the state variable need be considered. This process is repeated, considering a different constraint at every time instance. All the measured data are accumulated over time and help in predicting the state.

Video can also be pre-processed, perhaps using a segmentation technique, to reduce the computation and hence latency.

The Kalman filter is a recursive estimator. This means that only the estimated state from the previous time step and the current measurement are needed to compute the estimate for the current state.

In this paper only the predictor part of the kalman filter is used for object detection, no correction is done to the predicted objects since the need is only to locate the object and not to track it.

Bounding box

Component labeling [9] gives each blob on the foreground image map a unique label, a list is produced with all of the labels in the binary image map and the corresponding number of pixels which have this label. The list is sorted in descending order of blob size. A bundle rectangle is drawn for all the detected objects and the largest rectangle amongst them is taken to be the object which is nearer to the vehicle.

Regionprops is used to measure the properties of image regions (blob analysis) which measures a set of properties for each labeled region L.

Time to Collision(TTC)

In our system, TTC [4] is used to trigger a warning to the driver regarding the object/obstacle at the rear end.

Erez Dagan [1] from Mobileye Company gave us the way to compute the TTC.

TTC is calculated using equation (7).

First we need to measure the rate of 'optic inflation'.

If we define "optic inflation" to be "ds"

$$ds = dw/w \quad (6)$$

('w' is the width of the target vehicle in the image, and 'dw' the difference of widths of that vehicle between 2 frames) than it is shown that

$$TTC = dt/ds \quad (7)$$

where, dt is the time gap between the 2 frames.

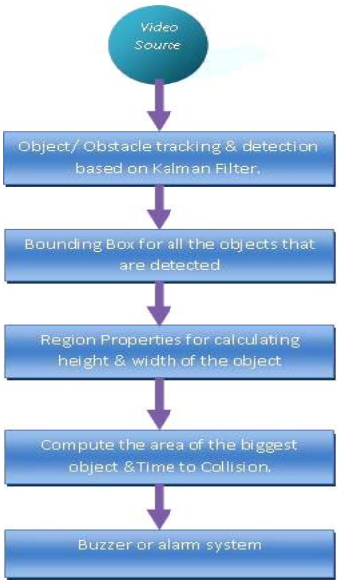


Fig.1: Steps followed in Backward Collision Warning (BCW) systems

As shown in fig.1 the Video captured is fed into the system from the camera at the rear end of the vehicle into the image processing System.

The System first slices the videos into multiple frames and then processes each frame accordingly. Firstly the camera calibration [9] is done to find the distance of the object from the vehicle. This distance will be used at the later stage to compute TTC [4] (Time to Collision). Based on the distance and the TTC an alarm is triggered to alert the driver about the obstacle at the rear end.

Secondly it uses Kalman filters [2] for tracking the object/obstacle which is at the rear end of the vehicle.

Bounding Box [7] is used to draw a rectangle around the objects detected in each frame.

The region props of the bounding box is used to find the area of the object, based on its height and width.

Sorting the areas in descending order of all the objects detected, will give the area of the biggest object in the top of the array. This is used to draw a bounding box with a red color. This is the biggest object in that particular frame. The rest of the objects are drawn with a bounding box of green color.

The buzzer or the alarm system alerts the driver when the car moves towards the object at the rear end. At a specific point using the area of the biggest object the system determines the threshold distance and alarms the driver when the car has reached the threshold distance. When the car reaches this distance from the obstacle, the alarm is triggered.

Experimental Results

In our system the threshold is taken to be 0.6 mtrs. The snapshots obtained during the experimentation for various conditions are shown in the figures below.

Images without Region of Interest (ROI)

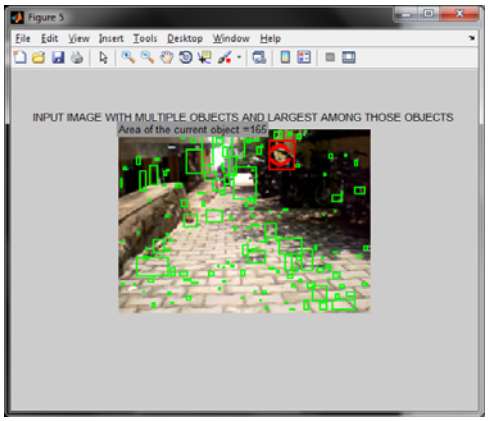


Fig2: Snapshot showing no Object or Region of Interest.

In fig.2 the vehicle is moving in the reverse direction and there are no objects or obstacles at the rear end. The red colored bounding box is the first biggest object that is detected, the other green bounding boxes are the objects that are small in area and are negligible.

Images where Region of interest are static

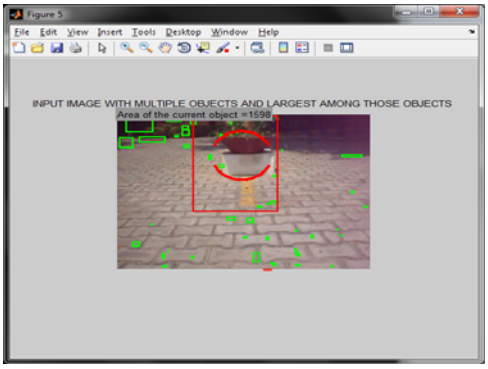


Fig3: sample largest bounding box extracted using region properties.

In the fig.3, the object or the obstacle is static in the ground. Hence the calculation of the area is a straight forward way. It is simpler to calculate TTC for such static objects.

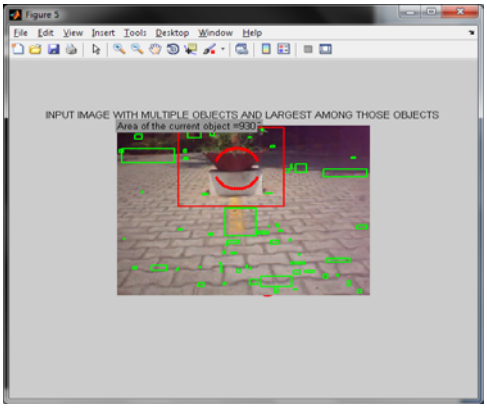


Fig 4: snapshot showing Vehicle approaching towards the object

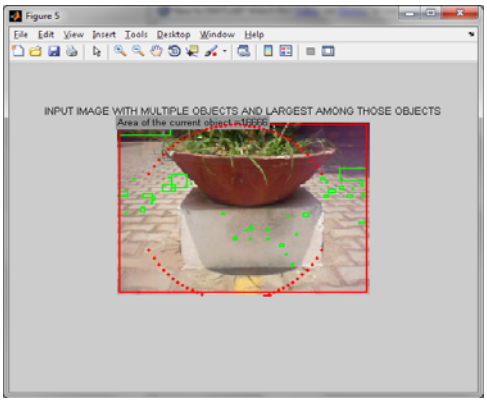


Fig 5: Snapshot showing Object is nearest to the car.

In fig.4 the vehicle is nearing the object in the rear and in fig.5 the car is too close to the object, at this point the threshold is crossed and the alarm buzzes.

Images where Region of interest are dynamic

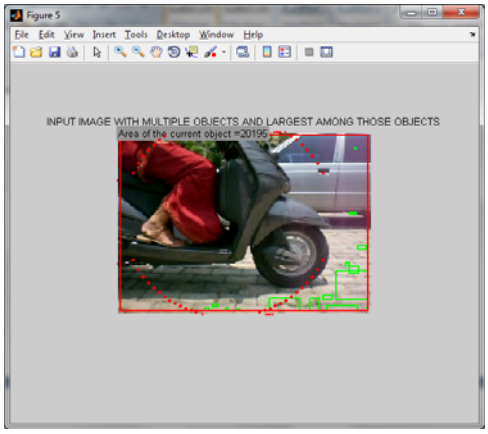


Fig.6: Snapshot showing the dynamic object moving in the rear end of the vehicle.

In test case fig.6, the dynamic object which is in motion is also detected.

Various kinds of datasets have been taken for checking the performance of the system in parking lots and even busy roads. The system works efficiently for various datasets and the results are obtained with accurate values.

.

Conclusion

The backward collision warning system can be implemented in all types of vehicles because of its cost effectiveness and affordable accessories which we collected the information in consultation with the automobile companies. Since most of the rear end driver related accidents is caused due to lack of attention by the driver[1], using this type of automatic obstacle detection system will not only assist the driver to safely move the car in the reverse direction, but also avoid the rear end collision.

The system can be implemented in all types of vehicles, both high cost and low cost and is found to be economically benefitted to the society. Since usage of high end radar systems are not affordable, our system not only helps in avoiding rear end collision but also in a cost effective manner, since the accessories used in the system are of low cost compared to that used in the existing backward collision warning systems (Eg. Range finders and optical sensors).

Henceforth a lot of work can still be done in this area to provide the robust technique in overcoming the backward collision. The methodology can be extended to take care of the orientation of the obstacles which are not in the perception of the camera.

References

[1] Mobil eye(Advance Warning System): A novel method of finding front end collision of vehicles in highways. http://www.mobileye.info/en/index.html

[2] The general descriptio about Kalman filter and its various types according to wikipedia site:http://en.wikipedia.org/wiki/Kalman_filter

[3] Meaning of collision detection according to wikipedia site:http://en.wikipedia.org/wiki/Collision_detection

[4] Jianzhu Cui,Fuqiang liu,Zhipeng Li,Zhen Jia, “Vehicle Localisation Using a Single Camera” 2010 IEEE Intelligent Vehicles Symposium, University of California, San Diego, CA, USA June 21-24, 2010

[5] E. Dagan, O. Mano, G.P. Stein and A. Shashua,”Forward Collision Warning with a Single Camera”, IEEE Intelligent Vehicles Symposium (IV2004) June 2004, Parma, Italy

[6] Christopher Rasmussen, ”Grouping Dominant Orientation for Ill- Structured Road Following,” Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol.1, pp. I470-I477, 2004.

[7] Xuezhi Wen, Hong Zhao, Nan Wang, Huai Yuan, “A Rear-Vehicle Detection System for Static Images Based on Monocular Vision”, School of Information Science & Engineering, Northeastern University Advanced Automotive Electronic Technology Research Center, Neusoft Co., Ltd. Shenyang, China, 2006 IEEE.

[8] R. Okada et al,”Obstacle detection using projective invariant and vanishing lines”, Proceedings of the 9th ICCV 2003

[9] Digital image processing- Rafael C Gonzalez, Richard E. Woods ; ISBN: 0-201-43577-2.

[10] White Paper: Camera Calibration and Stereo Vision: Peter Hillman, Square Eyes Software 15/8 Lochrin Terrace, Edinburgh E-MAIL: pedro@peterhillman.org.uk, www.peterhillman.org.uk October 27, 2005

Design and Development of Low Cost and Light Weight Microwave Filters by Using Metalized ABS Plastic as a substitute of Metalized Substrate and Metals

Jagdish Shivhare[1] and S.B. Jain[2]

[1]Department of Electronics and Communication, ITM University, Gurgaon, Sector-23A, Gurgaon-122 017, India
[2]Department of Electronics and Communication, Indira Gandhi Institute of Technology,
Indraprashth University Campus, Delhi, India
E-mail: jshivhare.isro@gmail.com

Abstract
The main objective to introduce the ABS (Acryonitrile Butadyne Styrene) plastic substrate in place of RT- Duroid substrate for microstrip filters is to reduce the cost. The cost of plated ABS plastic substrate is substantially less (Rs. 3/-sq.inch) compared to the cost of RT-Duroid (Rs.100/-sq.inch). Two microstrip (hairpin line) band pass filters at 1537.5 MHZ and 1575.42 MHz have been developed and tested. The performance of filters have been verified over temperature range, -10 deg. C to 60 deg. C.The ABS plastic blocks have been used in place of metal blocks of brass, copper and Commercial Aluminium to fabricate three dimensional microwave cavity filters for communication systems.The specific gravity of ABS plastic is 1.03 gms/cubic cm compare to 2.7, 6.3 and 9.6 gms/cubic cm of Com. Al, Copper and Brass respectively. Hence the weight of metallized ABS cavity filters will be 1/3rd, 1/6th and 1/9th of the com. Al, copper and brass cavity filters.

Index Terms: Acryonitrile butadiene styrene, dissipation factor, insertion loss, hairpin line, microstrip, poly tetra flouroehtelene, RT-duroid.

Introduction

With the development of the electronic industry in India, there is continuous need for new materials having better performance properties with stringent size and weight limitation. Such needs can be tailor made by plastic materials, which have very high rate of environmentally safe processablility and production process, maintaining the accuracy and level of functional properties. As new plating processes groom, metalized plastics play a bigger role as replacement of conventional materials on cost, weight and performance basis. Currently plated thermoplastic ABS is widely used in automotive and decorative uses but is untouched for passive microwave component applications in communication.

Choice of Material

At microwave frequencies the current density is maximum near the surface and it falls off exponentially with depth. Thus as long as there is a thin layer of silver, copper or any other highly conductive material on surface, the body can be of wood, plastic or anything else; without affecting the microwave propagation. In addition to size and weight the other factors determining the type of body material are machining tolerances, water absorption, ability to handle power, efficiency, problems in matching, shielding, and reliability.

Acrylonitrile Buadiene Styrene (ABS) plastic has excellent toughness, rigidity and gloss[1][2]. They are cheaper than engineering plastics like nylons, polyacetals and polycarbonates. The important physical and electrical properiates like tensile modulus, tensile strength, surface hardness, porosity, coefficient of thermal expension, thermal conductivity, heat distortion temperature, dielectric strength, dissipation factor, surface resistivity, etc. were found to be favorable for such applications. Moulding from ABS exhibit better and uniform impact strength in all the directions. Another important feature of ABS is that it is the only plastic which can be commercially electroplated. It is found that electroplated ABS parts show improvement in properties like surface hardness, tensile and flexural strength, heat resistance, chemical resistance, weather resistance etc. The electrical properties of ABS plastics are unaffected by temperature and humidity. Dielectric strength, power factor, and dielectric constant are reasonably good to allow it to be used in electronic or electrical applications like coil formers, connectors, wave guides etc[3].Thus for our study, ABS (platable) plastic material was selected for the design, development and fabrication of substrate type (instead of RT-Duroid) and cavity type (instead of metal) band pass filters in microwave frequency range for communication systems.

Coating/Plating Procedure

The copper thickness was built by electroplating. Silver plating is done after deposition of copper on plastic. The deposits were subjected to environmental tests such as humidity and corrosion resistance, thermal cycling, thermo vacuum, baking, hot and cold storage etc. tape test was used to check the adhesion of the coating after each test. Mechanical properties of the coatings were evaluated by micro hardness test, surface roughness and peel strength. To verify the

assumption of replacement of very costly substrates and metals with plated ABS plastic, some planar structured(hairpin line) and cavity type band pass filters (coaxial, helical and comb line) having different centre frequencies and bandwidths were designed and fabricated. The standard design tools and techniques were used for all types of cavity band pass filters but for hairpin line filters necessary design corrections were applied [6][7].

ABS with 10% butadiene is more suitable for electroplating than ABS with 16 to 27% butadiene. Several trials were conducted for electroplating on ABS plastic. The articles were immersed in the mixture of chromic acid and sulphuric acid to improve mechanical adhesion. Poor etching leads to skip plating or poor adhesion of the plating and possible blistering. Thus etched articles are not to be treated with sensitizer and activators stannous chloride and palladium chloride solution are used for this purpose. The deposited palladium nuclei on the plastic surface, initiates elecrtoless plating of copper or nickel or gold or other metals. We have carried out electroless copper deposition for our work. Finally deposited with electroplated copper or silver to get highly conductive surface.

Performance Evaluation

The transmission characteristics of metalized ABS plastic filters in the form of substarte and cavity were tested to compare with RT-Duroid and commercial Aluminum filters by using the network analyzer. The electrical parameters of metalized ABS filters; like centre frequency, insertion loss at the centre frequency, 3dB bandwidth, stop band attenuation were measured and compared with that made from RT-Duroid and Aluminum alloy in the temperature range from -20 deg C to -60deg C, as shown in the comparison tables. Thus this new plastic material is getting use in making different types of band pass filters for ground and space application [3].

Though, the dielectric constant and dissipation factor above 1 MHZ are not given in the literature, we have measured effective dielectric constant (Єeff) up to 10 GHz and verified practically by the performance of two hairpin line band pass filters at 1537.5 and 1575.42 MHz. The insertion loss is more due to higher dissipation factor. So if the insertion loss is not very critical, the very low cost microstrip hairpin line filters can be developed by using plated plastic substrate in place of RT-Duroid (#5880, Єr=2.22).The measurement method of Єeff also has been verified by cross checking the value of Єeff for RT-Duroid (#5880), for 50 ohms line up to 10 GHZ (table-1) on network analyzer HP-8510[8][15].

Design Procedure

The design procedure available for RT-Duroid #5880 has been applied to calculate the dimensions of resonators for hairpin line filters. The existing design tables and graphs are sufficient to carry out the design calculations.

Experiment Procedure

Two hairpin line microstrip band pass filters have been optimized on Network Analyzers HP8754 A and HP8510. The resonator lengths, practically found at 1537.5MHZ and1575.42MH have been verified the correctness of measured Єeff of ABS plastic substrate[5]. The correctness of measurement method is also verified with the help of value of Єeff, measured and available in the data sheet of # 5880, RT-Duroid, supplied by Rogers Corp. USA.We have measured Єeff for ABS plastic and RT-Duroid, having electrical lengths of 150 and 50 mm of 50 ohms microstrip line of each substrate material (table-1).In our experiments we used the microstrip filters fabricated on ABS plastic substrate by using positive/negative of RT-Duroid based filters. The centers of response were achieved at lower frequencies than that of RT-Duroid. The filters were optimized by trimming-out the resonator lengths. The band widths with respect to center frequencies were not similar to RT-Duroid based filters in both cases.

So, if the insertion loss and band width are not critical, low cost microstrip filters can be developed by using ABS-plastic substrate as an alternative to RT-Duroid substrate[9][10].

Conclusion

It can be concluded that metallised ABS plastic at UHF and SHF exhibits electrical behavior similar to that of metals. Although additional work is required before large scale use of ABS can be implemented by industry, the superior performance will undoubtedly make it the material of choice for future high performance microwave equipments in satellite earth stations.

Table 1: properties comparison (as per data sheet & catalogues.

Sr. No.	PROPERTY	RT – DUROID # 5880	ABS – PLASTIC # AP78EP
01.	DIELECTRI CONSTANT Er (relative)	(2.22 - 0.02) UPTO 10 GHZ (Eeff =1.89 for 50 line)	Er = (2.8 – 3.8) AT 1 MHz Eeff = (1.89 to 2.12) UPTO – 10 GHz measured on HP 8510 TABLE 1
02.	Dissipations factor (ten)	(0.0009 to 0.0010) UPTO 10 GHz	0.0024 at 9.2 GHz (measured)
03.	Specific gravity	2.20 gms/gm^3	1.05 gms / cm^3
04.	Heat distortion temperature	> 260^0 C	84^0 C
05.	Power handling capability	120 W	72 W
06.	Thermal Expansion (Linear)	48 x 10^{-6} mm/mm/^{0}C	70 x 10^{-6} mm/mm/^{0} C
07.	Tensile strength	450 Kg /gm^2	430 Kg /gm^2
08.	Volume resistivity	2 x 10^{13} to 2 x10^{14} ohm-cm	10^{13} to 10^{16} ohm/cm
09.	Elongation at Break	17.6 %	25 %
10.	Hardness R- scale	R - 88	R - 110
11.	Deformation under Load	(0.6 -1.0) %	(0.4 0.6) %
12.	TENSILE MODULES	30 X 10^3 Kg/gm^2	23 x10^3 Kg/gm^2

Table 2: Measurement of effective dielectric constant (feff) of ABS on network analyzer.

$$(\epsilon\ eff) = \frac{\text{Difference of electrical length i.e. (EL150-EL50)}}{\text{Difference of physical length i.e. (150-50)}}$$

Frequency of measurement in MHz	RT.DUROID #5880 Measured Electrical length in (mm) for		ϵ eff	ϵ eff	ABS-PLASTIC # AP78 EP Measured Electrical length in (mm) for		ϵ eff	ϵ eff
	150mm physical length of 50 Ohm line	50mm physical length of 50 Ohm line			150mm physical length of 50 Ohm line	50mm physical length of 50 Ohm line		
45.0	230.76	96.26	1.345	1.81	231.22	93.82	1.374	1.89
650.0	232.18	96.18	1.360	1.85	233.15	95.35	1.378	1.90
1500.0	234.49	97.79	1.367	1.87	234.61	95.71	1.389	1.93
2500.0	235.82	97.62	1.382	1.91	235.70	95.40	1.403	1.97
4500.0	236.75	97.15	1.396	1.95	237.14	96.14	1.410	1.99
7000.0	242.11	97.91	1.442	2.08	250.32	108.22	1.421	2.0
10000.0	251.43	106.53	1.449	2.10	263.07	117.47	1.456	2.1
From DATA sheet of REGOERS CORP. Er=2.22 ± 0.02, up to 10GHz. ϵ eff=1.89 for 50 ohm line. Dissipation factor: 0.0009. Measured ϵ eff is approximately equal to actual ϵ eff (1.89). Which verifies correctness or our test method of measurement. (Table-2)					From DATA sheet by ABSTRON Er=(2.8 – 3.3) at 1 MHz. ϵ eff(measure) =1.89-2.12 for 50 ohm line from 45MHz to 10GHz. Dissipation factor: 0.0024 at 9.0GHz. Measured by wave guide method. 19x19x3 mm sheet of ABS plastic.			

Table 3: Verification of correctness of our test method.

S.No.	For RT-DUROID # 5880 T ROGERS CORP. USA	For ABS-PLASTIC # AP78EP ABSTRON INDIA
01.	Thickness of substrate: 1-6mm	Chosen thickness for filters: 1-6mm
02.	As per DATA sheet: ϵ eff=1.89 for 50 ohm line upto 10GHz	Measured ϵ eff=1.89 to 2.12 from 45MHz – 10GHz
03.	Length of resonator ($\frac{\lambda}{4}$) at 1537.5 MHz & 1575.42MHz $\frac{\lambda}{4} = \frac{3\times10^{11}}{4\times1537.5\times10^{6}\sqrt{1.89}}$ =35.628 at 1537.5MHz $\frac{\lambda}{4}$ =34.483 at 1575.42MHz	Practically (found) lengths of hairpin line resonators are 31.0mm and 32.0mm at center frequencies 1537.5 & 1575.42 respectively. Therefore, $\sqrt{eff} = \frac{3\times10^{11}}{4\times1537.5\times10^{6}\times(\frac{\lambda}{4})}$ = 1.96 at 1537.5 MHz and 1.94 at 1575.42MHz
04.	By our test method, ϵ eff =1.87 at 1500MHz and varies form 1.89 to 2.12 for 45MHz to 10GHz. Thus the measured values of ϵ eff are very close to the actual ϵ eff (1.89 for 50 ohm line), which verifies the correctness of our test method.	By the same test method, ϵ eff=1.93 (Table-1) for which is very close to the values found practically, ϵ eff=1.94 & 1.96 at 1537.5MHz & 1575.42MHz. This also provides the proof of the correctness of our method adopted for measurements of ϵ eff.

Table 4: Achieved results at various temperature.

Para meter	Unit	Test temperatures: Room temperature		-10°C		-20°C		+50°C		+65°C	
		Com Al	ABS plastic	Com Al	ABS plastic	Com Al	ABS plastic	Com Al	ABS plastic	Com. Al	ABS plastic
Center Freq.	MHz	1636	1636	1636.5	1636.3	1639	1638	1635.6	1635.7	1632	1634
0.1dB BW	MHz	±10	±10	±9.5	±9.9	±9.2	±9.7	±9.6	±9.9	±9.2	±9.6
3.0dB BW	MHz	±20	±20	±19.5	±19.6	±19.1	±19.5	±20.0	±20.0	±19.3	±19.6

Table 5: design and development of various types of metal cavity bandpass.

Type of filter	Freq. Band MHz	Center Freq. MHz	Band width MHz	Insertion loss dB	I/O return loss dB	Stop band attenuation dBc	Size (LxBxH) MMxMMx MM	Weight grams
VHF/UHF								
Helical	52-55	53.5	±1.5	1.0	16	30dBc @60MHz	200x50x72	80
Helical	85-88	86.5	±1.5	1.0	16	30dBc @80MHz	200x50x72	80
Helical	320-328	324.0	±4.0	6.0	15	>30dBc @±8MHz	105x38x25	41
Helical	591-609	600.0	±9.0	2.2	20	>30dBc @±18MHz	100x23x33	28
Combline	1050-1350	1200.0	±150	1.5	15	>30dBc @±300MHz	155x45x30	55
L-BAND								
Co-axial	1530-1545	1537.5	±7.5	0.4	20	>30dBc @±90MHz	130x44x37	90
Co-axial	1626-1646	1636.0	±10	0.4	20	>30dBc @±90MHz	152x52x41	130
S-BAND								
Combline	2500-2690	2595.0	±85	1.5	16	>30dBc @±2000MHz	120x25x20	110
C-BAND								
Combline	4170-4200	4190.0	±20	1.5	15	>30dBc @±40MHz	107x21x18	30
Combline	4570-4610	4590.0	±20	1.5	15	>30dBc @±40MHz	126x23x17	40
Combline	5850-5930	5890.0	±40	2.0	16	>30dBc @±80MHz	95x18x12	50
Ext. C-BAND								
Combline	6725-7025	6835.0	±150	2.0	15	>30dBc @±500MHz	100x11x13	100

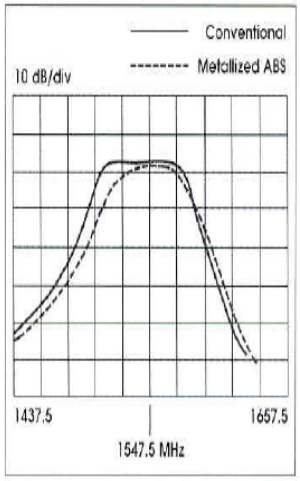


Figure 1 - Measurement of the 1537.5 MHz hairpin filter.

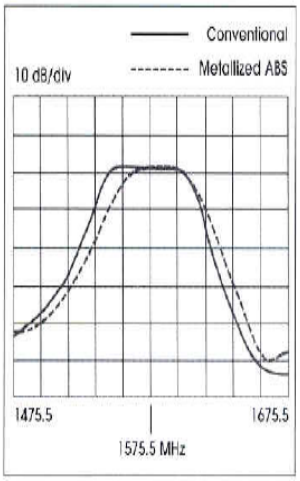


Figure 2 - Measurement of the 1575.42 MHz hairpin filter.

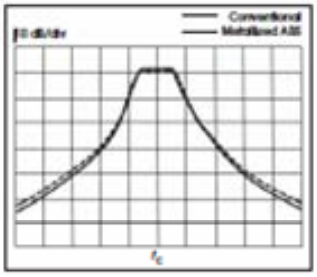


Figure 3 - Bandpass plots for the 1537.5 MHz coaxial cavity filter.

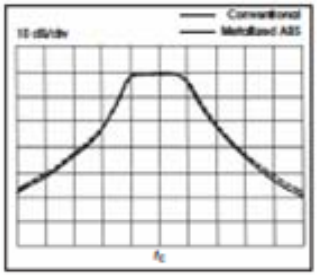


Figure 4 - Bandpass plots for the 1636.0 MHz coaxial cavity filter.

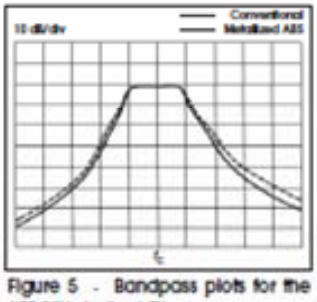


Figure 5 - Bandpass plots for the 600 MHz helical filter.

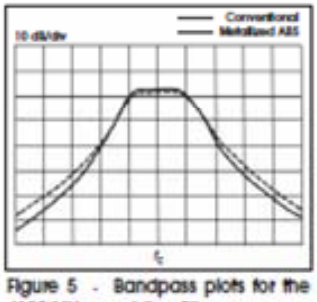


Figure 5 - Bandpass plots for the 4190 MHz combline filter.

References

[1] Data sheet for ABS plastic AP78EP by "ABSOTRON" Technical Collaboration with Sumitomo Chemical Engg. Co., Japan.

[2] High performance plated plastics" by Jim Rychwalski and Martin Bayes, Shipley Co., Newton, MA.

[3] Data sheet for ABS plastic AP78EP by "ABSOTRON" Technical Collaboration with Sumitomo Chemical Engg. Co., Japan.

[4] Zhu, Y.-Z, Y.J. Xie and Feng "Novel microstrip bandpass filters" Progerss in Electronics Research, PIER, 29-41, 2007

[5] Xiao, J.-K., S.-W MA, S.Zhang, and Y Li, "Novel

compact band pass filters"Journal of Electromagnetic Waves and Applications, Vol.21, No.10, 1341-1351, 2007

[6] Deng, P-H., Y. S. Lin, C.-H. Wang and C.H. Chen "Compact Microstrip bandpass filters with good stopband rejection"IEEE Transactions on Microwave Theory and techniques, Vol.54, No.2, 533-539, Feb 2006.

[7] Kazerooni, M. and A Cheldavi, "Simulation, analysis, design and applications of microstrip structure filters using multistrip method" Progerss in Electromagnetics Research" PIER, 63, 193-207, 2006

[8] D. Pozar, *Microwave Engineering,* Third Edition, Wiley, 2005. pp. 416-438.

[9] Jen-TasiKuo, Ming-Jyh Maaand Ping-han Lu, "Microstrip filter with Compact miniaturized hairpin line resonators" IEEE Microwave Theory and Guided Letters, Vol. 10, No.3, March 2005, pp 94-95

[10] H. Wang and L Zhu " Microstrip resonator with ultra-broad rejection bandwidth"Electronic Lett.Vol.40, No.9, pp.1188-1189, September 2004

[11] Hong J. S. and Lancaster M. J. " Microstrip filters for RF/microwave application" A Wiely–Interscience publications Canada, 2004

[12] Agilent Technologies, Inc.; information at www.agilent.com

[13] Sonnet Software; information at: www.sonnetusa.com

[14] Ansoft-HFSS-3D for Electromagnetic modeling: www.ansoft.com

[15] Zereve A. I." Handbook of filter synthesis" Wiley & Sons New York.

A Transliteration Keyboard Configuration with Tamil Unicode Characters

M.A.C.M. Raafi[*] and H. M. Nasir[#]

[*]*Department of Mathematical Sciences, South Eastern University of Sri Lanka*
E-mail: raafim@seu.ac.lk
[#]*Department of Mathematics, University of Peradeniya, Sri Lanka*
E-mail: nasirh@pdn.ac.lk

Abstract
Keyboard configurations for typing are available for many languages and for data processing tasks. The common keyboard used today is QWERTY keyboard. The QWERTY keyboard layout is specially designed for typing English alphabets and numerals. Typing for other languages needs these configurations which remap the QWERTY keys to fit for other languages. This configuration often faces difficulties due to large number of character sets in these languages other than English. To solve this issue, transliteration keyboard configuration is to be considered. Transliteration is a method by which one could read a text of a language in the writing method of another language. In this paper, phonetically we discuss about developing a transliteration keyboard configuration for Tamil language using Unicode encodings.

Introduction
Input devices are used to enter data and commands in the computer system for data processing work. One of the commonly used input devices is the keyboard which consists of letters, numerals and other special characters.

There are different types of keyboard system available in the computing environment. The standard keyboard is known as the QWERTY keyboard. This keyboard is specially designed to type English Language letters and related symbols. Use of other languages, such as Asian languages, the QWERTY keyboard is inconvenient.

Entering these Asian language characters using this QWERTY keyboard is impossible without a proper convenient configuration mapping for the English keys in the keyboard. Even with the configuration mapping, typing the letter of the language is difficult, because one has to memorize or be familiar with the keyboard mapping in the configuration.

Despite these limitations, transliteration is to be considered for typing texts to the benefit of end users. The transliteration is the process by which one reads and pronounces the words and sentences of one language using the letters and special symbols of another language. It is helpful in situations where one does not know the script of a language but knows how to speak and understand the language [1].

For example, one of the Asian languages, Tamil, can be introduced to English literate Tamils and non-Tamils with a transliteration scheme. There are 247 characters in Tamil: 12 vowels, 18 consonant, 216 compound alphabets and one Aayitha character. In Tamil word-processors the large numbers of compound alphabets are obtained by a sequential keying of the corresponding consonant and vowel. For example, the keystrokes for consonant k (க்) followed by vowel I leads to appearance of compound character ki (கி). Keyboard layouts of this kind have been called "phonetic". Tamil transliteration is phonetic keyboard system. Thus, the Tamil word for father (அப்பா) is written as appA (or appaa), mother (அம்மா) as 'ammA' (or as ammaa) in the transliteration program.

The following advantages are available normally in ourv transliteration system.

1. A user-friendly keystrokes; users easily type in more familiar way.
2. No need to memorize whole the mapping key strokes of the keyboard.
3. New person entering from some other language can type easily.
4. We don't need to change the font each time to type following characters special character and symbols such as: / , : < > |) (* & ^ % $ # @ ! ~ + ?.....etc
5. By introducing Unicode
    a. It can be displayed everywhere
    b. No matter about the language
6. No matter about the font
7. Wrong word format is being corrected.

Encoding Systems
Encoding scheme is a necessary part of the configuration of a keyboard layout for the transliteration program. The encoding is the system by which the characters in a set are represented in binary form in a file. In computers and in data transmission between them, i.e. in digital data processing and transfer, data is internally presented as octets, as a rule. Octets are often called bytes, but in principle, octet is a more definite concept than byte. Internally, octets consist of eight bits [6].

Tamil Character encodings
In Tamil, the forms of some of the letters differ from one to another for the same vowel sound. This is the reason for the inclusion of a high number of letters in the Tamil keyboards

designed so far. Tamil is a language, where in addition to the basic vowels (uyir) and consonants (mei), the compounded (uyirmei) characters, all have unique glyph forms. Some popular Tamil font encoding schemes are TSCII, TAM, TAB, ISCII and Unicode.

TSCII

The first and most popular one is the Tamil Standard Code for Information Interchange (TSCII), a glyph-based, 8-bit bilingual encoding. It uses a unique set of glyphs; the usual lower ASCII set. Roman letters with standard punctuation marks occupy the first 128 slots and the Tamil glyphs occupy the upper ASCII segment with slots 128-256.

TAM and TAB

TAM is a Monolingual encoding scheme (TAmil Monolingual) where TAB is a Bilingual encoding scheme (TAmil Bilingual). They were proposed by the Tamil Nadu Government. TAM is limited use in an OS environment.

ISCII

Indian Standard Code for Information Interchange, ISCII is a 8-bit /single byte umbrella standard, defined in such a way that all Indian languages can be treated using one single character encoding scheme. ISCII is a bilingual character encoding (not glyphs-based!) scheme. Roman characters and punctuation marks as defined in the standard lower-ASCII take up the first half the character set (first 128 slots). Characters for Indic languages are allocated to the upper slots (128-255) [5].

Unicode

Unicode is an international standard for multi-lingual word-processing. It is a two-byte encoding scheme which covers the entire world's common writing systems. It represents each character as a 2-byte number, from 0 to 65535. Each 2 byte number represents a unique character used in at least one of the world's languages. There is exactly 1 number per character, and exactly 1 character per number. It provisions over 65000 slots to handle nearly all world more than 50 languages simultaneously. Along with other Asian languages, for example Tamil has been assigned specific slots from U+0B80 to U+0BFF (which, in decimal, is from 2944 to 3071; 128 locations) in this multi-lingual standard [6].

Unicode encodes only basic vowels and consonant characters and a set of modifiers to represent situations where the vowel/consonant pair appear as a combination (uyirmei) in Tamil language. Unicode file stores textual information solely at this "character" level. It does not care about the actual form of the glyphs. Rendering of the glyphs corresponding to stored characters is left to softwares.

Once we get beyond the ASCII world, there are many different native encodings for different languages and operating systems. Converting between all of these is easiest with a central "common point", and that is Unicode.

Technically, Unicode is used wherever the characters used are all drawn from the Unicode set in other words, just about everywhere. Systems that use ASCII are also using Unicode, since Unicode contains the ASCII set and gives them the same code points they had in ASCII [6].

Unicode Code Charts

The code charts that follow present the characters of the Unicode Standard. Characters are organized into related groups called blocks. In the Unicode Standard, character blocks generally contain characters from a single script. In many cases, a script is fully represented in its character block. There are, however, important exceptions, most notably in the area of punctuation characters.

Literature Review

Transliteration of Asian language input is a subject of recent research. During the past several years, different methods have been introduced to prepare Indian language documents by entering the text through specific transliteration schemes. Data entry through transliteration is quite close to phonetic mapping of Indian language characters to the letters of the Roman alphabet.

The earliest and widely used transliteration scheme is what is known as Library Of Congress Transliteration Scheme. This uses roman alphabets with diacritics (horizontal bars or circles added above or below roman alphabets) to represent alphabets of Indian languages. Diacritical markers added to a letter or symbol show its pronunciation, accent, etc., typically indicating that a phonetic value is different from the unmarked state. The scheme is very general in scope and hence can be used in almost all world languages. Established Tamil research centers all around the world are aware of this scheme and most of them implement this scheme as such without modifications [5].

ADAMI was one of the early Tamil word-processors for MS-DOS PCs produced by Dr. K. Srinivasan of Canada in early eighties released in 1984 to recast such transliterated text into Tamil. The Tamil text is to be typed using a plain ASCII transliteration scheme. Upon compiling and execution of the linked macro, this romanized text page is recast on screen in equivalent Tamil. One needs to return to the romanized text mode to make the corrections if any. In a more recent version of this software called THIRU, a split screen, where the roman text being typed in the bottom half of the screen is continuously recast in the upper half in Tamil. ADHAWIN is another recent implementation of the same software for Windows-based PCs [5].

Murasu and Anjal word-processing packages are widely used in Malaysian, Singaporean and Tamil Newspapers and Magazines. These packages belong to the group of "romanized input and interpreted output" tools. The ‘inaimathi’ and related fontfaces used in these packages are of the 8-bit bilingual type. The first 128 (0-127) slots are filled by roman characters as in basic ASCII and the Tamil characters occupy the upper ASCII slots (128-255). By invoking the keyboard editor it is possible to access either of these two blocks. In the Tamil typing mode, the roman keyboard strokes and their relative sequence are continuously interpreted to present equivalent Tamil characters on screen. Thus we can type 'kathai' to get the equivalent Tamil word ‘கதை’ [8].

Keyboard Configuration Program

There are number of computer programs used to develop

transliteration keyboard configuration softwares such as Keyman, C, C++, Java. In our work we take Keyman as a keyboard configuration program. Keyman is a keyboard management utility that makes it practical to input many different languages. It is fully supports Unicode and allows us to creating our own keyboard layouts for use. It interprets and translates input from the computer keyboard according to a set of rules called a keyboard. These rules are stored in a keyboard file. It includes features such as an on-screen keyboard, phonetic and visual-order input methods.

Keyman includes full support for Unicode. It support input and output of any of the thousands of characters defined in Unicode. There are two applications included in Keyman Developer: TIKE and KMComp. TIKE, the Tavultesoft Integrated Keyboard Editor is a complete environment for designing, developing, testing, and packaging our keyboards for distribution.

KMComp, the command-line compiler, is a simple tool that lets us compile keyboards, packages, and installers from the command-line. This is useful if we want to use batch builds or Make files.

Keyboard File

Keyboard file is the most important component in a keyboard configuration. It contains the set of rules to represent the particular keyboard. As we want to create a new keyboard, we want to create a keyboard file. There are two ways to create a keyboard file:

The Keyboard Wizard

It gives us a simple interface to quickly create a keyboard using a visual representation of a computer keyboard. We can drag and drop characters from a character map, and create ANSI and Unicode keyboard layouts. We cannot access most of the programs more powerful features from the Keyboard Wizard, but it will be useful to get us started on our design. We can convert keyboards created in the Keyboard Wizard to standard program source files in TIKE.

The Keyboard Language

It provides the flexibility that is needed to write keyboards with complex character management, including constraints, dead keys, post-entry parsing, virtual key management (accessing any key on the keyboard), and other features.

A keyboard file is divided into two sections: the header and the rules. The header section defines the name of the keyboard, its bitmap, and other general settings. The rules are used to define how the keyboard responds to keystrokes from the user, and are divided into groups.

The keyboard header is the first part of a keyboard; it consists of statements that help Keyman identify the keyboard and set default options for it. Each statement in the header must be on a separate line and is usually written with capital letters. The body of the keyboard is another the most important part: it determines the behavior of the keyboard. The body consists of groups, which in turn contain one or more rules which define the responses of the keyboard to certain keystrokes.

Methodology

The keyboard program interprets and translates input from the computer keyboard according to a set of rules called a keyboard. Transliteration of Tamil has to fit the need for Tamil to be recognized as the only other known language comparable to the English language with a 26-letter keyboard. It is the plan of our work to develop simple methods to use Tamil in the computer and introduce Tamil through transliteration.

We have over 230 characters in Tamil language;13 vowels(uyir), 18 consonants(meis) and compound (uyirmeis) derived from these. Tamil is one of the Indian languages where many of the compound (uyirmei) alphabets have complex geometric structure (glyph) of their own. There are 12 vowels characters and one aayitham letter in Tamil language.

There are 18 Mei Letters(consonants) and 216 Uyir-Mei. The Mei characters are created with sign Anushvara (◌்). The Uyir-Mei letters are created by the combination of the above 12 Uyir letter with the 18 Mei characters(12X18).

Also there are 13 digits used in Tamil. These character digits are now not much used by people but these characters were used in early times. They are as follows:

௦	க	உ	ங	ச	ரு	ண	ெ	அ	க	ை	ஈ	ஆ
0	1	2	3	4	5	6	7	8	9	10	100	1000

Choosing the mapping for characters

We define the output characters to be produced by the keyboard. We select the appropriate keystrokes from the QWERTY keyboard to map the output characters. Some keystrokes are used to represent output characters while some keys are not. These keystrokes that do not represent any output are called dead keys. Dead keys produce null output.

Analyzing the Keystrokes and Assigning Keystroke

We want to analyze how to create all the Tamil Characters using this limited number of codes. Some characters have direct Unicode numbers so it can be assigned directly while some other characters; they don't have their own Unicode numbers. So, we have to assign them for Unicode characters by combining two or more other Unicode characters. It is being assigned a key or collection of key strokes to a particular character or combination of characters to represent Tamil characters. To represent a character one or more key strokes can be used.

The 247 letters in the Tamil alphabet are the product of 31 basic Tamil letters. 18 English letters have similar sound connection with 18 Tamil letters. It is only the 13 remaining Tamil letters that need a 'sound connection' with English. We can make the 'sound connection,' - that is, devise the new connections- by allocating letters that are in use either in combination or singly as follows;

+ "a" > U+0B85 ⟶ அ

U+0B85 + "a" > U+0B86 ⟶ ஆ

+"A" > U+0B86 ⟶ ஆ

+ "i" > U+0B87 ⟶ இ

U+0B87 + "i" > U+0B88 ⟶ ஈ

+ "I" > U+0B88 ⟶ ஈ

Conclusion

Usage of Tamil language in computers enters a new era with the emerge of the Unicode standard with the support of more modern platforms and applications. These days, most of the Tamil websites support Unicode and typography related techniques also switching into the new standard.

This paper is useful to people who are interested in developing their own transliteration softwares to type words and sentences for their word processing work and to do World Wide Web applications easily using QWERTY keyboard.

Also this study provides solutions for some existing problems with Tamil typography. Many non-Unicode Tamil fonts with stylish glyphs are available at present. Usage of such fonts in documents can give great appearance. But due to the unfamiliar keyboard mapping to these fonts, these are not widely used in typing of Tamil. It is possible to develop these stylish fonts into familiar keyboard configuration mapping, of course with the support of keyboard configuration environment. Then we can use it with our keyboard configuration.

It is also possible to extend this keyboard configuration to other platforms like Linux, Mac OS, Solaris, etc. as these are already supporting Unicode. Only thing to be done is to set up a keyboard layout in each Operating system's native format.

Appendix

Some Typing Example.

naan or nAn ⟶ நான்

avan ⟶ அவன்

manithan ⟶ மனிதன்

paadasaalai ⟶ பாடசாலை

paLkaLaikazakam ⟶ பல்கலைகழகம்

References

[1] Acharya, Multilingual Computing for Literacy and Education, SDL, IIT Madrass, India, http://acharya.iitm.ac.in/acharya.html, 2005.

[2] Addison-Wesley Pub Co, The Unicode Standard 3.0 (www.unicode.org), 1998.

[3] Elengo, Tamil 99 Keyboard Layout, www.cadgraf.com, 2000.

[4] Ilakkuvanar, S., Tholkappiyam in English.

[5] Kalyanasundaram, K., An Overview Of Different Tools For Word-ProcessingOf Tamil And A Proposal Towards Standardisation, Institute of Physical Chemistry, Swiss Federal Inst. of Technology, 1997.

[6] Mark Pilgrim, "Python and Unicode", http:// diveintophython.org/toc /index.html.

[7] Muguntharaj, Tamil-TSCIIANJAL, 1998.

[8] Muthu Nedumaran, Murasu Anjal, 2000.

[9] Ramalingam Shanmugalingam, jAzhan,Transliteration of Tamil to English for the Information Technology, 2002.

[10] Samaranayake, V. K., Nandasara, S. T., Dissanayake, J. B., Weerasinghe, A.R.,Wijayawardhana, H., An Introduction to UNICODE for Sinhala Characters, University of Colombo School of Computing, 2003.

Rotation Invariant Texture based Image Indexing and Retrieval

Suchi Srivastava and Suneeta Agrawal

Department of CSE, MNNIT, Allahabad, U.P., India,
E-mail: suchi31@gmail.com, suneeta@mnnit.ac.in

Abstract
In this paper, a method is proposed for Image Retrieval based on analysis of texture properties of an image. Texture is the primitive image descriptors in content based image retrieval systems. We first calculate directional properties of texture pattern in each image of our database by applying Radon transformation. The directional properties are used to rotate the image with the dominant orientation of the texture patterns, then apply the Haar wavelet transformation to extract the texture feature vector of size 10. A clustering method modified ROCK is used to cluster the group of images based on feature vectors of images of database. Then representative feature vector of cluster is calculated using the average of feature vectors of all images belonging in the corresponding cluster. Similar process is applied on the submitted query image. Finally Euclidian distance between query image and representative cluster feature vector is calculated. The cluster having minimum distance is extracted from the set of clusters as result. Our experiments are conducted on Brodatz texture with different orientation and successful matching results are obtained.

Keyword: Image retrieval, Radon Transformation, Haar Wavelet Transformation, ROCK Clustering, Brodatz Texture

Introduction

During the last decade, a new image retrieval approach, called Content-Based Image Retrieval (CBIR), emerged. In this approach, the content of an image is described using low-level features such as color, texture, and shape. Despite their advantages over the traditional text-base image retrieval systems, CBIR systems face a major problem commonly referred to as the semantic gap, whereby the description of the images using the low-level features is unable to capture the semantic intended by the user in his/her queries. Therefore, CBIR systems produce a large amount of false positives in the retrieval process. A significant improvement is obtained by integrating the spatial distribution of the visual features since it captures better the contents of the images and reduces the number of false positives. The exponential growth of image data that are being generated makes it imperative to use computers to save, retrieve and analyze images. The problem of image retrieval as been an active area of research since early 70's. In order to make the best use of information in images, we need to organize the images so as to allow efficient browsing, searching and retrieval. The basic two approaches for image retrieval are text-based and visual-base. Early image retrieval techniques were generally based on textual annotation of images rather than visual features. In other words, images were first annotated with text and then searched using a text-based approach from the traditional database management systems. Content-based image retrieval (CBIR) has been an active research topic in the last few years. Comparing to the traditional systems, which represent image contents only by keyword annotations, the CBIR systems perform retrieval based on the similarity defined in terms of visual features with more objectiveness. Although some new methods, such as the relevant feedback, have been developed to improve the performance of CBIR systems, low-level features do still play an important role and in some sense be the bottleneck for the development and application of CBIR techniques. A very basic issue in designing a CBIR system is to select the most effective image features to represent image contents. Texture features is of the great majority of content based image retrieval system. However the robustness, effectiveness, and efficiency of its use in image indexing are still open issues. In image preprocessing, the features used to represent texture information and the measures adopted to compute similarity between the features of two images are critically analyzed.

In the previous method extracted feature is of large in size. We have tried to reduce the size of feature vector and also make it rotation invariant.

This paper is organized as follows. In Section 2, we introduced our proposed work. Experimental results of texture based image retrieval are described in Section 3. Conclusion & future Scope are given in Section 4.

Proposed Work for Image Retrieval

Content-based image retrieval (CBIR) is a new but widely adopted method for finding images from vast image databases. In CBIR images are indexed on the basis of low-level features, such as color, texture, and shape that can automatically be derived from the visual content of the images. Here we propose an efficient approach for image retrieval based on texture descriptor features. The steps involved in this methodology are listed below:

- Radon transformation is used for texture orientation estimation.
- Modified Haar Wavelet transformation is applied for feature extraction for 3 iteration to choose only 10 feature vectors.
- Clustering the images based on feature vectors using modified ROCK clustering algorithm.

- Computing the representative cluster feature vector using average of feature vectors of all images in corresponding cluster.
- Computing the feature vector of the query image as and when presented.
- Comparing query image feature vector with representative cluster feature vectors, identifying the closest cluster for the query image and retrieves that cluster.
- Show the result.

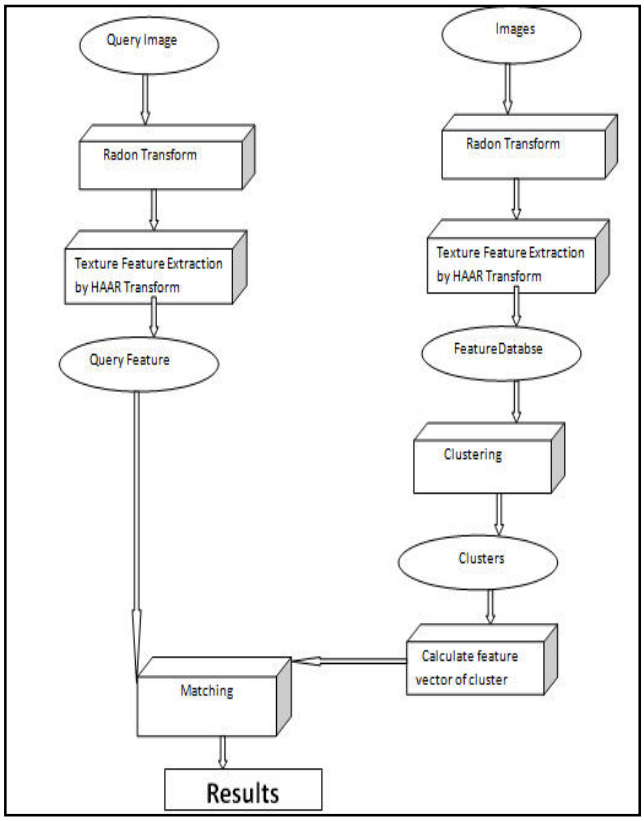


Figure 1: Proposed System Architecture

Texture Orientation Estimation using Radon transform

In order to make the image retrieval algorithm invariant to texture orientations, Radon transform is applied on the maximum circle region which is centered in the input image *f* (*x*, *y*) for a given set of angles to estimate the dominant orientation of the texture. Here, the region with circle shape is chosen since it has the least direction interference comparing with other shapes. For each given angle, the radon transform can be thought of as computing the projection of the image along the given angle. The resulting projection is the sum of the intensities of the pixels in each direction. As depicted in Figure 2, the Radon transform *R (ρ, θ)* of the input image *f(x, y)* can be defined as :

$$R(\rho, \theta) = \iint_{-\infty}^{\infty} f(x,y)\,\delta(\rho - x\cos\theta - y\sin\theta)\,dx\,dy$$

where $\rho = x\cos\theta + y\sin\theta$ is the perpendicular distance of a line from the original position and θ is the angle between the line and y-axis. $\delta(\,)$ is the Dirac delta function.

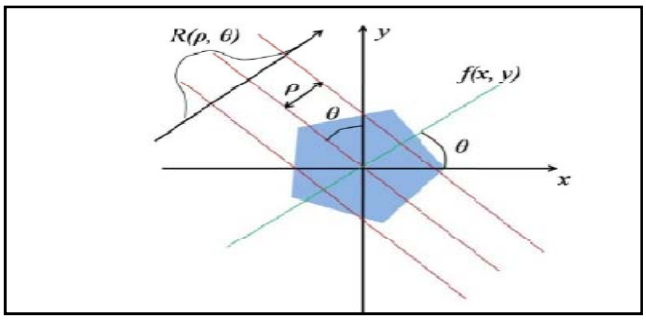


Figure 2: Radon transform of an image f(x,y)

Radon transform computes projections along θ, which varies from 0° to 180° in discrete steps of $\Delta\theta$. So for any $\Delta\theta$. The texture principal orientation can be estimated as the projection which has more straight lines. As shown in Figure 2, for the images with two or more main texture orientations, we determine the final dominant orientation by calculating the mean of the variance of projections at 6 neighbor angles around each local maxima variance. The orientation with largest mean value will be chosen as the final dominant orientation of the texture.

Feature Extraction

The objective of a feature extraction is data reduction by measuring certain features or properties which distinguish objects or their parts. Usually feature extraction is associated with another technique known as feature selection. The objective of the feature selection and extraction techniques is to reduce this dimensionality. During this process, the silent features which are essential for recognition are retained. In this paper Haar Wavelet Transformation[2] is used for extracting the texture feature.

Haar wavelets are used because they are the fastest to compute. One disadvantage of Haar wavelets is that it tends to produce large number of signatures for all windows in image. We proposed the modified Haar wavelet transformation that reducing the size of signatures.

In this feature vector computation process, we applied Wavelet Transformations only three times to get 10 sub images of input image in the following way.

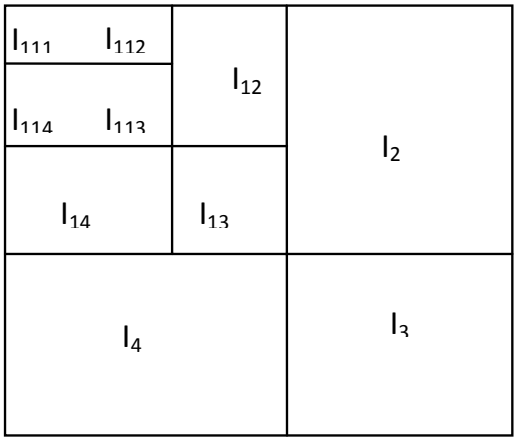


Fig-3: Haar Wavelet

Algorithm for calculating wavelet signatures
Let I be the image of size w×w.
Divide the image I into four bands I_1,I_2,I_3,I_4 based on Haar wavelet of size w/2×w/2.
Compute Signatures f_r for I_1, I_2, I_3, I_4 .
Now take the image I_1 and divide it into 4 bands namely I_{11},I_{12},I_{13},I_{14} of size w/4×w/4
Compute signatures f_r for I_{12},I_{13},I_{14}
Again take the I_{11} and divide it into 4 bands namely I_{111}, I_{112}, I_{113}, I_{114} of size w/8×w/8.
Now we obtain 10 signatures then stop the process.

The Wavelet signature (texture feature representation) is computed from sub image as follows,

$$f_r = \sqrt{\frac{c_{ij}^2}{i \times j}}$$

Where f_r is the computed Wavelet signature (texture feature representation) of the sub image, C_{ij} is the representation of the intensity value of all elements of sub image and i × j is the size of the sub image.

Image Indexing

The objective of image indexing is to retrieve similar images from an image database for a given query image (i.e., a pattern image). Each image has its unique feature. Hence image indexing can be implemented by comparing their features, which are extracted from the images. The criterion of similarity among images may be based on the features such as color, intensity, shape, location and texture, and above mentioned other image attributes. Current Image indexing techniques are of two types,
Textual (manual)
Content- based (automated)

In this paper, we use content based indexing method based on clustering. The basis of the clustering method in indexed image database is that, the images belonging to the same cluster are similar or relevant to each other when compared to images belonging to different clusters. We have used ROCK Clustering method for fast image retrieval.

ROCK Clustering Algorithm[4]-The ROCK algorithm is divided into three major parts
Draw a random sample from the data set
Perform a hierarchical agglomerative clustering algorithm
Label data on disk

In our case, we do not deal with a very huge data set. So, we will consider the whole data in the process of forming clusters, i.e. we skip step1 and step3.

Draw a random sample from the data set
Sampling is used to ensure scalability to very large data sets
The initial sample is used to form clusters, then the remaining data on disk is assigned to these clusters
In my case, I will consider the whole data in the process of forming clusters.

Perform a hierarchical agglomerative clustering algorithm
ROCK performs the following steps which are common to all hierarchical agglomerative clustering algorithms, but with different definition to the similarity measures:

a. places each single data point into a separate cluster
b. compute the similarity measure for all pairs of clusters
c. merge the two clusters with the highest similarity (goodness measure)
d. Verify a stop condition. If it is not met then go to step b.

Label data on disk

Finally, the remaining data points in the disk are assigned to the generated clusters. This is done by selecting a random sample Li from each cluster Ci, then we assign each point p to the cluster for which it has the strongest linkage with Li. As we said, we will consider the whole data in the process of forming clusters.

Computation of links

1. Using the similarity threshold θ, we can convert the similarity matrix into an adjacency matrix (A)
2. Then we obtain a matrix indicating the number of links by calculating (A x A) , i.e., by multiplying the adjacency matrix A with itself.

Query

Query by example allows the user to formulate a query by providing an example image. The system converts the example image into an internal representation of features. Images stored in the database with similar features are then searched. Query by example can be further classified into query by external image example, if the query image is not in the database, and query by internal image example, if otherwise. For query by internal image, all relationships between images can be pre-computed.

Experiment & Results

The efficiency of implemented method was evaluated on a set of 13 textures selected from the Brodatz album. The 13 textures are shown in Figure 4. A database of 91 images are constructed by rotating each texture on 0º, 30º, 60º, 90º, 120º, 150º, 200º degree.

To get desired result, intel® Core TM i3 CPU M350@ 2.27 GHz with 2 GB RAM , 300 GB hard Disk , MatLab7.11. Window 7 Operating System, Microsoft Office 2007 Pack are used.

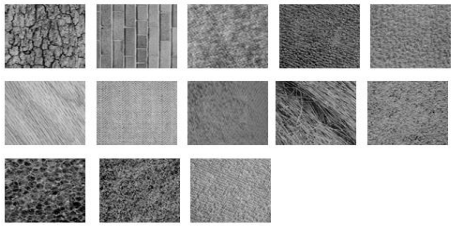


Figure 4: Samples of 13 different textures used in the test

Fig.5 shows the query image. Table 1 shows the feature vector values or feature vectors of sub images.

Fig 5: Query Image

Table 1- feature vectors of figure 5

Sub-Images	Feature Vectors
I_A	877.6354
I_B	1528.706
I_C	1582.835
I_D	47.76519
I_E	91.09537
I_F	90.70711
I_G	622.6505
I_H	43.56737
I_J	88.78239
I_K	85.02783

The clustered images from the database are shown in Fig 6. The fig.6 clearly represents matching images with the original (query) image and it has removed all non-relevant images.

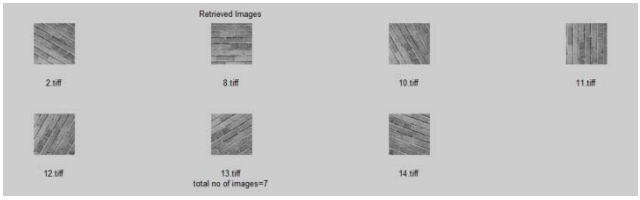


Figure 6: Clustered Image set according to query image.

Conclusion and Future Scopes

By using Radon transform we have made this algorithm rotation invariant. By deriving ten feature vectors or feature vectors from wavelet transformation in three iterations reduces overall time complexity than previous methods. The new method proposed in this paper for clustering effectively minimizes the undesirable results and gives a good matching pattern, that will be having zero or a minimum set of no relevant images.

The present system operates partially at the primitive feature level. The present system extracts only the Texture feature of an image. Here scale and illumination invariant method may also inbuilt to make this algorithm more efficient. This system can be enhanced to extract the other primitive features also.

References

[1] R. C. Gonzalez, R. E.Woods, Digital Image Processing, Pearson Education 2002.

[2] N Gnaneswara Rao ,Dr.V Vijaya Kumar, V Venkata Krishna, "Texture Based Image Indexing and Retrieval IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.5, May 2009

[3] Wasim Khan, Shiv Kumar, Neetesh Gupta, Nilofar Khan "Signature Based Approach For Image Retrieval Using Color Histogram And Wavelet Transform," International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307 (Online), Volume-1, Issue-1, March 2011

[4] Guha S.,Rastogi R., and Shim K.ROCK: A robust clustering algorithm for categorical attributes. In proceeding Conclusions of the IEEE International Conference on data engineering,Sydney,March 1999.

[5] Eric J. Stollnitz Tony D. DeRose David H. Salesin University of Washington "Wavelets for Computer Graphics: A Primer".

A Turning from Virtual Environment to Reality- Communication with Non Human Devices in Natural Way

Chhaya Kinkar[1], Richa Golash[1] and Akhilesh Upadhaya[2]

[1]Sagar Institute of Research Technology and Science- Bhopal , India
[2]Sagar Institute of Research and Technology – BhopaI , India
E-mail: chhayakinkar@gmail.com , golash.richa@gmail.com ,akhileshupdhaya@gmail.com

Abstract
Communication with devices in a natural way as human being do has been science fiction. Today with powerful smart microphones and microprocessor, science fiction is becoming a reality. This paper present a method/model where input to a device is speech signal, the device will sense it ,after which speech is converted into text & generated text is process for machine translation so that the device will work according to instruction given by human being orally. The proposed model/method also take care of surrounding noise, false operation by unauthorized person, & handle large amount of vocabulary to avoid complexity present in natural speech signal. Ambiguities of sentences are also solved by proper design of context free grammar.

Keywords: Machine translation, natural speech signal, context free grammar, Ambiguities

Introduction

Language- Language is a system for communication. It includes both verbal and written expression which help us to communicate our thoughts and feelings. Ability to speak & write and to communicate is one of the most fundamental aspect of human behavior. As the study of human languages developed the concept of communicating with non human devices was investigated. Presently human communicate with devices with the help of key board, remote, keypad etc & the search for an alternative method is going on. Speech reorganization technology is on of the solution to the problem .The idea to turn virtual environment into real is to design and build a system that will analyze, understand, and work according to human voice signal or we will call it as a speech signal.

As Language is a very fast and effective way of communicating. To use language means to express an unlimited amount of ideas, thoughts and practical information by combining a limited amount of words with the help of a some grammatical rules [1].We know that a good software professional can type about 300 key strokes (letters) per minute. Since the average speaking rate is about 150 words per minute (with some variance between the speakers and the languages) [2], that is if a system is build up which works according to instruction given by a person orally the over all data rate is very fast as compare with typing the data.

The over all proposed model /method will be divided in to 3 broad sections the first section will be reorganization of voice/speech signal, second section will content different algorithms & grammatical rules about the language, & generated speech is converted into text, third section will be machine transition of instruction given by human beings orally. To avoid any false working of device, speech is converted in to text.

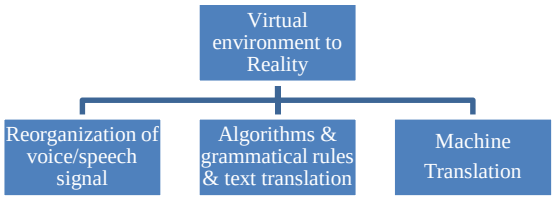


Fig 1: Proposed natural language communication system

Recognition of voice/speech signal

In US & some other countries speech group started in 2005 to developed several successful speech recognition services [3], all the services developed by the group can recognized speech with the accuracy of 90%, the reaming 10%.may cause false operation of the device. To increase this efficiency up to the 96% plus, the proposed model /method calculate WER, semantic Quality, OOV, latency with the help of genetic algorithms. The total proposed speech reorganization system is shown in figure 2 in the from of block diagram. Here a time synchronous finite state transducer decoder[2] is used for acquisition of voice/speech signal, after receiving voice/speech signal to avoid false operation background noise is eliminated from the received signal, & the signal is segmented in to some blocks like voiced & unvoiced sound signal etc, start & end point of speech is detected from these blocks.

Human sound can be characterized by pitch related to the frequency of the sound that is loudness, the physiological perception of the intensity of sound & Quality (a property possessed by sound by virtue of its harmonic content),all these features of speech signal are extracted from the received signal by applying windowing techniques along with soft computing approach.

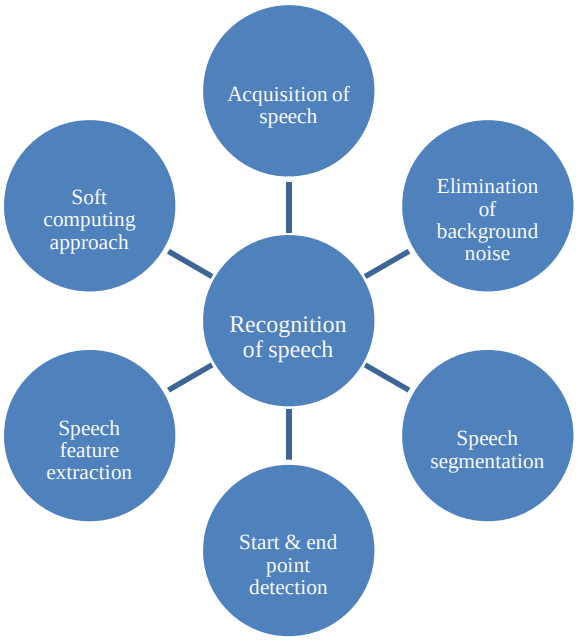


Figure 2: Block diagram of proposed speech recognition system

In this model/method WER, semantic Quality, OOV, latency are calculated as,

Word Error Rate (WER)

The word error rate measures misrecognitions at the word level. It compares the words output by the recognizer to those really spoken by the speaker. Every error (substitution, insertion or deletion) is counted against the recognizer.

$$WER = \frac{\text{(Number of Substitution + Insertions + Deletions)}}{\text{Total number of words}}$$

Semantic Quality

Individual word error will also effect the generated text. For example, misrecognition of the plural form of a word (missing "s") would also change the generated text therefore tracking of the semantic quality of the recognizer is imported.

$$\text{Semantic quality} = \frac{\text{No. of correct recognition}}{\text{Total number of spoken words}}$$

Out-of-Vocabulary (OOV) Rate

The out-of-vocabulary rate tracks the percentage of words spoken by the speaker to that are not present in vocabulary. It is important to keep this number as low as possible. Any word spoken by speaker that is not in vocabulary will ultimately result in a recognition error. Furthermore, these recognition errors also cause errors in next word due to the acoustic misalignments.

Latency

Latency is defined as the time from when the speaker finishes speaking & the text appear on the screen. Many factors contribute to latency as (a) the time it takes the system to detect end-of-speech, (b) the total time to recognize the spoken query etc.

Algorithm & text generation

When we consider general purpose devices/machine, normally the key pad contain letters from A to Z & numbers from 0 to 9 [4], hence the vocabulary designed in this model/method consist of digits 0 to 9 and more than hundred general purposed used words like start, stop, enter, erase, help, yes, no, go, repeat, do, select, channel, delete etc. The enormous number of sentences of natural language are of average length for example, a language of 10,000 words consist of the number of different string of n words or less would be $10^{4(n+1)}$ ·However not all of these are sentence, but if we take 50% redundancy, we can give $10^{2(n+1)}$ as a reasonable estimate for the number of sentences [5]. In purposed method/model recognition routines are designed using this concept. Algorithms & grammatical rules are constructed based upon the structural properties of grammar of natural language. A grammar consist of a finite set of sentence formation rules that is grammatical rules. All of which draw their symbol from a single finite source that is vocabulary.

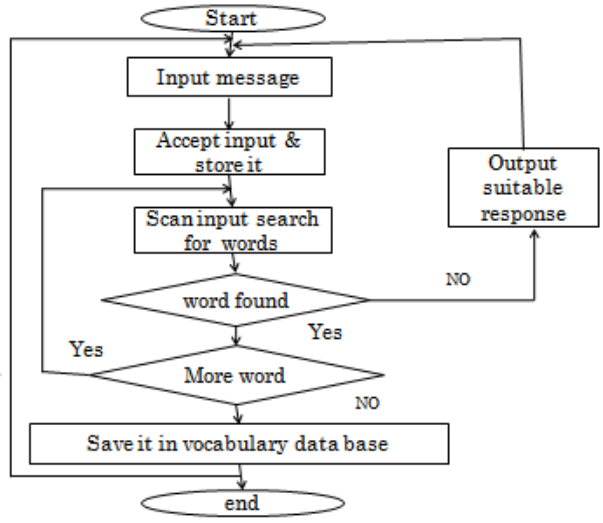


Figure 3: Flowchart for design of vocabulary

The grammatical rules are applied in a given linear order for generation of text.

Machine Translation

The pronunciation of same word vary from person to person & country to country , In purposed model/method to avoid this speech is first converted into text.& if the generated text is right, the text is process for machine translation.

Rule-based statistical technique is used for machine translation of voice/speech signal. This technique is hybrid, large-scale system & is capable of performing lexical and phrasal translations directly from segmented data. The over all machine translation process for English words, in the from of

block diagram is shown in figure 4

Segmented speech signal is tag with end marker to indicate end point, after which a token is assign to every segment of speech signal. In parsing phase a lexicon value is assign to every token. The assignment of lexicon value is completely depends on function of word translations ,all lexical translation information is learned automatically from data & algorithms. The translation time is fast enough for device operation. After completion of machine translation respective interrupt signal is generated to perform the specific operation.

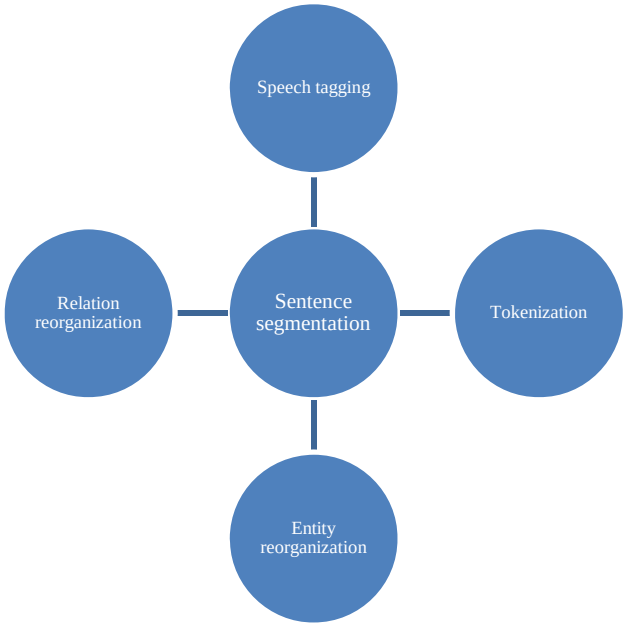


Fig 4: Block diagram for machine translation of oral instruction

Conclusion

As we are moving toward an environment full of devices/ machine, therefore it has become our first requirement that this environment must be eco friendly. The approach in this paper of handling devices/machines naturally is a step towards it.

References

[1] Susanne Wagner (Halle) “Intralingual speech-to-text-conversion in real-time: Challenges and Opportunities”EU-High-Level Scientific Conference Series,MuTra 2005 – Challenges of Multidimensional Translation: Conference Proceedings

[2] B.-T. Zhang and M.A. Orgun “Speech Recognition for Mobile Devices at Google” PRICAI 2010, LNAI 6230, pp.8–10, 2010._c Springer-Verlag Berlin Heidelberg 2010

[3] Etienne Barnard, Johan Schalkwyk, Charl van Heerden, Pedro J. Moreno” Voice Search for Development” Human Language Technologies Research Group, Google Research, New York, NY, USA INTERSPEECH 2010

[4] L.A.Smith “Selection of speech recognization features using a geneticalgorithms” proc IEEE Int.cont.Acoust.,Speech and signal processing

[5] G.H.Matthews”Analysis by synthesis of sentences of Natural languges”Inter national conference on machine translation of language and applied languae analysis.National physis laboratory,Taddington,U.K.,5-8Sept 2007

[6] Md. Rabiul Islam, Md. Fayzur Rahman “Noise Robust Speaker Identification using PCA based Genetic Algorithm” International Journal of Computer Applications (0975 – 8887) Volume 4– No.12, August 2010

[7] Brandon Ballinger, Cyril Allauzen, Alexander Gruenstein, Johan Schalkwyk “On-Demand Language ModelInterpolation for Mobile Speech Input” Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA INTERSPEECH 2010

[8] Johan Schalkwyk, Doug Beeferman, Fran_coise Beaufays, Bill Byrne,Ciprian Chelba, Mike Cohen, Maryam Garret, Brian Strope “Google Search by Voice: A case study” Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA, INTERSPEECH 2010

[9] Dmitriy Genzel “Automatically Learning Source-side Reordering Rules for Large Scale Machine Translation”Google, Inc.

[10] Michael Riley, Cyril Allauzen, and Martin Jansche,” An Open-Source, Weighted Finite-State Transducer Library and its Applications to Speech and Language” Google, Inc.

[11] Christian Krapichlera, Michael Haubnera, Andreas Löscha,Manfred K. Langb, Karl-Hans Englmeiera “ Human-machine interface for a VR-based medical imaging environment”

[12] M. Aissiou and M. Guerti “Genetic Algorithms Application forthe Automatic Recognition of the Arabic Stop Sounds” Journal of Applied Sciences Research, 3(5): 358-366, 2007 © 2007, INSInet Publication

Performance Sensitivity of FWM Effects to System Parameters in High Capacity WDM

Shankar Duraikannan[#1] and P. Rajeswari[#2]

[1]*Lecturer, Engineering, Asia Pacific University College of Technology and Innovation, Kuala Lumpur, Malaysia.*
[2]*Assistant Professor, Department of ECE, Sri Sai Ram Institute of Technology, West Tambaram, Chennai, India.*
E-mail: [1]itzdshankar@yahoo.co.in [2]raji_ece2009@yahoo.com

Abstract
Wavelength Division Multiplexing (WDM) is a boon of Light wave Technology as it increases the transmission capacity and enhances the transmission distance by using optical amplification. The capacity of the present system is strongly influenced by the noise generated by the amplifiers and interactions due to nonlinearities. According to many literatures, Stimulated Raman Scattering (SRS) and Four Wave Mixing (FWM) are dominant nonlinear effect where the other nonlinear effects like Self Phase Modulation (SFM), Cross Phase Modulation (CFM) and Stimulated Brillouin Scattering (SBS) can be suppressed. This paper focus on simulation and theoretical issues of fiber nonlinearities and impact of FWM on WDM systems. The Effect of FWM on channels with different values of inter channel spacing, input power, length, effective core area has been analyzed. From the results obtained, it is found that the FWM effect is more dominant at the middle wavelengths compared to end wavelengths and also the FWM noise power is proportional to the power of the channel, length and effective core area i.e., with the increase in any one of the above mentioned parameters, the FWM noise power increases. FWM have become significant at high optical power levels and when the capacity of the optical transmission line is increased, which has done by decreasing the channel spacing. The simulation results confirm that the fiber nonlinearities play decisive role in the WDM system.

Keywords: Wavelength division multiplexing, Fiber nonlinearities, Four-wave mixing.

Introduction

We are moving toward a society which requires that we have access to information at our fingertips *when we need it, where we need it, and in whatever format we need it*. The information is provided to us through our global mesh of communication networks, e.g., today's Internet and asynchronous transfer mode (ATM) networks, whose current implementations do not have the capacity to support the foreseeable bandwidth demands. Fiber-optic communication is a method of transmitting information from one place to another by sending light through an optical fiber. Fiber-optic communication systems have revolutionized the telecommunications industry and played a major role in the advent of the information age. Often the optical fiber offers much higher speed than the speed of electronic signal processing at both ends of the fiber. Because of its advantages over electrical transmission, the use of optical fiber has largely replaced copper wire communications in the developed world.

The main benefits of fiber are its exceptionally low loss, allowing long distances between amplifiers or repeaters and its inherently high data-carrying capacity, such that thousands of electrical links would be required to replace a single high bandwidth fiber. FTTH deployed with Passive Optical Network technology seems to be the best solution to alleviate the bandwidth bottleneck in the access network.

Fiber optic technology with WDM can also be considered as a saviour for meeting our needs because of its potentially limitless capabilities like huge bandwidth (50 terabits per second (Tb/s)), low signal attenuation (as low as 0.2 dB/km), low signal distortion, low power requirement, small space requirement, and low cost [2][3].The wavelength division multiplexing has dramatically increased the network capacity. This allows the transport of hundreds of gigabits of data on a single fiber for distances over thousands of kilometers, without the need of optical-to-electrical-to-optical (O-E-O) conversion.

For efficient recovery of received signal, the signal to noise ratio at the receiver must be considerably high. Fiber losses will affect the received power eventually reducing the signal power at the receiver. Hence optical fibers suffer heavy loss and degradation over long distances. To overcome these losses, optical amplifiers were invented which significantly boosted the power in the spans in between the source and receiver.

In long haul transmission, EDFA's (Erbium Doped Fiber Amplifier) are used to compensate the signal attenuation instead of optoelectronic / electro optic conversions and DSF (Dispersion Shifted Fibers) to overcome chromatic dispersion. Booster amplifier, in-line amplifier, Pre- amplifier are mostly preferred for long distance transmission. Singh et al made the comparative study of all the amplifiers and concluded that inline amplifier is most preferable as it requires minimum power for the given probability of error [10].

All these attempts are made to maintain high bit rate however, optical amplifiers introduce amplified spontaneous emission (ASE) noise which is proportional to the amount of optical amplifications they provide and some undesirable nonlinear interactions such as Four Wave Mixing and Stimulated Raman Scattering are created and accumulated as the optical signals propagate along the length of the fiber [5]-

[9]. Optical fiber nonlinearities affect the system parameters like total transmission distance, channel spacing, power per channel and so on [1]-[4]. Low loss in optical fibers is still a critical requirement in long distance optical systems to efficiently recover the signal at the receiver.

Impact of fiber Nonlinearities on transmitted optical power

Refractive Index of the fiber is both intensity and frequency dependent. Nonlinear Kerr effect is dependence of the refractive index of the fiber on the power that is propagating through it. This effect is responsible for the generation of non linear effects like four wave mixing (FWM), stimulated Raman scattering (SRS).

Stimulated Raman Scattering (SRS)

SRS is first observed in 1972 and it belongs to inelastic scattering. Optical phonons participate in Raman scattering. The SRS is scattering of a photon by one of the molecules to a lower-frequency photon, while the molecule makes transition to a higher energy vibrational state. Incident light acts as a pump for generating the frequency-shifted radiation called the Stokes wave. The scattered signal has a wavelength longer than incident light due to which longer wavelength channels are amplified by the depleting shorter wavelength channels. As the signal propagates along a long haul fiber, lower wavelength channel gets completely depleted due to SRS resulting in the degradation of SNR. Careful Examination on SRS amplification and depletion power is to be done. The SRS effect in DWDM fiber optic system is examined by many Authors [15]-[19].

Ignoring walk-off effects, Singh and Hudiara [11] have given a model to calculate SRS without any assumptions.

Modified power due to SRS is given by

$$P_m[k] = P_t[k] - P_t[k] \sum_{i=k+1}^{N} D[j,k] + \sum_{j=1}^{k-1} P_t[j]\, D[j,k] \quad (1)$$

$$\text{for } k=1,2,3...N$$
$$D[k,i] = 0 \text{ for } i>N$$
$$D[j,k] = 0 \text{ for } k = 1$$

$$D[i,j] = (\lambda_j/\lambda_i)\, P_t[j] \{(f_i - f_j)/ 1.5 \times 10^{13}\} g_{rmax} \{ L_e(\lambda_j) \times (10^5/b) \times A_e\}$$
$$\text{for } (f_i - f_j) \le 1.5 \times 10^{13} \text{ Hz and } j>i$$
$$D[i,j] = 0$$
$$\text{for } (f_i - f_j) > 1.5 \times 10^{13} \text{ Hz and } j \le I \quad (2)$$

In Eq.1, $P_m[k]$ represents the total power transmitted to K_{th} channel, second term gives the total power depleted from the K_{th} channel by amplifying the higher wavelength channels and the third term indicates the total power received by the K_{th} channel from the lower wavelength channels.In Eq.(2), g_{rmax} is the peak Raman gain coefficient (cm/W), λ_j, λ_i are the wavelengths (nm)of jth and ith channels, f_i, f_j are the centre frequencies (Hz) of the i_{th} and j_{th} channels, A_e is the effective core area of optical fiber in cm^2 and the value of b varies from 1 to 2 depending up on the polarization state of the signals at different wavelength channels.

Four Wave Mixing (FWM)

Four-wave mixing (FWM) is a major source of nonlinear crosstalk for WDM light wave systems. Consider three waves co propagate along the fiber with the frequencies f_i, f_j, f_k, FWM is the generation of new wave with the frequency $f_{ijk} = f_i + f_j - f_k$ resulting from the interaction of three waves. If the newly generated wavelength falls in the window of original transmitting channel wavelength, it causes cross talk between the channels propagating through fiber resulting in severe degradation of the WDM channels. Probability of this match increases for equally spaced channels [20]-[23]. In fact, these spurious signals fall right on the original wavelength which results in difficulty in filtering them out.

FWM is one of the major limiting factors in long haul optical communication system which uses low channel spacing or low dispersion fiber as medium. The WDM has concept of propagating the different wavelength channels separated by a particular spacing between each of them in terms of nanometers causes interaction between them weakly. This weak interaction becomes significant in the long range of distance between transmitter and receiver. For a system with in-line amplification, the FWM effect will be more severe [11]. FWM occurs in DWDM systems in which the wavelength channel spacing are very close to each other. This effect is generated by the third order distortion that creates third order harmonics.

The power of the newly generated FWM component at the frequency f_{ijk} [12] [13] with in-line amplification is given by

$$P(f_{ijk}) = k^2 P^3 e^{-aL} [\{(M+1)L_e\}/A_e]^2 \ \eta_{ijk} (d_{ijk})^2 P \quad (3)$$
$$\text{Where } k=32\Pi^2(X/n^2 c\lambda)$$

$$\eta_{ijk} = \{\alpha^2/[\alpha^2 + (\Delta\beta_{ijk})^2]\} \times [1 + \{4e^{-aL}/(1 - e^{-aL})\}^2] \sin^2(\Delta\beta_{ijk} L/2) \quad (4)$$

$$\Delta\beta_{ijk} = (2\Pi\lambda^2/c)(|f_i - f_k| |f_j - f_k|)\{D + (dD/d\lambda)(\lambda^2/2c)(|f_i - f_k| + |f_j - f_k|)\} \quad (5)$$

where n is refractive index of the fiber, λ is centre wavelength, X is third-order non linear electric susceptibility, P is power injected in the channel, a is fiber attenuation coefficient, M is number of amplifiers, L_e is effective system length, A_e is effective area of fiber, d_{ijk} is degeneracy factor, D is dispersion coefficient, α is total fiber attenuation, L is system length and η_{ijk} is FWM efficiency.

$$L_e = (1 - \exp(-\alpha L))/\alpha \quad (6)$$

Two factors strongly influence the magnitude of the FWM products. The first factor is the channel spacing, where the mixing efficiency increases dramatically as the channel spacing becomes closer. Fiber dispersion is the second factor and the mixing efficiency is inversely proportional to the fiber dispersion, being strongest at the zero-dispersion point. In all cases, the FWM efficiency is expressed in dB, and more negative values are better since they indicate a lower mixing efficiency.

Amplified Spontaneous Emission (ASE)

WDM system makes use of the optical in-line amplifiers to reduce the fiber loss in long distance transmission. Due to in-

line amplifiers, ASE noise is generated and accumulated as the coupled light travels along the fiber. It influences BER (Bit Error Rate) and capacity of the channel [1]

Chraplyvy [14] provided a model for the calculation of ASE

$$P_{ase}=2nsp\ (G-1)\ hfB_oM \quad (7)$$

Where h is Plank's constant ($6.63x10^{-34}$ Js), f is centre frequency, G is gain of amplifier, B_o is equivalent rectangular optical bandwidth in Hz, nsp is population inversion parameter, M is number of amplifiers.

Considerations for simulating the effects of FWM

The simulations are carried out using Matlab to determine the FWM noise power in an WDM network as per the mathematical model stated in (3) with the input power, system length, interamplifier spacing values fixed and by varying the interchannel separation in steps of 0.01 x 10^{-6}m.

Input power per channel = 3mW,
Fiber attenuation coefficient = 0.205 dB/km,
Number of channels = 4,
System length =1000 km,
Effective area of the optical fiber =$5.3x10^{-7}$ cm^2,
Dispersion coefficient = 3,
Degeneracy factor = 6,
Interamplifier spacing = 100 km.

Results and Discussion

The simulation results of the program are presented in graphical form. Fig.1 shows the variation of FWM noise power Vs wavelength. The FWM noise power at 1.5 x 10^{-6} m with a channel spacing of 0.01 x 10^{-6} m is -51.1dBm and with 0.04 x 10^{-6} m is -55.5 dBm. It is proven from the figure that the FWM noise power is more severe in the middle wavelengths when compared to end wavelengths (High and Low) and by putting the maximum spacing between the channels, causes low interaction between them and results in low FWM noise power.

Fig 2, Fig 3, Fig 4 and Fig 5 depicts the variation of FWM noise power in dBm Vs length (L) in km for different values of channel input power (1mw, 10mw, 20mw, 30mw) and interchannel spacing (0.01 x 10^{-6}m, 0.02 x 10^{-6}m, 0.03x 10^{-6}m, 0.04 x 10^{-6}m). As shown in Fig 2, when more power is launched through the fiber, the system gets affected by FWM noise power.

Another important relation from above result was the dependency of the FWM power with the system propagating length. The FWM power at 10Km for input channel power of 1mW was around -41dBm and the corresponding FWM power at the system length of 100Km was around -37dBm which is equal to the FWM power produced at 11Km at channel input power of 30mW. For better system performance right value of input channel power and the regenerator distance has to be chosen.

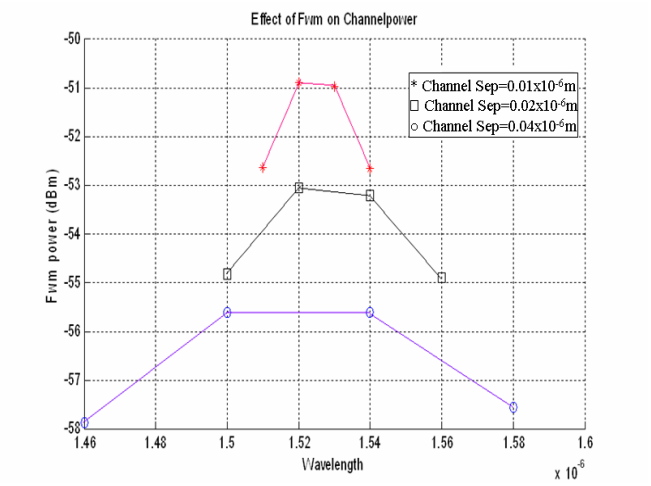


Figure 1. FWM Noise Power Vs Wavelength

For different values of interchannel spacing (10nm, 20nm and 40nm) in WDM transmission system with Fiber attenuation coefficient = 0.205 dB/km, Effective area of the optical fiber =$5.3x10^{-7}$ cm^2, Dispersion coefficient = 3, Degeneracy factor = 6.

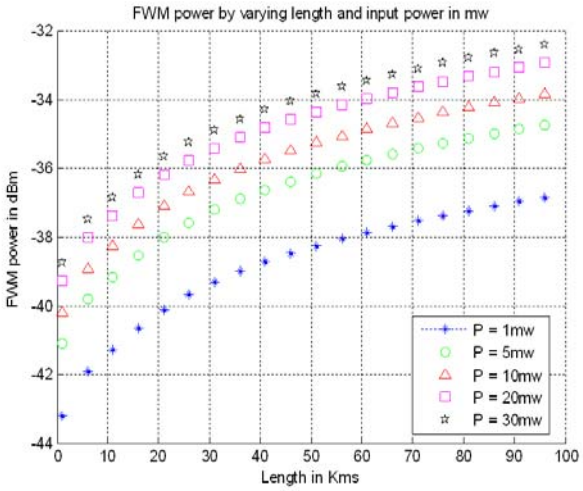


Figure 2. FWM Noise Power in dBm Vs Length in Kms

For different values of input power (1mw, 5mw, 10mw, 20mw, 30mw) with an interchannel spacing of 0.01 x 10^{-6}m in WDM system with Fiber attenuation coefficient = 0.205 dB/km, Effective area of the optical fiber =$5.3x10^{-7}$ cm^2, Dispersion coefficient = 3, Degeneracy factor = 6.

On comparing Fig.2 and Fig.3, the FWM power at a length of 10Km for input channel power of 1mW with an interchannel spacing of 0.01 x 10^{-6}m was around -41dBm whereas the FWM power with same length and same input channel power with an interchannel spacing of 0.02 x 10^{-6}m is around-40dBm. The increase in interchannel spacing can control the effect of FWM.

According to the simulation results shown in Fig.2, Fig.3, Fig.4 and Fig.5, FWM power increases with the increasing

system length and input injected channel power assuming all channels were transmitted in the same power and decreases with increase in channel spacing between the channels. Fig6. illustrates the FWM noise/crosstalk power variation for different values of system length while considering third order and fifth order dispersion parameters. The FWM power variation was from -36.2dBm to -33.4 dBm for L = 5 Km. Fig6. shows the dependency of FWM on input power and length. Therefore FWM noise power is proportional to input power and Length.

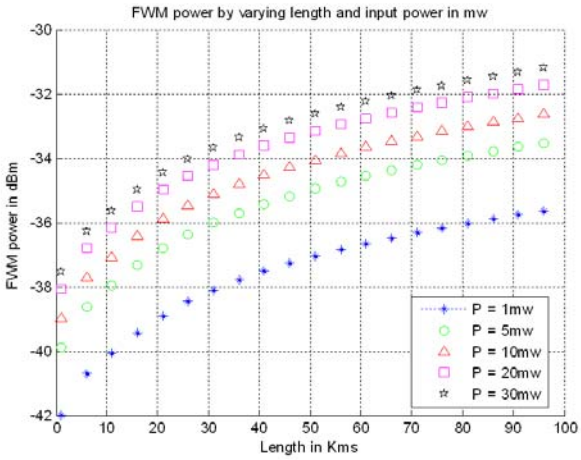


Figure 3. FWM Noise Power in dBm Vs Length in Kms

For different values of input power (1mw, 5mw, 10mw, 20mw, 30mw) with an interchannel spacing of 0.02 x 10^{-6}m in WDM system with Fiber attenuation coefficient = 0.205 dB/km, Effective area of the optical fiber =5.3x10^{-7} cm^2, Dispersion coefficient = 3, Degeneracy factor = 6.

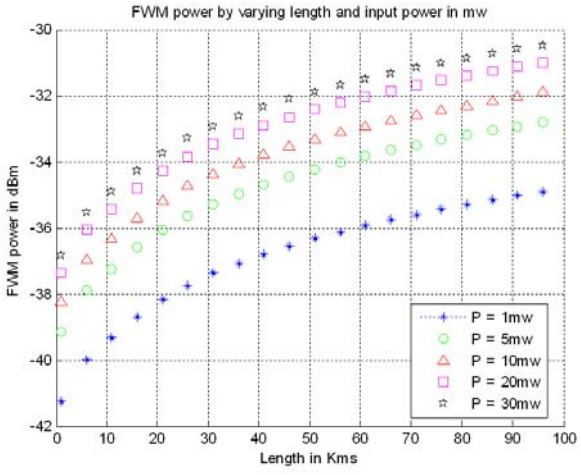


Figure 4. FWM Noise Power in dBm Vs Length in Kms

For different values of input power (1mw, 5mw, 10mw, 20mw, 30mw) with an interchannel spacing of 0.03 x 10^{-6}m in WDM system with Fiber attenuation coefficient = 0.205 dB/km, Effective area of the optical fiber =5.3x10^{-7} cm^2, Dispersion coefficient = 3, Degeneracy factor = 6.

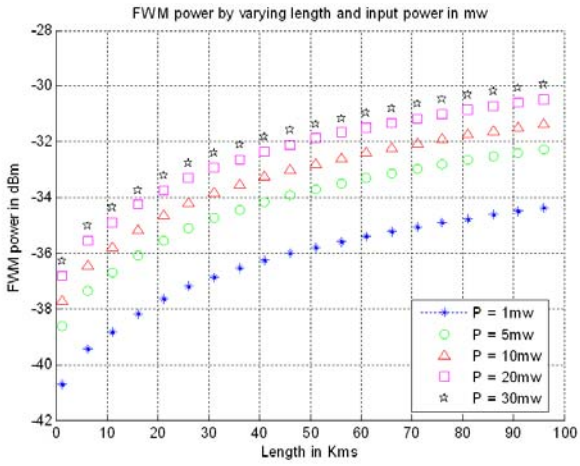


Figure 5. FWM Noise Power in dBm Vs Length in Kms

For different values of input power (1mw, 5mw, 10mw, 20mw, 30mw) with an interchannel spacing of 0.04 x 10^{-6}m in WDM system with Fiber attenuation coefficient = 0.205 dB/km, Effective area of the optical fiber =5.3x10^{-7} cm^2, Dispersion coefficient = 3, Degeneracy factor = 6.

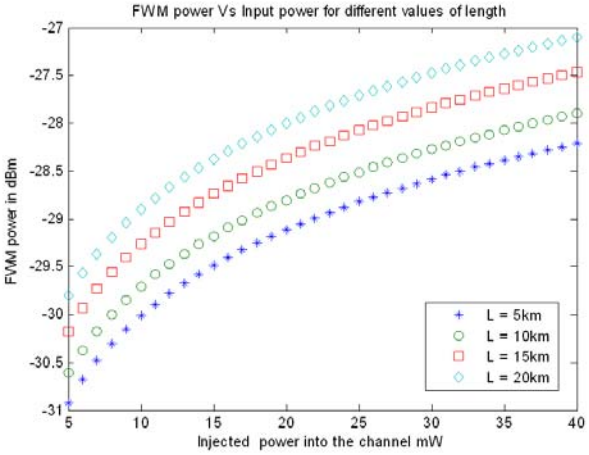


Figure 6. FWM Noise Power in dBm Vs Injected power in mW

For different values of Length (5Km, 10Km, 15Km, 20Km) in WDM transmission system with Fiber attenuation coefficient = 0.205 dB/km, Effective area of the optical fiber =5.3x10^{-7} cm^2, Dispersion coefficient = 3, Degeneracy factor = 6.

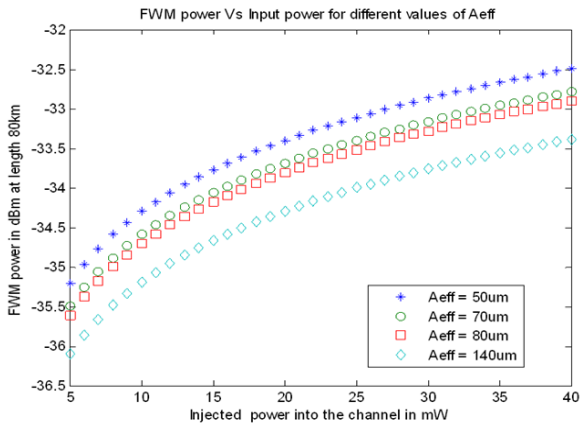


Figure 7. FWM Noise Power in dBm Vs Injected power in mW

For different values of Length (5Km, 10Km, 15Km, 20Km) in WDM transmission system with Fiber attenuation coefficient = 0.205 dB/km, Effective area of the optical fiber $=5.3x10^{-7}$ cm^2, Dispersion coefficient = 3, Degeneracy factor = 6.

Fig.7 shows the FWM noise power variation for different values of Effective core area. The interaction between the channels will be more when the channel spacing is low which increases the FWM noise power. For A_{eff} = 50µm, the FWM power was high which varies between -35.2dBm to -32.5dBm and for A_{eff} = 140µm the FWM power was low compared to all others. The decision of spacing between the channels was decided by considering the available channel capacity, bandwidth and number of channels that need to propagate through the same fiber.

From the simulation results shown as figures 2-7, we infer that even in coarse DWDM systems with larger spacing of channels, the transmission capacity of the system is affected by FWM noise due to its dependency on system parameters.

References

[1] G. Kaur, M.L. Singh, *"Effect of four wave mixing in optical communication"*, Optik- International journal of light and Electron optics 120(2009) 268-273.

[2] P.R.Trischitta, W.C.Marra, *"Applying WDM technology to undersea cable networks"*, IEEE Communication Magazine 36(2) (1998) 62-66.

[3] N.S.Bergano, *"Wavelength division multiplexing in long haul transoceanic transmission systems"*, Journal of light wave technology 23 (12) (2005) 4125-4139.

[4] P.Bayvel, R.Killey *"Optical fiber Telecommunications IV-B systems and impairments"*, Chapter 13. Nonlinear Optic Effects in WDM transmission, Academic Press, 2002, pp.611-641.

[5] A.Yu, M.J.O'Mahony, *"Optimization of wavelength spacing in a WDM transmission system in the presence of fibre nonlinearities"*, IEEE proceedings-J Optoelectron 142 (4) (1995) 190-196.

[6] M.Wu, W.I.Way, *"Fiber nonlinearity limitations in ultra dense WDM systems"*, Journal of Lightwave Technology 22 (6) (2004) 1483-1498.

[7] D.G. Schadt, *"Effect of amplifier spacing on four wave mixing in multichannel coherent communications"*, Electron. Lett. 27 (20) (1991) 1805–1807.

[8] A.R.Chraplyvy, *"Limitations on lightwave communications imposed by optical-fibre nonlinearities"*, J. Lightwave Technol. 8 (1990) 1548–1557.

[9] D. Marcuse, A.R. Chraplyvy, R.W. Tkach, *"Effect of fibre nonlinearity on long-distance transmission"*, J. Lightwave Technol. 9 (1991) 121–128.

[10] S.P. Singh, S. Kar, V.K. Jain, *"Effect of four-wave mixing on optimal placement of optical amplifier in WDM star networks"*, Fibre Integrated Opt. 25 (2006) 111–140.

[11] M.L. Singh, I.S.Hudiara, *"A piece wise linear solution for nonlinear SRS effect in DWDM fibre optic communica- tion systems"*, J. Microwave Optoelectron.3 (4) (2004) 29–38.

[12] M.J. O'Mahony, D. Simeonidou, A. Yu, J. Zhou, *"The design of a European optical network"*, J. Lightwave Technol. 13 (5) (1995) 817–828.

[13] M.W. Maeda, W.B. Sessa, W.I. Way, A. Yi-Yan, L. Curtis, R. Spicer, R.I. Laming, *"The effect of four wave mixing in fibers on optical frequency division multiplexed systems"*, J. Lightwave Technol.8 (9) (1990) 1402–1408.

[14] A.R. Chraplyvy, *"What is the actual capacity of single mode fibres in amplified lightwave systems"*, IEEE Photon. Technol. Lett.5z (1993) 665–668.

Management Information Security Systems in Libraries: Mathematical Approach

Dr. Mohammed Imtiaz Ahmed

Pt. Ravishankar Shukla University, Raipur, C.G. 492010, India
E-mail: imtiazexplores@gmail.com

Abstract
I have proposed a new theory of Management Information Security (MIS) System in Libraries (MISSL). MISSL is based on the directions of Diffie and Hellman's Secret Key Exchange Protocol (DHSKEP) [1] and it is represented by the formula

$$d^{k_1 k_2} = (d^{k_1})^{k_2} = (d^{k_2})^{k_1}$$

.

Keywords: MIS, MISSL, DHSKEP.

Introduction

In 1942, Ferguson & Whitelaw [2] proposed computerized management information system in libraries. Lancaster [3, 4] put the idea on the measurement and evaluation of library services in the year 1977. In the Same year Allen [5] & Anaszewicz [6] presented a new concept on managing the flow of technology and management information from integrated systems and the role of decision support system. In 1980, McClure [7] designed a technique about Information for academic library decision making. Runyon [8] extended some developments as their title "On towards the development of a library management information system" in the year 1982. Lakos [9] gave a new thought as implementing a library management information system in the year 2000. I [10] also did the work in the development of the aforesaid and wrote the thesis which entitled as "Management information systems in libraries".

In this paper, I have proposed a new theory on MISSL, which is based on DHSKEP.

Management Information System (MIS) in Library

Library can be referred as the 8th ocean of the collection of the creative information's and I have strong faith in the understanding that the Libraries are the carriers of civilization. Without books, history is silent, literature dumb, science crippled, thought and speculation at a standstill. Without libraries, the development of civilization would have been impossible; they are the engines of change. My work is little contribution in the ocean of the knowledge. I hope and believe that proper use of the modern tools/ methods of library management will help in the optimum dissemination of knowledge.

Management of the library is the basic and core activity which helps the academic community in identifying and accessing knowledge resources in a university. Living in an age of information explosion it is estimated that the amount of information in the world doubles every 20 months. Libraries, as centres of learning are experiencing unprecedented rates of change, both internally and from external sources. Therefore, Libraries have to transform themselves into organizations that support the values of quality and quality management.

This also means that libraries should build organizations that support learning. Libraries that focus on customer needs increase their ability to provide quality service to their customers. By concentrating on their ability to learn and create solutions, the learning organization "is continually enhancing its capacity to create its future".

A changing user population, technology enhancement, transformation of the scholarly communication system, digital libraries, new approaches to planning and assessment throughout the library are propelling the new environment. It has now become inevitable for Libraries to redefine their vision, mission, values, structures, and systems support behaviour that is performance and learning focused.

Today, Libraries are to be examined in terms of their planning system, financial management, buildings utilization, the state of automation, participation in cooperative activities and collection management, staffing, evaluation process, reader's services, instruction, resources, budget etc. Success of an academic library is increasingly dependent on the most effective utilization and strategic management of new technologies in libraries.

The concept of the user-centred library emerged in the late 1980s and early 1990s, fostered by strategic planning, total quality management, the external demands for accountability and measurable outcomes, and rapidly changing information and budgetary environments. Libraries must move from defining quality by the size of the inputs— and especially from valuing staff and collection size as "goods" in and of themselves. Early opinions of the potential uses of computers in libraries varied. The decision to introduce library automation was usually based on a re-appraisal of the objectives of the library to make the most effective use of existing library manpower and resources for the benefit of library users, to allow for integration of all aspects of information management relevant to the library's work now and in the future

It is believed that the versatility and power of I T which includes accommodation of increased workload, achievement

of greater efficiency in improving existing services, ability for generation of new services, facilitating cooperation and in providing for an integrated approach without regard to format, location or medium through which it is served, which can be called one - stop information shopping, can stand in good stead in the quest for quality and productivity in information services and products. Library services need to reach to the Readers with the use of the technology to provide online access to globally generated information and to provide un-interrupted world- wide access to the library resources searchable from anywhere, anytime, by anyone.

We find that global changes through the information and communication technologies (ICT), have had an impact on the functioning of academic libraries. The developments in ICT have changed the Reader's expectations from the academic libraries in many ways particularly the e-learning process. ICT holds the key to the success of modernizing information services. Not only does ICT introduce new ways of information handling, it also brings about change in the very structure of information and its communication. Concepts like universal bibliography, accessibility to and availability of documents, irrespective of location, highly personalized services matching user needs/interests with document databases, full text searches, storage and retrieval with speed and accuracy, etc. have all been accomplished to a great extent.

Integrated library system, or ILS, is another enterprise resource planning system for a library, used to track items owned, orders made, bills paid, and patrons who have borrowed. An ILS is usually comprised of a relational database, software to act on that database, and two graphical user interfaces (one for patrons, another for staff). Most ILSs separate software functions into discrete programs called modules, which are then integrated into a unified interface. Examples of modules include: acquisitions (ordering, receiving, and invoicing materials), cataloguing (classifying and indexing materials), circulation (lending materials to patrons and receiving them back), serials (tracking Journals and newspaper holdings), and the OPAC (online public interface for users). Each patron and item has a unique ID in the database that allows the ILS to track its activity.

ILSs were often known as library automation systems or automated systems in the 1970s and early 1980s. Before the advent of computers, libraries usually used a card catalogue to index its holdings. Computers were used to automate the card catalogue, thus the term automation system. Automation of the catalogue saves the labour involved in sorting the card catalogue, keeping it up-to-date with respect to the collection, etc. Other tasks automated include checking out and checking in books, generating statistics and reports, acquisitions and subscriptions, indexing journal articles and linking to them, as well as tracking interlibrary loans.

Since the late 1980s, windows and multi-tasking have allowed business functions to be integrated. Instead of opening separate applications, library staff could now use a single application with multiple functional modules.

As the Internet grew, ILS vendors offered more functionality related to the Internet. Major ILS systems now offer web-based portals where library users can log in to view their account, renew their books, and be authenticated to use online databases.

One word, ‘INTERNET’ has completely changed the way Libraries operate. Today's libraries are having a paradigm shift towards web-based e-resources. The conventional bibliographic resources are now fast supplemented by the e-resources. It is huge task for librarians to maintain a supply chain that moves shoulder to shoulder with a global information rate that doubles at every 20 months.

Management Information Systems (MIS) [Figure-1] have emerged as a solution to this capacity expansion requirement of Academic Libraries. According to McClure (1990), Management information systems are tools designed to improve management decisions. MIS is applied in libraries to track performance, monitor the results of innovation, identify problems and opportunities, evaluate alternative options, and conduct strategic planning. MIS assists library staff in daily decision making process to maintain better accountability and control of resources to monitor budget allocations, to improve overall library effectiveness, to improve long-term planning and to facilitate performance measures activities.

Generically an MIS can be defined as any reporting technique, manual or automated, which provides the key members of an organization with data used in its operation. Heim has defined a Management Information System as: 'the process and structure used by an organisation to identify, collect, evaluate, transfer, and utilise information in order to fulfil its objectives. It is a system that provides management with information to make decisions, evaluate alternatives, measure performance, and detect situations requiring corrective action'.

MIS is an interdisciplinary tool that has an amalgamation of computer sciences, information sciences, management sciences and Engineering sciences.

MIS is an information system that integrates data from all the departments it serves and provides operations and management with the information they require. MIS refers broadly to a computer-based system that provides administrators with the tools for organizing, evaluating and efficiently running their departments. Management of the library and information systems is the basic and core activity which helps the user community in identifying and accessing knowledge resources in a university.

The goal of MIS is to motivate staff to enhance their skill and expertise in conventional and e- library associated services and operations. The impact of MIS is enormous and global in its magnitude. It is now all set to become an integral part of all aspects of the library management. It has the potential to profoundly affect the library operations, information sources, services, and staff skills requirements and users expectations.

Amos Lakos, Charles R. McClure and other library professionals and educationists recognized the need for systematic application of management information systems in libraries some years ago. However, systematic application of some kind of MIS in the library environment has been and remains rare. The MIS's function is to provide library managers and staff with data, information, analysis and tools that enhance the effectiveness and efficiency of library services and assist in the decision - making process.

The objectives of an MIS are to assist library staff with the daily decision making process, to maintain better accountability and control of resources, to monitor budget allocations, to improve overall library effectiveness by focusing on outcomes to generate internal and external reports to improve long-term planning and to facilitate performance measures activities.

The four main objectives for Management Information systems have been defined as: (1) to facilitate the decision making process in the library by providing the managers with accurate, timely, and selective information that assists them in determining a specific course of action. (2) to provide for the objective performance measurement and assessment of selected relevant areas of the library. The areas are to be determined during strategic planning. (3) to provide pertinent information about the library's internal and external environments, and (4) to provide information on alternative strategies and contingency plans.

In essence, an integrated Management Information System is very important because it can be used to provide supporting information to determine: (1) Efficiency: is the library doing things right? (2) Effectiveness: is the library doing the right things? And (3) Competitiveness: is the library heading in a direction which is consistent with the environment, that is, does the library have a strategy, and is it certain that it is the correct one?

The library should strive to innovate in the use of new technologies to enhance the usefulness of the MIS in the organization. Special attention should be given to the use of the Web as integrating and enabling tool, especially for collaborative work, particularly important in a consortia environment. Libraries have used MIS to track performance, monitor the results of innovation, identify problems and opportunities, evaluate alternative options, and conduct strategic planning.

Library Automation Stands as a major prerequisite for the Management Information System. The present study aims to focus on Library Automation status for successful implementation of MIS. The library world is moving into network-based environment. Data management is the essence of such environment that enables fast transmission, retrieval and dissemination of information for better customer (reader) service. The entire process originates with management commitment and vision based leadership. Feedback from customers (readers) provides a basis for continuous improvement.

Besides the MIS operations, the most vital element for a successful MIS project is the customer focus. The evolution of MIS is to provide for better customer services with a commitment for continuous improvement through the customer's feedback on its services.

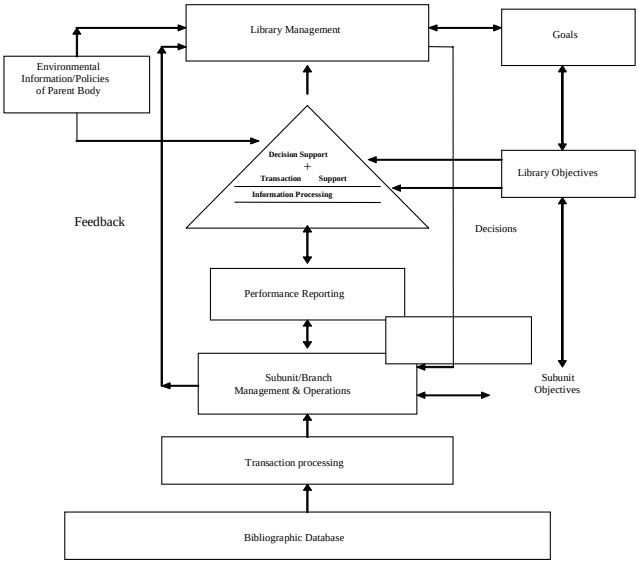


Figure – 1

In the present study, the MIS based libraries maintain customer (reader) focus as their central idea during all performances. The objective of an MIS based Library is to be an effective library through:

1. Providing convenience and justice to its readers.
2. Attract non-readers to become readers.

The five cardinal laws of Library Management, as enunciated by Dr S.R. Ranganathan (1892 – 1972), state that:

- Books are for use
- Every reader his book
- Every book its reader
- Save the time of the reader
- Library is a growing organism.

Living in the age of information explosion, MIS based libraries, powered by automation; I humbly propose (on the lines of the 36th Chamber of Shaolin) the sixth law for library management that is: "One library for all"

Maintaining this sixth law in letter and spirit allows us to further the scope of our performance to be extended through integration with all neighbouring libraries through resource sharing that again provides achievement of aforesaid library objectives of convenience and justice to readers, and, attracting non-readers to become readers. Figure 2 below exhibits the macro model of our concept of the MIS based university libraries.

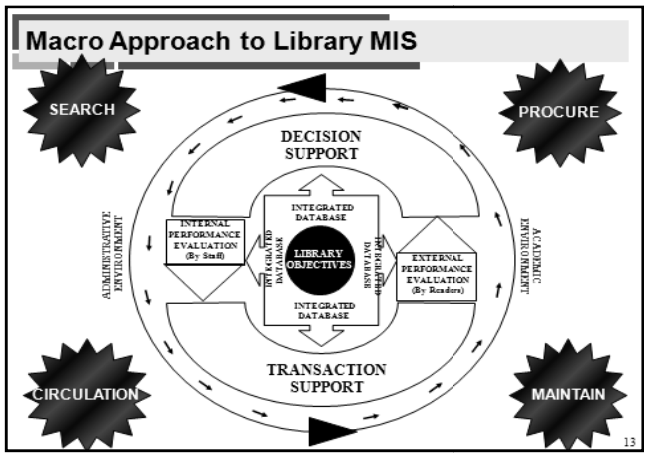


Figure – 2

Like the conventional libraries, MIS based university libraries too are surrounded with the administrative and academic environment forces of the university that govern the finance, selection of bibliographic resources, and staffing decisions of the library. As the major driving force at its core, stand the aforesaid library objectives, supported with an integrated database [Figure-3] covering the Membership Data, Bibliographic Data, Circulation Data, and the Library Maintenance Data.

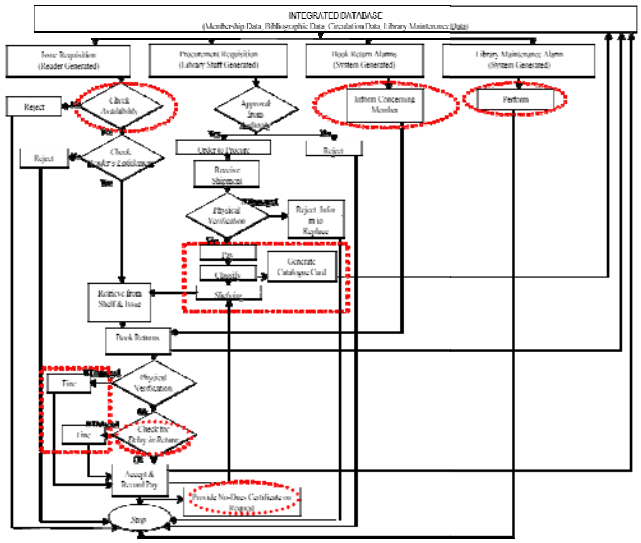


Figure – 3

But all the above works are devoted only on MIS. I propose a new theory on MISSL, behind this is a natural thought, which is inspired by protection, because I think that, without security of MIS in Library, how we will link the order of knowledge authentically. So I have used DHSKEP for all the MIS, not only for all the direct users but also this will be applied for indirect users. In the below section, I have demonstrated the proposed scheme, by the simple formulation.

Diffie-Hellman secret key exchange protocol (**D-H**) is a cryptographic protocol that allows two parties that have no prior knowledge of each other to securely agree on a shared secret key over an insecure communications channel. Then they use this key to encrypt subsequent communications using a symmetric-key cipher.

Proposed MISSL

Let,
d: Library Data,
u: User,
l: Librarian,
k_1: Secret Key of User,
k_2: Secret Key of Librarian,

By DHSKEP, User and Librarian may exchange their secret keys k_1 and k_2 respectively with the library data d as the following formulation;

$$d^{k_1 k_2} = (d^{k_1})^{k_2} = (d^{k_2})^{k_1}$$

Conclusion

Digital Crime is one of the most challenging for the civilization. We are always committed to provide the digital facilities to the users with our best, but some users misuse these facilities and share these data for the crime or non-constitutional activities. So my theory of MISSL can be very useful to prevent all the cyber crimes. MISSL means a discipline and constitution of library.

References

[1] Diffie W., Hellman M.E., New directions in cryptography, Transactions on information theory, 22, 1976, 644-654.
[2] Ferguson, S. & Whitelaw, M. Computerized management information system in libraries. Australian libraries journal. 41; 1942, pp. 184-198.
[3] Lancaster, F. W. If you want to evaluate your library… London Library Association. 1988.
[4] Lancaster, F. W. The measurement and evaluation of library services. Washington, D. C. : Information resources press. 1977.
[5] Allen, T. Managing the flow of technology: technology transfer and the dissemination of technological information. Cambridge, Ma: MIT press. 1977.
[6] Anaszewicz, R. Management information from integrated systems and the role of decision support system. Washington, Benton foundation. 1977.
[7] McClure, Charles R. Information for academic library decision making: the case for organizational information management. Westport, co.: Green Wood. 1980.
[8] Runyon, R. S. Towards the development of a library management information system. Collage and research

libraries. 42, 6; 1981, pp. 539-548.
[9] Lakos, Amos. Implementing a library management information system: update and lessons from the tri-university group of libraries experience. Proceedings of the 3rd Northumbria international conference on performance measurement in libraries and information services, Newcastle upon Tyne: department of information and library management, University of Northmbria. 2000, pp. 91-98.
[10] Ahmed, Md. Imtiaz, Management information systems in libraries, Ph.D. Thesis, Pt. Ravishankar Shukla University, Raipur, C.G., INDIA, 2009.

Biography of Author

Dr. Mohammed Imtiaz Ahmed holds M.Sc. in Chemistry, MLISc .and Ph.D. in Library and Information Science. Author is presently working at Pt. Ravishankar Shukla University, Raipur – 492010 (Chhattisgarh) India. Has 31 years of extensive experience in Library. Has 22 years of teaching experience and 8 years experience as In charge University Librarian. Author has attended several training programmes at DRTC, Bangalore, INFLIBNET Ahmadabad & IIM, Ahmedabad Has contributed extensively for various seminars, and research journals (75 papers).Presently working on MIS in Libraries.

OFDM Technique for Multi-carrier Modulation (MCM) Signaling

[1]H. Umadevi and [2]K.S. Gurumurthy

[1]Assistant Professor, Dr.AIT, Bangalore, India
E-mail: umadevi.ait@gmail.com
[2]UVCE, Bangalore University, Bangalore, India
E-mail: drksgurumurthy@gmail.com

Abstract
OFDM is novel multicarrier modulation (MCM) technique. It has strong advantage of being a generic transmission scheme whose actual characteristics can be widely customized to fulfill several requirements and constraints of an advanced communication system. It adopts wavelet packet function as carriers which have the characteristic of good orthogonality and time-frequency localization. It can be seen from both theoretical analysis and software simulation that multi-carrier modulation and demodulation technique based on wavelet packet transform has unique advantage and great potential in improving the performance of communication system.

This paper demonstrates the operation of a Wavelet Packet based multi-carrier modulation (WP-MCM) scheme. The wavelet packets are derived from multistage tree-structured paraunitary filter banks by choosing the right tree structure which would minimize the bit error between the desired and received signal for a particular channel condition. The performance of the system is simulated and analyzed for the AWGN channel. Through simulation results, we demonstrate the efficacy and the flexibility of the proposed wavelet packet based mechanism. The Bit Error rate (BER) performance is shown to be comparable, and even at times better, to conventional Fourier based OFDM. Comparison of different family of wavelets has been carried out and Meyer wavelet seems to be the most suitable wavelet through simulation results.

Keywords: OFDM, Wavelet Packet Multicarrier Modulation, AWGN, CDMA, WCDMA, Orthogonality, BER, Meyer Wavelet, SINR.

Introduction

Recently, intense interest is focused on modulation techniques which can provide broadband transmission over wireless channels for applications including wireless multimedia, wireless local loop, and future generation mobile communication systems such as CDMA, WCDMA, 3G.

While standard single carrier modulation techniques (PSK, QAM ...) take advantage of a flat (narrowband) channel, multicarrier modulation is a technique to deal with non-flat broadband channels. It splits up the channel into a large number of sub channels which all can be considered flat, so standard QAM or PSK can be used in each sub channel. Multi-carrier modulation (MCM) technology was firstly brought forward in the 1960s, which was used to modulate signals. Multi-carrier modulation (MCM) is a spectral efficient modulation scheme which transforms the single high-speed serial signal to multiple parallel low-speed signals with different carriers, and then combines these signals to one serial signal for the further transmission. By transmitting simultaneously *N* data symbols through *N* carriers the symbol rate is reduced to the one of the original symbol rate, and therefore the symbol duration is increased by N times. This leads to a transmission system which is robust against channel dispersions/fading, impulse noise and multipath interference. At the receiver port, it firstly demodulates the received signal to multiple low-speed signals with the help of the relevant carriers, and then transforms the multiple parallel low-speed signals to the high-speed original signal. The one-way symbol duration of the MCM is longer than that of the single-carrier modulation, which can effectively counteract the inter-symbol interference (ISI) and signal-to-interference-plus-noise-ration(SINR) caused by multipath transmission. MCM technique carries out the integral of numbers of symbol duration, which can effectively counteract pulse interference by dispersing effect of interference. Thereby, multi-carrier modulation technology is one effective high-speed transmission technology in wireless environment. Multicarrier modulation techniques, including orthogonal frequency division multiplex (OFDM) and wavelet packet division are among the promising techniques.

The Orthogonal Frequency Division Multiplexing (OFDM) is a MCM technique that is widely adopted and most commonly used today. In OFDM system, the modulation and demodulation can be implemented easily by means of IDFT and DFT operators. In such a system, however, the input data bits are actually truncated by **a** rectangular window and the envelope of the spectrum takes the forms of sinc (w) which create rather high sidelobes. This leads to rather high interference when the channel impairments can‘t be fully compensated. Time synchronization errors originating from misalignment of symbols at demodulator is a serious OFDM design consideration. This is because they cause Inter Symbol Interference (lSI) and Inter Carrier Interference (ICI) which severely degrade the OFDM performance. A lot of research energy has been expended to address this problem.

Wavelet transformation has recently emerged as a strong candidate for digital modulation. WPM was first proposed by Lindsey [1] in 1997 as an alternative to OFDM. The fundamental theories of OFDM and WPM have many

similarities in their way of functioning and performance but there are some significant differences which give the two systems distinctive characteristics. OFDM makes use of

Fourier bases while WPMCM uses wavelet packet bases which are generated from a class of FIR filters called paraunitary filters. OFDM signals only overlap in the frequency domain while the wavelet packet signals overlap in both, time and frequency. Due to time overlap WPM systems cannot use cyclic prefix (CP) or any kind of guard interval (GI) that is commonly used in OFDM systems. OFDM utilizes CP to overcome interference caused by dispersive channels. The greatest motivation for pursuing WPM systems lies in the freedom they provide to communication systems designers. Unlike the Fourier bases which are static sines/cosines, WPM uses wavelets which offer flexibility and adaptation that can be tailored to satisfy an engineering demand. Different wavelets result in different subcarriers leading to different transmission characteristics [2].

In this paper we investigate the BER performance degradation of OFDM and WPMCM systems. Several well-known wavelets such as Haar, Symlets, discrete Meyer and Biorthogonal wavelets are applied and studied. To simplify the analysis the channel is taken to be additive white Gaussian noise (AWGN) and perfect frequency synchronization is assumed. The paper is organized as follows: theory on OFDM and MCM are given from section II-VII. The system block of Meyer based WPMCM is outlined in Section VIII.WPMCM transmitter, AWGN channel and WPMCM receiver is outlined in section IX, X and XI respectively. Finally section XII shows results obtained by computer simulations and is followed by section XIII which concludes the paper.

OFDM and Multicarrier Modulation

Recently, a worldwide convergence has occurred for the use of *Orthogonal Frequency Division multiplexing(*OFDM) as an emerging technology for high data rates. In particular, many wireless standards (Wi-max, IEEE802.11a, LTE, DVB) have adopted the OFDM technology as a mean to increase dramatically future wireless communications. OFDM is a particular form of Multi-carrier transmission and is suited for frequency selective channels and high data rates. This technique transforms a frequency-selective wide-band channel into a group of non-selective narrowband channels, which makes it robust against large delay spreads by preserving orthogonality in the frequency domain. Moreover, the ingenious introduction of cyclic redundancy at the transmitter reduces the complexity.

Multicarrier modulation splits the broadband channel into a large number of (narrowband) subchannels. The total bitstream is divided over these subchannels. These bits are modulated per subchannel onto a subcarrier with standard narrowband modulation techniques like PSK or QAM. The sum of all the modulated subcarriers forms the composite multicarrier signal that is sent over the channel.

When the subcarriers are orthogonal, the subchannels may overlap without interfering each other, resulting in a high spectral efficiency (compared to e.g. frequency division multiplexing, where all the subchannels are separated by guard bands to prevent interference). The generation of these subcarriers is done in the digital domain, so that only one global local oscillator is needed instead of one for each subcarrier. Normally the Fourier Transform is used. An IFFT multiplexes the different mapped subcarriers to a composite signal that is modulated onto the global carrier and sent over the channel. At the receiver, the signal is demodulated by the local oscillator; the sub channels are demapped by applying an FFT to the composite signal and taking a decision in each sub channel. Because of the orthogonality of the transform, the different sub channels do not interfere.

Modulation of a subcarrier is split into two processes: mapping the bits to the constellation of the modulation technique, followed by modulation onto the one global carrier after performing the transform, on all mapped symbols. The channel is of course not ideal. The signal suffers from ISI (inter symbol interference), SINR and ICI (inter carrier interference), which comes from the loss of orthogonality due to the channel effects. To replace the Fourier Transform by a transform that is less susceptible to all these channel effects, that can easily compensate for the resulting effects is the Wavelet Transform. Its longer basis functions allow more flexibility in the design of the waveforms used, and can offer a higher degree of sidelobe suppression. This is very important, since loss of orthogonality then results in less interference. Also narrowband interferers in the channel corrupt less subchannels, so less capacity is lost when such interferers are present [3].

Wavelet Vs. Wavelet Packet

Wavelet packet Transform offers a richer signal analysis than Wavelet Transform. Wavelet packet tree allows focusing on special parts in time-frequency domain in a more detailed way than is possible with ordinary wavelet transform [4].

A wavelet packet is a generalization of wavelets in that each octave frequency band of the wavelet spectrum is further subdivided into finer frequency bands by using the two-scale relations repeatedly. The translates of each of these wavelet packets form an orthogonal basis. We can decompose a signal into many wavelet packet components.

A signal maybe represented by a selected set of wavelet packets without using every wavelet packet for a given level of resolution. The good frequency characteristics and greater flexibility offered by wavelet packet transform make it an attractive choice for a high data rate transceiver in fading channel conditions.

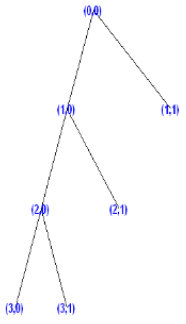


Figure 1: Wavelet tree

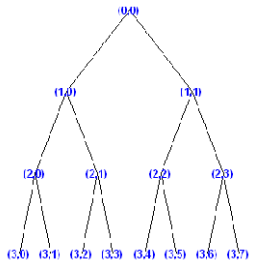


Figure 2: Wavelet packet tree

The WPT branch has a uniform frequency resolution.

Uniformity comes due to the same manner of decomposition in both low and high frequency components.

Comparing with Wavelet Transform, filter bank implementation of Discrete Wavelet Transform (DWT) performs iterative decomposition only on the low pass filter output. Thus we see non-uniformity in the frequency resolution of DWT. The output of each wavelet packet node corresponds to particular frequency band whereas outputs at wavelet packet nodes in the same level have evenly spaced frequency bands.

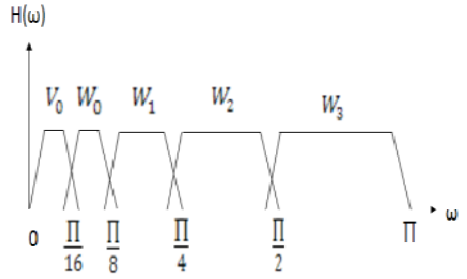


Figure 3: Frequency bands spanned by DWPT``

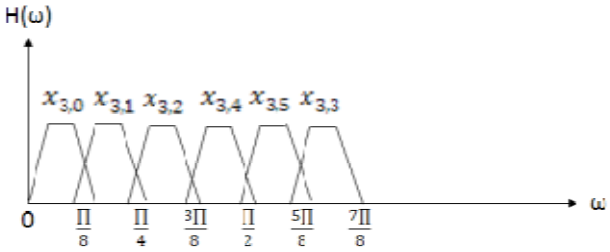


Figure 4: Frequency bands spanned by DWT

Generation of Wavelet Packet Bases

The subcarrier signal waveforms in traditional MCM implementations, such as OFDM, are sine/cosine basis functions. In WP-MCM the sub-carrier waveforms are derived from poly-channel tree structures built by cascading multiple two-channel filter banks. A two-channel filter bank consists of a set of four perfect reconstruction filters (two high pass and two low pass) which allow the decomposition and reconstruction of a signal without amplitude or phase or aliasing distortion . The two-channel filter bank has the property of splitting the signal into two lower resolution versions – namely the coarse (low pass) and the detail (high pass). When the decomposition into course and detail components is continued iteratively, it leads to the generation of wavelet packet bases. When the perfect reconstruction filters used satisfy an additional property known as paraunitary condition, they lead to wavelet packet bases with impulse responses that are mutually orthogonal to one-another and to their duals. The wavelet packet sub-carriers (to be used at the transmitter) are generated from the synthesis filters (*H'* and *G'*). The synthesis procedure at each level consists of binary interpolation (upsampling) by 2, filtering and recombination. And the wavelet packet duals (to be used at the receive analysis filters (*H* and *G*) through the analysis procedure which consists of filtering, decimation (down-sampling) by 2 and decomposition at each stage [5].

WPMCM, Wavelets and Filter Criteria

By adapting the filters one can conceive a WPMCM transceiver that best handles a system specification. The design of wavelets is bounded by multiple constraints. The constraints include properties such as orthogonality, compact support, symmetry, and smoothness and are usually stated in terms of the scaling filter $h[n]$ [6].

1) *Wavelet Existence and Compact Support*: This property ensures that the wavelet has a finite number of non-vanishing coefficients and the filters are of finite length. Wavelet existence imposes a single linear constraint on h[n]

$$\sum_{n=0}^{L-1} h[n] = \sqrt{2} \qquad (1)$$

2) *Paraunitary Condition*: The paraunitary condition is essential for many reasons. First, it is a prerequisite for generating orthonormal wavelets. Second, it automatically ensures perfect reconstruction of the decomposed signal i.e. the original signal can be reconstructed without amplitude or phase or aliasing distortion. To satisfy the paraunitary constraint the scaling filter coefficients have to be orthogonal at even shift.

$$\sum_{n=0}^{L-1} h[n]h[n-2m] = \delta[m]\begin{cases} 1 & if\ m=0 \\ 0 & otherwise \end{cases} \qquad (2)$$

3) *Flatness/K-Regularity*: This property is a rough measure of smoothness of the wavelet. The regularity condition requires that the wavelet be locally smooth and concentrated in both time and frequency domains. It is normally quantified by the number of times a wavelet is continuously differentiable. The simplest regularity condition is the *flatness* constraint which is stated on the low pass filter. A LPF is said to satisfy *K*th order flatness if its transfer function $H(\omega)$ contains *K* zeroes located at the Nyquist frequency

($\omega = \pi$). For any function $Q(\omega)$ with no poles or zeros at ($\omega = \pi$) this can be written as

$$H(\omega) = \left(\frac{1+e^{j\omega}}{2}\right)^{K} Q(\omega) \qquad (3)$$

Parameter *K* is called the regularity order and for a filter of length *L* its degree is limited by $1 \leq K \leq L/2$. In the time domain we can impose regularity condition.

$$\sum_{n=0}^{L-1} n^k (-1)^n h[n] = 0 \qquad for\ k = 0,1,\ldots,K-1 \qquad (4)$$

K-Regularity condition enforces *K* constraints on $h[n]$.

Choice of Wavelet Bases for WPMCM

The nature of the subcarrier signal waveforms greatly influence the performance of the MCM system and the wavelet basis and hence the filter pairs .In an ideal scenario the filter banks used to generate the wavelets have zero transition bands. Under such an ideal scenario the wavelet packet bases derived from a level-*i* decomposition have confined spectral footprints with bandwidth (1/2*i*) times that of the Nyquist frequency. However, available wavelet families are derived from filter banks that have a wide transition band and hence the resultant wavelet sub- carriers have a dispersed spectrum with footprints spilling into neighbouring regions. The wider the transition bandwidth the greater the dispersion of the carrier's spectral footprint and therefore the greater the difficulty in isolating those subcarriers that fall in the adjacent spectra. This greatly reduces the efficiency of the system. It is therefore important to design filter banks that have narrow transition bands. With regard to the applicability to WP-MCM systems, the desirable properties of the wavelet bases are: [5]

- They must be time-limited and smooth
- Must be well confined in frequency.
- The wavelet packet bases and their duals must be orthogonal (or at least linearly independent) to one another to enable perfect reconstruction.
- The carriers must be orthogonal (or at least linearly independent) to one another in order to have unique demodulation.
- Desirable wavelet functions have both compact support & symmetry with respect to the centre.
- Symmetric wavelet functions decay very fast.

Considering these requirements, among several available wavelets such as: Coiflets, Daubechies, Haar, Symlets, we have chosen the Meyer wavelet as it has proved to be the most suitable wavelet through simulation results that will be elaborated in further discussion. Let us see some properties of the Meyer wavelet:

- It is frequency band limited function whose Fourier transform is relatively smooth providing faster decay in time.
- It's scaling function has a compact support & is defined as

$$\phi(t) = \frac{\sin(\pi(1-\beta)t) + 4\beta t\cos(\pi(1+\beta)t)}{\pi(1-(4\beta t)^2)t} \quad (5)$$

where, ϕ (t) represents the scaling function and β represents the scaling factor. Usually β=1/3.It's wavelet function is given by

$$\psi(t) = \frac{\sin[2\pi(1-\beta)(t-1/2)+8\beta(t-1/2)\cos[2\pi(1+\beta)(t-1/2)]}{\pi\{1-[8\beta(t-1/2)]^2\}(t-1/2)} \quad (6)$$

where Ψ(t) represents the wavelet function.

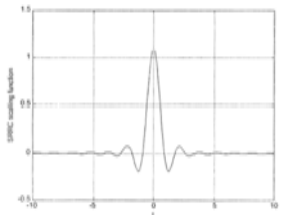


Figure 5 : Scaling function

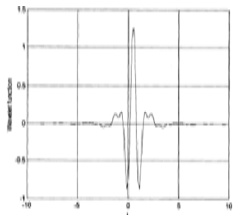


Figure 6 : Wavelet function

Perfect Reconstruction

An important issue in designing the multicarrier system is to obtain perfect reconstruction when the channel is not ideal. For perfect reconstruction in most Wavelet Packet Transform applications, the original signal can be synthesized by wavelet filter co-efficients. To actually implement the transform, one has to consider its end-effects. At the beginning and the end of a packet being sent, there is a sudden transition between no signal and signal-values creating unwanted high frequency peaks Also, implementing the Wavelet Transform comes down to performing convolutions which extend signal length (the number of non-zero signal-values). If we keep the extra values, and perform the inverse transform, we again introduce extra values at the edges. But we get perfect reconstruction and al1 the extra values at the edges are zero and can be discarded. In a way, these extra values can be seen as a small drawback and loss in performance [7].

In order to achieve the perfect reconstruction of original signal, certain wavelet conditions are to be satisfied out of which first condition says that the reconstruction is perfect and second says that it is aliasing free reconstruction. For perfect reconstruction the wavelet based transmultiplexer utilizes transmitting and receiving filter banks.

Wavelet Packet Based Modulation Scheme

The wavelet packet theory can be viewed as an extension of Fourier analysis. The basic idea of both transformations is the same: projecting an unknown signal on a set of known basis functions to obtain insights on the nature of the signal. Wavelet Packet modulation is an orthogonal multi-carrier modulation technique which is based on the wavelet packet transform. In the transmitter, a set of high speed input signals is converted into several low speed data streams by S/P (Serial to Parallel) conversion, all the sub-carriers are modulated by QAM or PSK. These wavelet packet sub-carriers (used at the transmitter end) are generated from the synthesis filters. Then by Inverse discrete wavelet packet transform (IDWPT), the wavelet packet modulation signal is obtained. At the receiver, by discrete wavelet packet transform (DWPT), QAM or PSK demodulation and P/S conversion, the original transmitted data can be obtained. And the wavelet packet duals (used at the receiver end) are obtained from the analysis filters given by equation:

$$S(n) = \sum_{u}\sum_{k=0}^{N} a_{u,k}\xi^{k}_{log_2(c)}(n-uN) \quad (7)$$

In equation (7), N denotes the number of subcarriers

while *u* and *k* are the symbol and subcarrier indices, respectively. The constellation symbol modulating *k*th subcarrier in *u*th symbol is represented as *au, k*. The sub-index log2 (*N*) denotes the levels of decomposition required to generate *N* subcarriers.

Time and frequency limited wavelet packet bases ξ (t) can be derived by iterating discrete half-band high g[n] and low-pass h[n] filters, recursively defined as:

$$\xi_{l+1}^{2p}(t) = \sqrt{2}\sum_{n} h[n]\,\xi_{l}^{p}(t - 2^{l}n)$$

$$\xi_{l+1}^{2p+1}(t) = \sqrt{2}\sum_{n} g[n]\,\xi_{l}^{p}(t - 2^{l}n) \qquad (8)$$

In equation (8), l denotes the level in the tree structure, superscript p denotes the sub-carrier index at given tree depth and h[n]and g[n] represent the low pass and high pass analysis filters respectively [6].

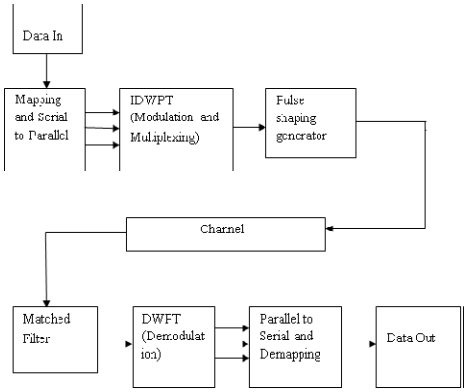


Figure 7: Block diagram of WPMCM

WPMCM Transmitter

WPMCM employs Inverse Discrete Wavelet Packet Transform (IDWPT) at the transmitter side. The IDWPT is implemented by wavelet packet synthesis filter bank which combines different parallel streams into a single signal. As shown in figure 8. the up sampling and downsampling operations by a factor 2 are represented by ↑2 and ↓2 respectively, while filter _g' stands for high-pass filter and filter _h' stands for low pass filter.

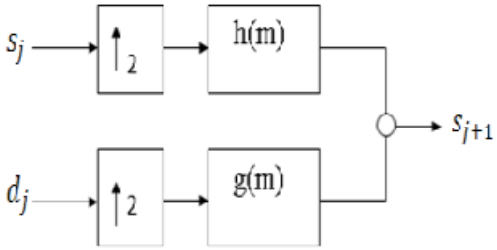


Figure 8: Synthesis or reconstruction of signal. The up arrow represents the interpolation by 2. h(m) and g(m) denote the frequency responses of the low and high pass reconstruction filters, respectively.

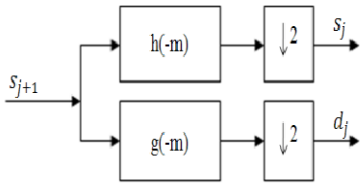


Figure 9: Analysis or decomposition of signal. The down arrow represents the downsampling by 2. h (-m) and g(-m) denote the frequency responses of the low and high pass decomposition filters, respectively

Because WPMCM transceivers are realized by an iterative method we can easily change the number of subcarriers and their bandwidth. By performing an addition alteration of two-channel filter bank at all outputs the subcarriers number is doubled or more generally the number of subcarriers is given by $N=2l$.

The subcarriers in WPMCM transceivers are completely determined by filters *H* and *G* and therefore by applying different set of filters we get different subcarriers which in turn lead to different transmission system characteristics. By just altering the filter coefficients the WPMCM transceivers are capable to achieve different values for bandwidth efficiency frequency concentration of subcarriers, low sensitivity to synchronization errors, low PAPR, etc. WPMCM signal in the discrete time domain can be expressed as:

$$X[n] = \sum_{u} a_{u,k} \sum_{k=0}^{N-1} \xi_{2\log N}^{k}(n - uN) \qquad (9)$$

Where k denotes the subcarrier index. u denotes the WPMCM symbol index. The constellation symbol modulating *k*th subcarrier in uth WPMCM symbol represented by [8].

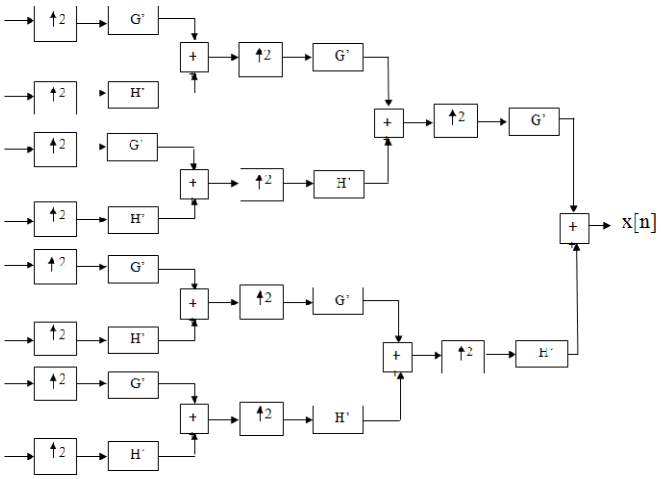


Figure 10: Synthesis using IDWPT where, G'=low pass synthesis filter and H'=high pass synthesis filter and x[n] =reconstructed signal

AWGN Channel

After implementing system and confirming that we get perfect reconstruction (with negligible round off errors), the first thing

we studied was the behaviour when simple standard AWGN is present on the channel. This is expected to give exactly the same results as in the single carrier case, since the subcarriers should not interfere with each other .The bi-orthogonality of the basis function is clearly not suited here. If everything was orthogonal as in the FFT case, AWGN would stay AWGN on each subcarrier, and subcarriers would remain orthogonal, resulting in the same performance as in the single carrier case. The fact that the basic functions used to create the different waveforms in each sub-channel are not orthogonal to each other, causes the AWGN to become correlated within the sub-channel, which is of course no longer orthogonal . Also the basis functions are not orthogonal to basis functions in other sub-channels, which also translate in some kind of interference of the AWGN between different sub-channels.

This seems to be price to pay for using bi-orthogonal basis functions. Resistance against other channel impairments, like narrowband interference, should therefore offer more performance gain in order to consider wavelets as an alternative to the standard Fourier Transform.

WPMCM Reciever

WPMCM employs Discrete Wavelet Packet Transform (DWPT) at the receiver side. The composite signal is afterwards decomposed at the receiver using wavelet packets analysis filter bank or so called DWPT. The receiver demodulates the data by employing time reverse diversion of waveforms used by the transmitter. If we assume that the WPMCM transmitter and receiver are perfectly synchronized and channel is ideal, the detected data at the receiver is shown below.

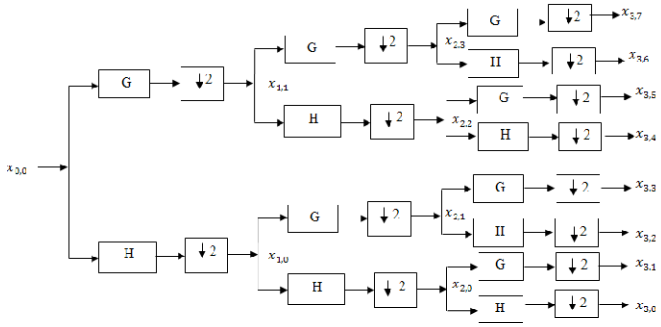


Figure 11: Recovery of data symbols using analysis filter bank. Analysis done using DWPT where G=low pass analysis filter H=high pass analysis filter

Simulation Results

The proposed system of the WPMCM system is shown in Fig.7.The simulation results of this system are obtained by using MATLAB version 7 R2009a. The comparison between WPMCM and single carrier modulation systems such as QAM, PSK and DPSK is shown in fig. 12, 13 and 14.The comparison between Meyer based WPMCM is compared with 4-OFDM, 16-OFDM and 32-OFDM over AWGN channel environment in fig.15, 16 and 17 respectively It can be clearly seen that the BER of WPMCM is lesser than that using OFDM over the same SNR.By comparing the BER for different wavelets such as Meyer wavelet, Haar wavelet, symlet wavelets, Meyer wavelet shows very less BER. we can conclude from the results shown in Fig. 18 that Meyer wavelet is most suitable wavelet than other orthogonal wavelets.

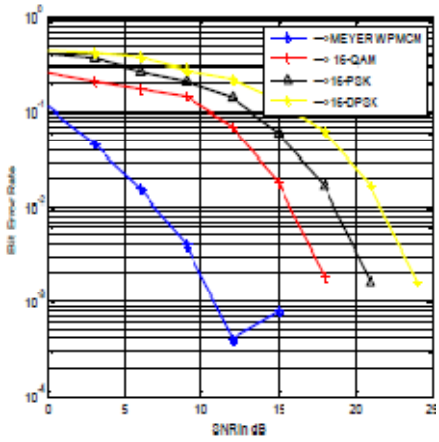


Figure 12: Bit Error Rate comparison between Meyer WPMCM (using 4- QAM), 16-QAM, 16-PSK and 16-DPSK

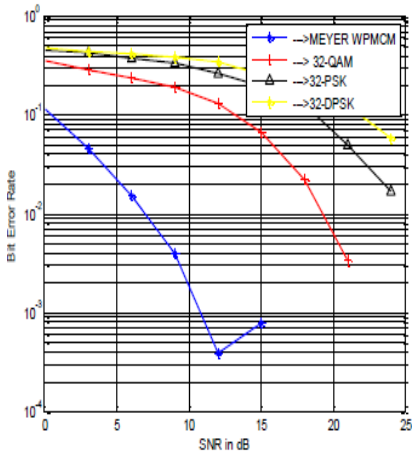


Figure 13: Bit Error Rate comparison between Meyer WPMCM (using 4- QAM), 64-QAM, 64-PSK and 64-DPSK

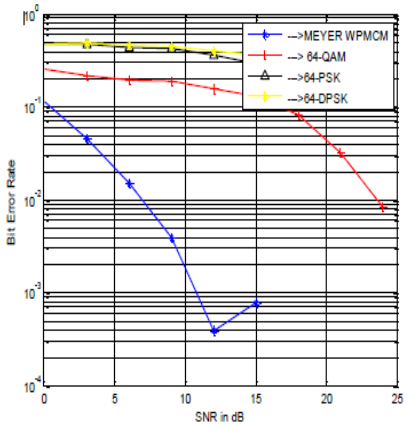


Figure 14: Rate comparison between Meyer WPMCM (using 4- QAM), 64-QAM, 64-PSK and 64-DPSK

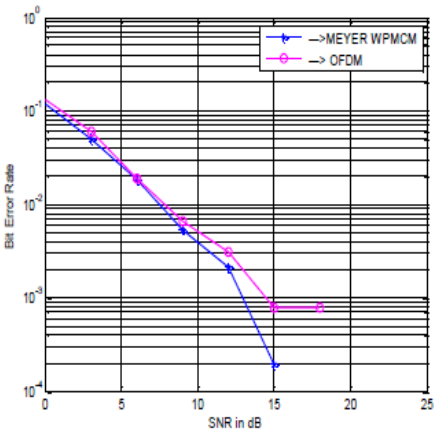


Figure 15: Bit Error Rate comparison between Meyer WPMCM (using 4- QAM) and OFDM (using 4-QAM)

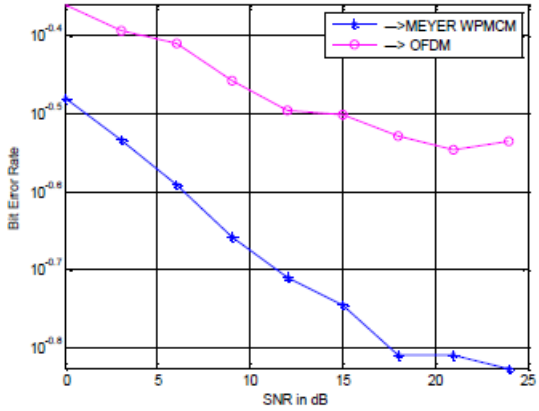


Figure 16: Bit Error Rate comparison between Meyer WPMCM (using 16- QAM) and OFDM (using 16-QAM)

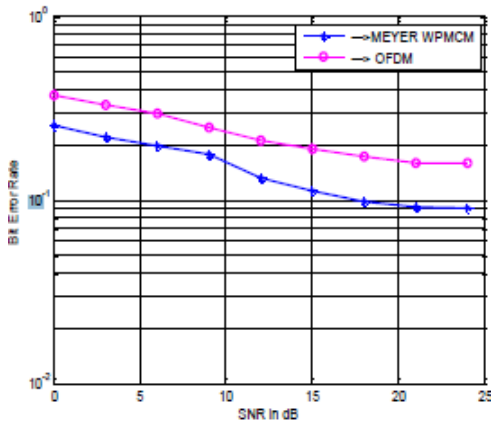


Figure 17: Bit Error Rate comparison between Meyer WPMCM (using 32- QAM) and OFDM (using 32-QAM)

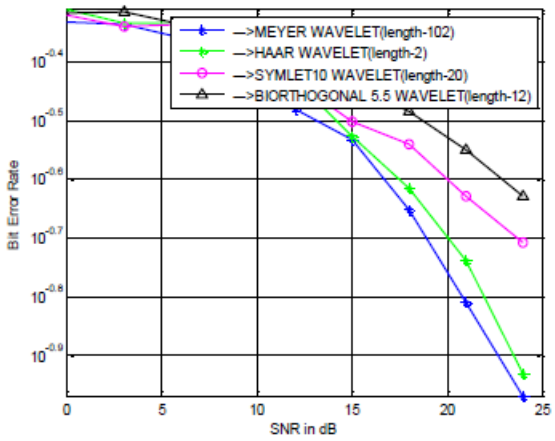


Figure 18: Bit Error Rate comparison between Meyer wavelet, Haar wavelet, Symlet wavelet and Biorthogonal wavelet (using 64-QAM)

Table II demonstrates the text transmission and reception.

Original text	**WPMCM**
SNR=12 db for WPCM. OFDM as well as 16 QAM	No of characters with errors Num=0 Rate of error, rt=0
OFDM	**16-QAM**
No of characters with errors Num=1 Rate of error, rt=0.0034	No of characters with errors Num=45 Rate of error, rt=0.1536

Conclusion and Future Scope

OFDM and WPMCM have recently emerged as strong candidates for multicarrier systems. WPMCM shares most of the characteristics of an orthogonal multi carrier system and in addition offers the advantage of flexibility and adaptation. The important points during simulation and discussion of the results are given below:

1. This paper presents an introduction to OFDM, WPMCM. WPMCM transmitter and receiver are described and the roles of main signal processing blocks are explained. After comparing WPMCM with various other transmission techniques like QAM, PSK, DPSK, OFDM, it can be inferred that WPMCM has a lower BER. Hence it proves to be a more promising modulation technique for future communication systems
2. The performance results of WPMCM lead us to conclude that this new modulation scheme is viable alternative to OFDM to be considered for today‘s communication systems. WPMCM remains nevertheless a strong competitor which offers a lot of flexibility and adaptability thanks to its simplicity and elegance.

These properties can make it suitable for the design and development of communication systems for the future (Cognitive Radio and 4G).Wavelet Packet Modulation with

adaptive filter gives very good performance even under very low SNR conditions. These wavelet packets are more immune to inter symbol interference (ISI) and inter channel interference (ICI). By altering the design specifications a wavelet based system that is more robust against synchronization errors could be developed without compromising on spectral efficiency or receiver complexity. The design and development of new wavelets which handle timing offset is in itself a separate topic for discussion and a future topic for research.

References

[1] Lindsey A.R, ―Wavelet packet modulation for orthogonally multiplexed communication, ―IEEE Trans On Signal Processing, vol.45, pp.1336-1339, 1997.

[2] Gao Xingxin, Lu Mingquan and Feng Zenming, ―Optimal Wavelet Packet Modulation, Over Multipath Wireless Channels, || 2002 International Conference on Communication, Circuits and Systems and West Sino Exposition Proceeding, Piscataway, NJ, USA, 2002, vol.1, pp 313-1317.

[3] C. Van Bouwel, J. Potemans, S. Schepers, B. Nauwelaers and A. Van de Capelle, "Wavelet Packet Based Multicarrier Modulation", Proc. IEEE Benelux Symposium on Communications and Vehicular Technology, Leuven, Belgium, 19 October 2000

[4] Sobia Baig, Fazal-ur-Rehman, M. Junaid Mughal ―Performance Comparison of DFT, Discrete Wavelet Packet and Wavelet Transforms, in an OFDM Transceiver for Multipath Fading Channel||.

[5] M.K.Lakshmanan, I. Budiarjo and H. Nikookar - Wavelet Packet Multi-carrier Modulation MIMO Based Cognitive Radio Systems with VBLAST Receiver Architecture||, *IEEE* International Research Center for Telecommunications and Radar (IRCTR), Department of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology Mekelweg 4, 2628 CD Delft, The Netherlands.

Biometric Measurement of Human Emotions

Dr. Clive Chandler and Rachel Cornes

Faculty of Computing, Engineering and Technology, Staffordshire University, Staffordshire, England
E-mail: c.chandler@staffs.ac.uk cu000775@student.staffs.ac.uk

Abstract
No aspect of our mental life is more important to the quality and meaning of our existence than emotions" (Sousa, 2010).

It is not surprising that emotions have been a point of interest for researchers for centuries (Sousa, 2010), however, even after considerable research effort over many years, there is no definitive answer to the question 'What is an emotion?' (Scherer, 2005). As computer programs and application become more complex User experience and the ability to interact successfully has become crucial, and as such the Human Computer Interaction (HCI) becomes critical to that success. One increasingly important aspect of HCI is the role played by the user's emotional state on such interactions. But as emotions are difficult at best to define, the goal is to identify a method by which they can be analysed and predicted thus enabling a possible improvement to interface interactions, and user experience. Biometric analysis offers one solution to this complex situation. Although there are many techniques utilized in biometric analysis it is imperative that whichever method is employed has the minimum effect without causing distress or interruption to the user's interaction. This paper reviews current biometric techniques and research available and suggests a system whereby human emotions could potentially be measured during software use.

Index Terms: Biometrics, Emotional Analysis, HCI, User Analysis

Introduction

Computers have become commonplace in modern life (Corrin et al., 2010) and as they increase in popularity, they also increase in complexity. As the complexity rises, software becomes more sophisticated and Human-Computer Interaction (HCI) has become a focal point - is new software going to be easy or frustrating for users to learn? Is a new computer game going to be fun to play, or scary where it is intended to be? Will a person involved in a real-time stressful job, such as a pilot or police officer, be able to cope with the stress involved in the job? Does a film have the desired effect on its viewing audience? To answer these questions, it is important to attempt to understand the user's emotional state.

Asking a select test group what they were feeling at the time could be one way to find out this information, albeit not a very accurate one as: memory can distort retrospective reports on emotions; and if the experiments are stopped as an emotion is being felt so questions can be asked, the results could be affected (Hagar & Ekman, 1983).

Are there other, more accurate, ways to find a user's emotional state? Looking at a general overview of biometrics and human emotions could help determine if there are such possibilities.

Researchers have been investigating methods whereby the process of defining the emotions felt, as they occur, can be automated. This research has been applied across many fields and in different scenarios: Human-Computer Interaction (Haag et al., 2004); measuring stress levels in Real-Time situations (Ruzanski et al., 2006); measuring user interaction with entertainment technology (Nacke et al., 2009); and eliciting the intended emotional response from users playing computer games (Callele et al., 2006).

Although Callele et al. (2006) make no attempt to utilise biometrics in their analysis of emotions others as indicated above rely on the use of biometrics to measure the emotion, as it is being felt, for example, skin conductivity, electromyography, respiration, and facial expression recognition, in particular have been used with some success.

Biometrics

Biometrics "refers to the automatic recognition of individuals based on their [unique] physiological and/or behavioural characteristics" (Jain et al., 2004)and in computing is usually used for security applications that focus on either verification (Is this person who they claim to be?) or identification (Who is this person?). Whilst in the medical profession biometrics refers to collecting the "measurements of biological and/or medical phenomena" (Vielhauer, 2006), for example, measuring a patient's blood pressure.

For the purpose of this paper, a more traditional and non-subject specific definition will be used which is "composed of the two Greek terms "bios" for life and "metros" for metric." (Vielhauer, 2006). Put simply, the measurement of life, or more specifically, the measurement of human life.

Although there is a body of research on the use of biometrics in a variety of domains, as yet the techniques have gained limited success and the question still remains as to which biometric technique could yield the most useful information for interaction design, providing as little disturbance as possible to the user whilst still measuring something as esoteric as emotions.

Given that the goal is to identify some reaction which will form the basis for the measurement of emotions whilst still leaving the user unimpeded or interrupted during their interaction with whatever software they may be using; such biometrics as blood pressure and heart rate which while

indicating a reaction would also prove too intrusive and would affect the performance of the user.

Therefore the most useful possibilities would appear to be;

- Finger Prints
- Iris
- Facial Recognition
- Skin Conductivity (GSR)
- Facial Thermography

Finger Prints

Fingerprints, which are made from ridges and valleys that are formed in the womb (Jain et al., 1999; O'Gorman, 1999; Lockie, 2009), have been used for many centuries (Jain et al., 2004), are considered to be unique between people, including identical twins, (Vielhauer, 2006) and between each finger of the same person (Jain et al., 1999).

However as they are generally used for security identification and verification it is arguable that they would not react to human emotional differences.

Iris

As with the fingerprint, the biometrics of the iris is generally used within security for verification and identification purposes. This is done in a similar way to the fingerprint. First an image of the iris is taken, and the points of interest, which in this case are made up from the "rich pattern of furrows, ridges, and pigment spots" (Bowyer et al., 2008), are checked for quality before extracting and storing this information in the form of a template. Once stored, this template can be checked against future images taken by the system to identify or verify a user. At this moment in time there appears to be no research on any changes which may or may not occur to the iris due to emotional state when measured biometrically.

Facial Recognition

Facial recognition is the measurement of key features of a face, such as "relation and size of eyes, nose and mouth" (Vielhauer, 2006) and has been used as far back as the ancient Egyptians to keep track of the workforce claiming provisions (Lockie, 2009).

As there are many disciplines interested in facial recognition, there are several implementations on the market, for example Eigenface, Feature Analysis, Neural Network, and Automatic Face Processing (Nanavati et al., 2002), that work using 2-D images, 3-D images, colour images, infra-red images, or a combination of them (Weng & Swets, 1998). Most systems will need a still image to compare to an image previously stored, such as a driving licence photograph, but there are some systems that work using a time-varying image sequence (video). This makes it possible to track faces and recognise facial expressions (Weng & Swets, 1998).

More recent advancements in Facial Reading applications have enabled real time measurement of critical facial nodes which indicate emotional state nominally from MIT media Lab (2011) and Noldus (2011).

Skin Conductivity (GSR)

The epidermis layer of human skin is highly resistive while the dermis and subcutaneous tissue is highly conductive (Giakoumis et al., 2010). As a human feels emotionally aroused, sweat glands open, making the skin more conductive (Healey, 2008) by creating a pathway from the skin's surface to the more conductive skin below (Giakoumis et al., 2010). Measuring the conductivity of skin is referred to as measuring the Galvanic Skin Response (GSR), or skin conductivity.

Measuring skin conductivity is utilized in several applications: lie detectors (Lykken, 1959); measuring negative emotions, such as boredom (Giakoumis et al., 2010); and measuring interest (arousal) in a phone conversation (Iwasaki et al., 2010).

Facial Thermography

Each person emits a heat pattern that is characteristic to that person so that it can be used as a biometric measurement (Jain et al., 2004). Facial thermography is a non-intrusive method for identifying a person from their characteristic facial heat emissions that has been found to have accuracy levels reaching 99.25% (Bhowmik et al., 2010). Face, hand and hand vein thermography are all used in biometrics (Jain et al., 2004).

Biometric Measurement of Emotions

While theorists do not agree on what constitutes the elements of emotion, there are measurements indicating that when a human has an emotion, there is also a physical reaction taking place; for example, when frightened a human being's heart rate increases, muscles tense, and palms become sweaty (Haag et al., 2004). It has been proposed that these reactions can be measured to determine exactly which emotion is being experienced and to what intensity (Schut et al., 2010; Pantic et al., 2007; Amershi et al., 2006; Haag et al., 2004). As can be seen from these papers, there are different methods to measure emotions: Skin Conductivity (SC), or Galvanic Skin Response (GSR); electromyography (EMG) of the facial muscles; respiration (RESP); and facial expression recognition being among them. The most commonly explored of these being facial expression recognition (Haag et al., 2004).

Skin Conductivity (GSR)

Conductivity of the skin has been used to measure emotions in several research projects over recent years (Haag et al., 2004; Giakoumis et al., 2010; Khalfa et al., 2002; Iwasaki et al., 2010). Sensors are applied to the fingers of the non-dominant hand, as shown in

Figure **1**, registering when the skin conductivity changes, or when an emotion takes place.

Skin conductivity is one of the bio-sensory methods used as the physiological change occurs as an emotion takes place (Healey, 2008). In and of itself the measurement of skin conductivity is not enough to map emotions felt, as emotional responses vary between different people and as one response could be mapped back to more than one emotion, so it is usual to utilize GSR techniques in conjunction with another measurements such as: pupil dilation or heart rate (Tapus & Mataric', 2007); electromyography, blood-volume pressure, electrocardiography, and respiration (Haag et al., 2004); although, this is not always the case (Khalfa et al., 2002).

Figure 1: Skin Conductivity Sensors (Haag et al., 2004)

Even though Khalfa et al. (2002) relied solely on skin conductivity in their research, they suggest in their conclusion that skin conductivity is better used to measure emotional arousal (calm to excited) rather than valence (negative to positive). They do, however, state that this is a good method to use in order to differentiate between conflict/non-conflict situations, such as anger and fear. One of the problems that Khalfa et al. (2002) identify with using skin conductivity is the influence of external factors such as temperature. This can be countered by calibration beforehand.

Electromyography

This technique measures the frequency of muscle tension of a specific muscle. This can be carried out in one of two ways: intra-muscular EMG or surface EMG. Intra-muscular EMG measures the frequency of muscle tension by inserting small needles into the muscle to be measured and subsequently getting the participant (usually a patient) to tense that muscle (eMedicineHealth, 2011). Surface EMG is measured by placing a sensor on the skin above the muscle to be measured, this is the method used by Haag et al. (2004) and Healey & Picard (1998), and is shown below

Figure 2: EMG sensor applied to the masseter muscle (Haag et al., 2004)

As with skin conductivity, EGM is used in conjunction with other techniques; skin conductivity; blood volume pressure; and respiration (Haag et al., 2004; Healey & Picard, 1998). One disadvantage to using EGM on facial muscles is that the results may be altered due to, for example, talking and coughing (Mandryk & Atkins, 2007).

Facial Expression Reading

Facial expressions are the most expressive way in which humans portray their emotions (Sarode & Bhatia, 2010); it is, therefore, not surprising that facial expression recognition is one of the more researched methods of ascertaining human emotion (Haag et al., 2004). There are three approaches to facial expression recognition: analytic, holistic and hybrid.

Sarode & Bhatia (2010), among others (Lu et al., 2008), take an analytic approach and determine facial expressions by vision-based facial gesture analysis – taking the parameters of the face in a neutral state, shown in Figure 3 as a base line to work from, then all future facial parameters taken for that person are compared to the initial measurements.

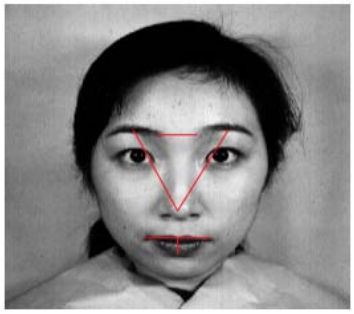


Figure 3: Parameters of the Face (Sarode & Bhatia, 2010)

Their study focused on recognising four of the six emotional states as laid out by Ekman (1993). It has been observed that if the edges of the mouth are above the original position then the emotion can be classified as ‘happy’; if the centre of the mouth is raised and the centre of the eyebrows are raised simultaneously, then that will indicate the emotion can be classified as ‘sad’; if the eyebrows are raised and the mouth is lowered then the emotion indicates ‘surprise’; and if the centre of the eyebrows are lowered, the edges of the mouth are closer together and the nose is lowered then the emotion would be interpreted as ‘anger’. This is illustrated in figure 4 below:

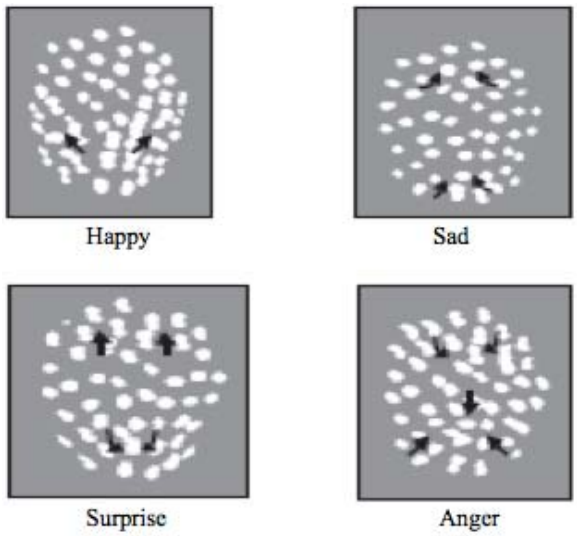


Figure 4: The Cues of Facial Expressions (Sarode & Bhatia, 2010)

This technique has been shown, Figure, to be fairly accurate for these four emotions with 'happy' and 'surprise' having an accuracy level of approximately 83%, with 'sad' and 'angry' being slightly less accurate at only 78% .

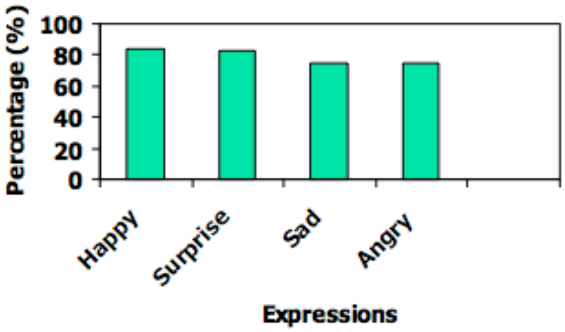


Figure 5: Accuracy of Facial Expression Recognition (Sarode & Bhatia, 2010)

Unlike the analytical approach, the holistic approach does not focus on individual features of the face; instead the face is examined as a whole. This technique is used in many research projects, for example (Lai et al., 2001; Etemad & Chellappa, 1997). These projects have resulted in a higher accuracy than Sarode & Bhatia (2010) with the levels at 95.56% and 99.2%, respectively. Part of the problem that researchers face is that the appearance of a face can change, e.g. with the addition of glasses or a different hair style, and that when acting naturally, a human will change the orientation of their face making it difficult, but not impossible (Lai et al., 2001), to make a comparison with the original neutral measurements taken.

Hybrid techniques are also abundant in research (Pantic & Rothkrantz, 2004; Ahonen et al., 2006); but, while a hybrid approach would be more robust against head orientation, a holistic approach is computationally less complex as a feature set does not have to search for and locate feature sets (Etemad & Chellappa, 1997).

One drawback with just using facial expression recognition is that a fraudulent emotion will still get recognised as being real with an 85% accuracy level, as in the case with an actor or actress portraying emotions (Busso et al., 2004). This can be overcome by using another technique alongside facial expression recognition, such as skin conductivity or facial thermography.

Biometric Emotional Analysis System

Given the previous investigation into suitable setups and analysis techniques at Staffordshire University a system has been developed using the latest face reading software from Noldus and a wireless GSR device developed at MIT Media labs in the USA, the combination of these two biometrics offer the possibility of measuring the responses of participants to use of software packages and may well lead into better guidelines for the design of effective user interfaces.

One project is in the area of Games design and it is hoped the system can be used to develop a Fear Index to aid in the consumer information offered to parents. The system is also being developed as a mobile setup for use in areas where the user works and thus having a more realistic analysis.
Initial results seem hopeful.

References

[1] Ahonen, T., Hadid, A. & Pietikainen, M., 2006. Face Description with Local Binary Patterns: Application to Face Recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 28(12), pp.2037-41.

[2] Amershi, S., Conati, C. & Maclaren, H., 2006. Using Feature Selection and Unsupervised Clustering to Identify Affective Expressions in Educational Games. In *Workshop on Motivational and Affective Issues in ITS.*., 2006. 8th International Conference on ITS.

[3] Bhowmik, M.K. et al., 2010. Human Face Recognition using Line Features. *CoRR*, abs/1007.0638.

[4] Bowyer, K.W., Hollingsworth, K. & Flynn, P.J., 2008. Image Understanding for Iris Biometrics: A Survey. *ScienceDirect: Computer Vision and Image Understanding*, 110, pp.281-307.

[5] Busso, C. et al., 2004. Analysis of Emotion Recognition Using Facial Expressions, Speech, and Multimodal Information. In *Multimodal Interfaces.* PA, 2004. ACM.

[6] Callele, D., Neufeld, E. & Schneider, K., 2006. Emotional Requirements in Video Games. In *International Requirements Engineering Conference.*, 2006. IEEE Computer Society

[7] Corrin, L., Bennett, S. & Lockyer, L., 2010. Digital Natives: Everyday Life Versus Academic Study. In *International Conference on Networked Learning.*, 2010. Springer.

[8] Ekman, P., 1993. Facial Expression and Emotion. *American Psychologist*, 48(4), pp.376-79.

[9] eMedicineHealth, 2011. *Electromyography (EMG).* [Online] Available at: HYPERLINK "http://www.emedicinehealth.com/electromyography_emg/page3_em.htm" \l "EMG Preparation" http://www.emedicinehealth.com/electromyography_emg/page3_em.htm#EMG_Preparation [Accessed 14 July 2011].

[10] Etemad, K. & Chellappa, R., 1997. Discriminant Analysis for Recognition of Human Face Images. *Journal of Optical Society of America*, 14(8), pp.1724-33.

[11] Giakoumis, D. et al., 2010. Identifying Psychophysiological Correlates of Boredom and Negative Mood Induced During HCI. In *1st International Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications.* Valencia, 2010. INSTICC Press.

[12] Hagar, J.C. & Ekman, P., 1983. The Inner and Outer Meanings of Facial Expressions. In Cacioppo, J.T. & Petty, R.E. *Social Psychophysiology: A Sourcebook.* New York, USA: The Guilford Press. p.Chapter 10.

[13] Haag, A., Goronzy, S., Schaich, P. & Williams, J., 2004. Emotion Recognition Using Bio-Sensors: First

Steps Towards an Automatic System. *Lecture Notes in Computer Science*, 3068, pp.36-48.

[14] Healey, J.A., 2008. *Sensing Affective Experience*. Springer.

[15] Healey, J. & Picard, R., 1998. Digital Processing of Affective Signals. *Proceedings of ICASSP*.

[16] Iwasaki, K., Miyaki, T. & Rekimoto, J., 2010. AffectPhone: A Handset Device to Present User's Emotional State with Warmth/Coolness. In *Proceedings of the 1st International Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications*., 2010. INSTICC PRESS.

[17] Jain, A.K., Bolle, R. & Pankanti, S., 1999. Introduction to Biometrics. In *Biometrics: Personal Identification in Networked Society*. 3rd ed. USA: Kluwer Academic Publishers. pp.1-42.

[18] Jain, A.K., Ross, A. & Prabhakar, S., 2004. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology.*, 14(1).

[19] Khalfa, S., Isabelle, P., Jean-Pierre, B. & Manon, R., 2002. Event-Related Skin Conductance Responses to Musical Emotions in Humans. *Neuroscience Letters*, 328, pp.145-49.

[20] Lockie, M., 2009. *Biometric Technology*. 2nd ed. England: Heinemann.

[21] Lu, G., Li, X. & Li, H., 2008. Facial Expression Recognition for Neonatal Pain Assessment. In *Neural Networks & Signal Processing*. China, 2008. IEEE.

[22] Lykken, D.T., 1959. The GSR in the Detection of Guilt. *Journal of Applied Psychology*, 43(6), pp.385-88.

[23] Mandryk, R.L. & Atkins, M.S., 2007. A Fuzzy Phsiological Approach for Continuously Modeling Emotion During Interaction with Play Technologies. *International Journal of Human-Computer Studies*, 65, pp.329-47.

[24] MIT Media 2011, http://www.affectiva.com/q-sensor/ (accessed 15 July 2011)

[25] Nacke, L.E. et al., 2009. Playability and Player Experience Research. In *Breaking New Ground: Innovation in Games, Play, Practice and Theory*., 2009. Digital Games Research Association (DiGRA).

[26] Nanavati, S., Thieme, M. & Nanavati, R., 2002. *Biometrics: Identity Verification in a Networked World*. 1st ed. John Wiley & Sons, Inc

[27] Noldus 2011, Face Reader software, http://www.noldus.com/human-behavior-research/products/facereade (Accessed 15 July 2011)

[28] O'Gorman, L., 1999. Fingerprint Verification. In *Biometrics: Personal Identification in Networked Society*. 3rd ed. USA: Kluwer Academic Publishers. pp.43-64.

[29] O'Gorman, L., 1999. *Fingerprint Recognition in Biometrics*. Springer US.

[30] Pantic, M., Pentland, A., Nijholt, A. & Hunag, T.S., 2007. Human Computing and Machine Understanding of Human Behaviour: A Survey. *Human Computing*, pp.47-71.

[31] Pantic, M. & Rothkrantz, L.J.M., 2004. Facial Action Recognition for Facial Expression Analysis from Static Face Images. *Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 34(3), pp.1449 - 1461.

[32] Ruzanski, E. et al., 2006. Stress Level Classification of Speech Using Euclidean Distance Metrics in a Novel Hybrid Multi-Dimensional Feature Space. In *ICASSP*. USA, 2006. IEEE.

[33] Sarode, N. & Bhatia, S., 2010. Facial Expression Recognition. *International Journal on Computer Science and Engineering*, 2(5), pp.1552-57.

[34] Scherer, K.R., 2005. What Are Emotions? And How Can They Be Measured? *Science Social Information*, 44(4), pp.695-729.

[35] Sousa, R.d., 2010. Emotion. In E.N. Zalta, ed. *The Standford Encyclopedoa of Philosophy*. 2010th ed. http://plato.stanford.edu/archives/spr2010/entries/emotion

[36] Tapus, A. & Mataric', M.J., 2007. *Socially Assistive Robots: The Link between Personality, Empathy, Physiological Signals, and Task Performance*. Los Angeles,: Association for the Advancement of Artificial Intelligence University of Southern California.

[37] Vielhauer, C., 2006. *Biometric user authentication for IT security: from fundamentals to handwriting*. 1st ed. New York, USA: Springer Science+Business Media, Inc.

[38] Wang, Y., Agrafioti, F., Hatzinakos, D. & Plataniotis, K.N., 2008. Analysis on Human Electrocardiogram for Biometric Recognition. *Journal on Advances in Signal Processing*, 2008, pp.1-12.

Enhanced Personalization and Customization Approach for Emerging Marital Market

Anand Singh Rajawat[1], Upendra Dwivedi[2] and Dr. Akhilesh R. Upadhyay[3]

[1&2]JJT University, Jhunjhunu, Rajasthan, India
[3]Professor and Head, Dept. of Communication Engg., SIRT, Bhopal, India.
E-mail: [1] rajawat_iet@yahoo.in, [2]ud1985@gmail.com, [3]akhileshupadhyay@yahoo.com

Abstract
The online matrimonial market is growing rapidly, thousands of matches on matrimonial portals has been happening every single minutes. The major advantages of online matrimonial markets are the ability to reach a large number of customers of matrimonial websites at very low costs, to avail desired online information, to perform profile matching on basis of input given by customers. Nowadays, by using intelligent programs required information can be checked, processed and carry out profile matches more quickly. This research enables the Indian online marital market using personalized and customization approach for improved one-to-one interaction and transactions. The core component used to generate the personalized web pages is the item based collaborative filtering recommendation replica. The benefits enabled by the research will be for the organization in India, how to put together the mechanism to renovate itself to the digital market and gain competitive advantage by using electronic trade technology especially about the matrimonial market.

Keywords: Collaborative Flittering, Personalization, Customization, Intelligent System, Web-mining.

Introduction

Web personalization refers to compose or modify according to requirement so that it must be suitable for a particular person's or organization's needs. Personalization is defined as the ability to provide content and services adapted to individuals based on knowledge about their preferences and behavior. Web personalization is about personalizing aspects of web resources - the content itself, links, web page structure and navigation. Customization takes place when users are able to modify a web site's look and feel for example, after registering with the Excite and Yahoo! sites, users can create their own customized start pages by choosing their preferred layout, content, and color scheme. Customer's information is obtained through the registration process, such as customer names. It is also used to create personalized greetings within the customized start pages. Thus, these sites combine customization and personalization features to provide users with the information they need, quickly and easily customization and personalization features. Both personalization and customization are powerful tools in the battle for customer loyalty. Marriages, it is said, are made in heaven. For many Indians, they are now increasingly being made on the Internet through matrimonial portals. Although still a fledgling industry, online matrimonial matchmaking, a uniquely Indian phenomenon, is seen by many to be brimming with potential. The reason the sheer reach and convenience that the internet provides as more and more people go online, they find that the medium lends itself very well to matchmaking because it takes away geographical limitations and is more efficient and more effective than the traditional avenues.

One of the most promising such technologies is collaborative flittering [1, 2]. Collaborative filtering works by building a database of preferences for items by users. A new user is matched against the database to discover neighbors, which are other users who have historically had similar taste to items that the neighbors like, then recommended to new user as he will probably also like them. Collaborative filtering has been very successful in both research and practice and also in both information filtering applications and e-commerce applications. For the online consumer decision-making process the goal of marketing research efforts is to understand the consumers online decision making and formulate an appropriate strategy to influence their behavior. E-commerce offers companies the opportunity to build one-to-one relationships with customers that are not possible in other marketing systems product. Customization, personalized services, and getting the customer involved interactively (e.g., in shaadi.com, order tracking, and so on) are all practical in cyberspace. In this paper, we address the profile matching and the matrimonial market, the collaborative filtering which is a personalization method that uses customer data to predict, based on formulas derived from behavioral sciences. The method and formulas item-based collaborative filtering used to execute collaborative filtering the two stages are prediction and recommendation [3]. The researchers develop a web and design it to be two-side personal online web pages which marriage portal can interact with a company and a company can do so. Each can learn about the other side about the products or services such as cast of the bride and groom or profile matching and the couple or companies in real time get customized products or services.

Online Matrimonial market

The industry's growth and the optimism of the players are fuelled by multiple factors. Take India's demographics. It is estimated that there are around 450 million people in India

currently below the age of 21. On the socio-cultural front, the dominant tradition is that of arranged marriages, where the parents or family elders find a suitable match for the young adults. Traditionally this has been done through contacts via family and friends, individual marriage brokers, marriage bureaus and classified advertisements in newspapers. Matrimonial portals are a fairly recent channel. match the demographics and the tradition of arranged marriages and there is clearly a huge market for match-making whatever the medium. With its reach, convenience and relative privacy, the internet provides a superior alternative to any other medium. Website personalization and customization offer internet users a sense of familiarity being an interactive network. Advantages of the online matrimonial market over the traditional one are listed in table.1.

Table 1: Online Matrimonial market Character.

Cost	Can be very inexpensive
Life cycle	Long
Place	Global
Context updating	Fast, simple, inexpensive
Space for details	Large
Ease of search by partner	Quick and easy
Reliability	High
Communication speed	Fast
Ability of profile matching	Easy, fast

The Internet spawns a diversity of personalities in the online community. These personalities show uniqueness by the way they use the Internet, and also by the various features and applications on websites that are continuously being added and modified. Examples of evolving Internet features are personalization and customization, which are website applications that make each users visit personalized to information known about him from his online identity. Users need to simply log on to a matrimonial portal and upload their profiles, sharing as much or as little information as they choose. they can then search for partners according to their individual preferences. They also have the option of exploring the medium by registering without any charge and then becoming paid users only if they see value in the portal by way of ease of use and suitable responses all this at the click of a mouse. organizational behavior and human resource management at the IIM, Bangalore, points out that the increasing mobility of younger professionals and the breaking down of traditional family networks are also responsible for driving the traffic on matrimonial portals. "Today's young adults see this as the cool, new-generation medium, one that puts them in control of choosing their life partners and at the pace that they want," he says. The bulk of the players' revenues come from subscription fees. On offer are various membership plans that differ according to the length of time a profile is posted, and features like level of personalization, special highlighting of the profile, access to verified phone numbers etc.

WEB Personalization Techniques

According to [1] and [4], web personalization techniques are classified in seven classes. The handicraft decision technique, Hyperlink-based technique, content based filtering, and traditional collaborative filtering, model based techniques, hybrid techniques and semantic techniques. The handicraft decision technique means that website managers establish decision rules according to the statistics of users or session history. To take advantage of these rules, the Journal of Theoretical and Applied Information Technology ,recommendation system provides particular contents and web structures to particular sorts of users. This kind of system functions easily, but its efficiency is low and it is difficult to renew in a timely fashion. The hyperlink-based technique generally uses an algorithm related to diagram theory to discover the most representative elements provided by the user input or information request. Search engines mostly use this technique. The famous goggle search engine is one notable example. Content based filtering uses an individual approach which relies on user's ratings and item descriptions. Items having similar properties as items positively rated by user are being recommended to the user. The most common problem of content based filtering is the new user problem. This problem occurs when a new user is added to the system, hence has an empty profile (without ratings) and cannot receive recommendations. Traditional collaborative filtering uses ratings from user's neighborhood. Neighbors are users who provided similar ratings for same items. Item is being recommended to the user according to the overall rating of the neighbourhood for that item. Problems in collaborative filtering occur when new content item is added to the system, because the item cannot take place in personalization without being rated before. Model based techniques represent an improvement in scalability issues, because part of data is pre-processed and stored as model, which is used in the personalization process. Hybrid personalization techniques combine two or more personalization techniques to improve the personalization process. In most cases, content based filtering is combined with traditional collaborative filtering. Collaboration via content is an example of a hybrid personalization technique, where user profiles contain item descriptions based on similarity of user's. Traditional personalization techniques can provide very suitable solution for couture web pages according to user's preferences. On the other hand, traditional web personalization has limitations in accuracy of modeling user's behavior In this paper

Customize Online Matrimonial Portal Technique

We proposed below are the tips you can do it yourself immediately to personalize the matrimonial website. all these stuff are basic five steps. Start with your logo, Take the smart domain name, personalize it with your concept, get a new home page banner, implement the banner code in the site, Start with your logo you can very easily customize the website from admin by changing the logo. This would give you whole new look to the website. take the smart domain name A good domain is the most important step to personalization of the matrimonial website. It enhances the search engine rankings as well, but do not be excessive dependent on the same. find a

small, keyword rich, easy to remember domain name as this also increases the recall value of the website. Personalize it with your concept personalize the site with text on home page, logo in the email design, packages that suits your market , few pages like “about us”, “contact us”, & “terms”. This gives a site a more credibility in front of users. Mentioning a phone number, address also adds more to the website. also customize the matrimonial directory section & article section with some viewable stuff. Also customize the home page meta tags this instantly helps in search engine & keyword enhancements on the site or any third party web stats service code so that you can get insight of the service. Further from the admin section of the website you can enable or disable or change defaults as per you’re requirement of your matrimonial place Get a new home page banner home page thing picture occupies significant space and this can make a turnaround the whole look of the website. the benefits a user will get on the registration. Implement the banner code in the site next thing to do is implement the third party banner code,this starts propagating the fresh, audience targeted content on the website & gives whole new look& immediately starts generating revenue with the traffic coming. We had derived this list so that new business planning to launch a online matrimonial website business, could estimate the work involved.

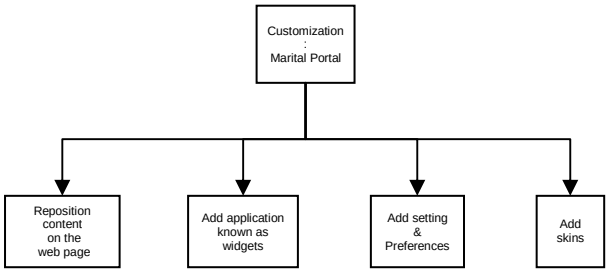


Figure 1: Types of marital customization.

Currently websites offer a variety of customization methods. Reposition content on the page - Boxes containing content can be moved anywhere on the page (or even removed). Refer the Redbridge Council homepage for an example. Add applications known as 'widgets' - These are small applications often built in HTML and JavaScript that can be used to display content feeds (such as RSS) or perform more advanced functions. The most popular widgets on Google's customizable web page include a simple clock, a local weather summary, a daily horoscope and a Wikipedia search. Add settings and preferences - Examples include setting how many news headlines are shown and setting your location to get relevant weather reports. Add 'skins' - These can be used to change the overall appearance of the web page, including its color scheme. Provide a reset button BBC homepage beta can click a button to reset their page to its original configuration. Give users the option to lock their configuration so that content can't be moved or removed by accident. Make it simple to arrange content.

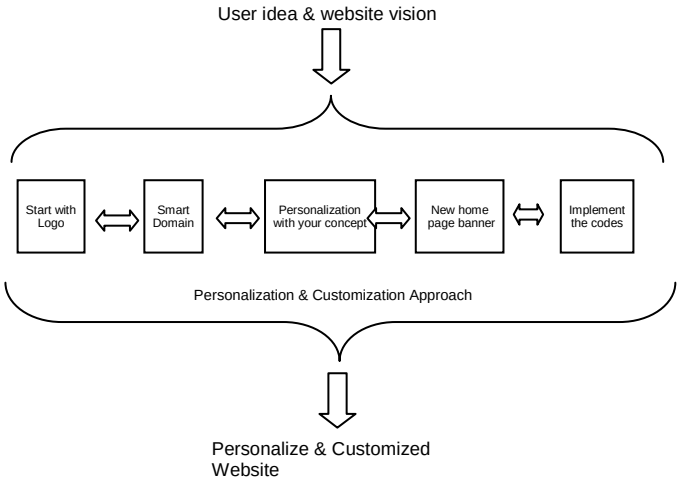


Figure 2: Personalization and Customization of Website

On MSN's homepage page content can't easily be moved from one column to another. Yahoo! and iGoogle's drag and drop method is much easier to use. The iGoogle’s content can be rearranged easily by simply dragging it around the page. Consider using an open application platform to make it easy for developers to migrate existing applications to your website.

Benefits of Adding Customization to Your Web Site, you could attract new users and keep existing users more engaged with your site by adding customization. This is due to three factors: Personalization, Choice & Prioritization and Entertainment. Making your website customizable by users could provide benefits for both you and your site visitors. However, before investing resources in developing such features carefully consider whether customization is appropriate for both your website and your users.

Collaborative Filtering

Recommended systems apply data analysis techniques to the problem of helping users find the items they would like to purchase at e-commerce sites by producing a predicted likeliness score or a list of top-N recommended items for a given user. Item recommendations can be made using different methods. Recommendations can be based on demographics of the uses, overall top selling items, or past buying habit of users as a predictor of future items. Collaborative Filtering (CF) [3,7] is the most successful recommendation technique to data. The basic idea of CF-based algorithms is to provide item recommendations or predictions based on the opinions of other like-minded users. The opinions of users can be obtained explicitly from the users or by using some implicit measures. CF algorithms represent the entire m x n user-item data as a ratings matrix, A. Each entry ai,j in A represents the preference score (ratings) of the ith user on the jth item. Each individual ratings is within a numerical scale and it can as well be 0 indicating that the user has not yet rated that item.CF approaches assume that those who agreed in the past tend to agree again in the future. For example, a collaborative filtering or recommendation system for music tastes could make

predictions about which music a user should like given a partial list of that user's tastes (likes or dislikes) [5]. CF methods have two important steps, (1) CF collects taste information from many users, and this is collaborating phase. (2) Using information gleaned from many users predictions and recommendation of users interest were automatically generated, and this is filtering phase. Researchers have devised a number of collaborative filtering algorithms that can be divided into two main categories, User-based and Item-based algorithms [6]

Item-based Collaborative Filtering Algorithms the item-based approach looks into the set of items the target user has rated and computes how similar they are to the target item i and then selects k most similar items $\{i_1, i_2 \ldots, i_k\}$. At the same time their corresponding similarities $\{si_1, si_2 \ldots, si_k\}$are also computed. Once the most similar items are found, the prediction is then computed by taking a weighted average of the target user's ratings on these similar items. We describe these two aspects, namely, the similarity computation and the prediction generation in details here. One critical step in the item-based collaborative filtering algorithm is to compute the similarity between items and then to select the most similar items. The basic idea in similarity computation between two items i and j is to first isolate the users who have rated both of these items and then to apply a similarity computation technique to determine the similarity sij here the matrix rows represent users and the columns represent items. There are a number of different ways to compute the similarity between items. Here, we present three such methods. These are cosine-based similarity, correlation based similarity and adjusted-cosine similarity.

Cosine-based Similarity: In this case, two items are thought of as two vectors in the m dimensional userspace.The similarity between them is measured by computing the cosine of the angle between these two vectors. Formally, in the m×n ratings matrix, similarity between items i and j, denoted by

Sim (i,j)is given by

$$sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|_2 * \|\vec{j}\|_2}$$

Where "." denotes the dot-product of the two vectors.

Correlation-based Similarity: In this case, similarity between two items i and j is measured by computing the Pearson-r correlation. To make the correlation computation accurate we must first isolate the co-rated cases (i.e., cases where the users rated both i and j). Let the set of users who both rated i and j are denoted by U then the correlation similarity is given by

$$sim(i,j) = \frac{\sum_{u\in U}(R_{u,i}-\bar{R}_i)(R_{u,j}-\bar{R}_j)}{\sqrt{\sum_{u\in U}(R_{u,i}-\bar{R}_i)^2}\sqrt{\sum_{u\in U}(R_{u,j}-\bar{R}_j)^2}}$$

Here Ru,i denotes the rating of user u on item i, Ri is the average rating of the i-th item.

Adjusted Cosine Similarity: one fundamental difference between the similarity computation in user-based CF and item-based CF is that in case of user-based CF the similarity is computed along the rows of the matrix but in case of the item-based CF the similarity is computed along the columns, i.e., each pair in the co-rated set corresponds to a different user. Computing similarity using basic cosine measure in item-based case has one important drawback the differences in rating scale between different users are not taken into account. The adjusted cosine similarity offsets this drawback by subtracting the corresponding user average from each co-rated pair. Formally, the similarity between items i and j using this scheme is given by

$$sim(i,j) = \frac{\sum_{u\in U}(R_{u,i}-\bar{R}_u)(R_{u,j}-\bar{R}_u)}{\sqrt{\sum_{u\in U}(R_{u,i}-\bar{R}_u)^2}\sqrt{\sum_{u\in U}(R_{u,j}-\bar{R}_u)^2}}$$

Here Ru is the average of the u-th user's ratings. The most important step in a collaborative filtering system is to generate the output interface in terms of prediction. Once we isolate the set of most similar items based on the similarity measures, the next step is to look into the target user's ratings and use a technique to obtain predictions. Here we consider two such techniques.

Research Methodology

The large number of available Profile online makes simple both for man and woman to search on the Internet.

Figure 3 shows the intelligent program in the online profile matching market works for both people and company. The system gives user a chance to find a match that would best suit their cast and subcat on the web site (written the web page with Dream Weaver and program with php). At the heart of the system are its matching capabilities. For the matchmaking process a database written on MySQL Community Server 5.1 stores all the profile submitted by different people. Another database stores the profile matching applications fed into the system. The system matches cast and sub cast. It also automatically does a ranking based on the matches using the personalization method namely item based collaborative filtering recommendation model.

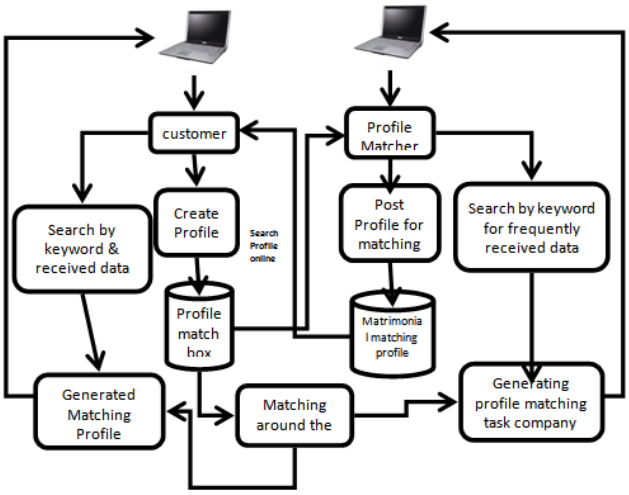


Figure 3

Outcome and Consideration

The Indian matrimonial market website shown the first home page in Figure 1. The matrimonial system gives the profile matcher a chance to find a partner that would best suit their cast age qualification etc. The web page: the user submits the information is shown in Figure 2. The attributes (or the key words) of the users that some of these, which must fill in are the field of the match (the cast categories), the location of the partner in India or other

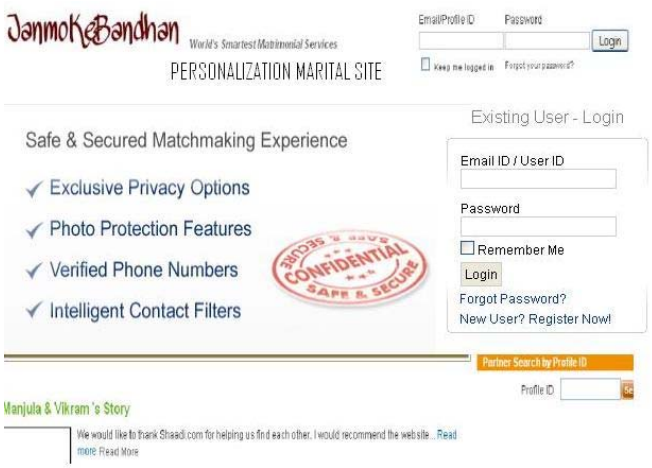


Figure 4: The home page of the web site provinces of country (the geographic regions) etc.

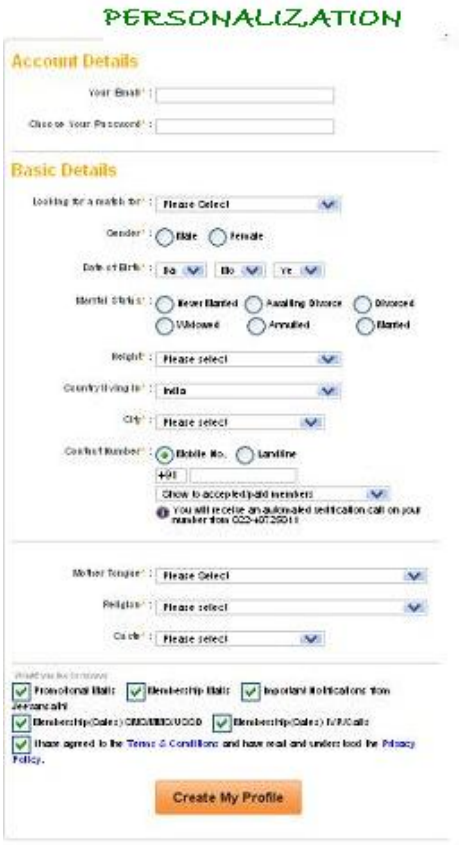


Fig 5. The page that the users must fill in

Fig 6. Search profile

The proposed system can help users find data that match specific profile. The system creates the part that the company must search the detail of the profile, the web page shown in Figure 3. The database about the matching profile then is offered for the correct partner match. The company must fills in some of these, education, age, and job, height, dob, sex, cast, sub cast, religion, mother tongue by the user. The intelligent program receives the information then solicits information from the user and then builds the profile from previous purchase pattern from the databases. In this stage the collaborative filtering based recommended systems using item based recommendation model was used to find out the suitable outputs proposing to the user. Besides of these, the system also provides many channels for both users and matrimonial company, such as the sudden listed of profile for match

making and of people for people hunter when they are the registered users. This system also provides many features such as the users can post available profile descriptions and advertise their services in e-mails and others on website.

Acknowledgment

We would like to express our gratitude to all those who gave us the possibility to complete this paper. We want to thank the Department of Engineering of the JJT University for giving me permission to commence this paper in the first instance, to do the necessary research work and to use departmental data. We are deeply indebted to our supervisor Prof. Dr. Akhilesh R. Upadhyay from the JJT University whose help, stimulating suggestions and encouragement helped me in all the time of our research work for our Phd. and writing of this paper.

Conclusion

The differences and similarities between personalization and customization work together toward the same goal making the Internet experience more pleasant for each user. These technologies are justified in that they benefit the Internet user and the company or organization managing the matrimonial website, as the user wills more than likely return after sensing the advantages of personalization and customization technologies.

References

[1] K.W. Wong, C.C. Fung, X. Xiao, K.P. Wong, "Intelligent Customer Relationship Management on the Web", Proceedings of the IEEE International Region 10 Conference, Melbourne, Australia, 5pp. Nov 2005.

[2] S. Y. Ho and K. Y. Tam, "An Empirical Examination of the Effects of Web Personalization at Different Stages of Decision Making", International Journal of Human-Computer Interaction, Vol. 19, No. 1, pp. 95-112, 2005.

[3] "Item-Based Collaborative Filtering Recommendation Algorithms", ACM Proceeding 1-58113-348-0/01/0005, 2001.

[4] Pew Internet Project, "Online Job Hunting : A Pew Internet Project Data Memo". July 17, 2002.

[5] T.Brice, and M.Waung, "Web Site Recruitment Characteristics : America's Best Versus America's Biggest", SAM Advanced Management Journal 67,no1, 2002.

[6] Goldberg, D., Nichols, D., Oki, B.M. and Terry, D. Using Collaborative Filtering to Weave an Information Tapestry.Communications of the ACM, 35, 12 (1992), pp. 61-70.

[7] Resnick, P., Iacovou, N., Sushak, M., Bergstrom, P., and Riedl, J. GroupLens: An open architecture for collaborative filtering of netnews. In Proceedings of the 1994 Computer Supported Cooperative Work Conference. (1994) ACM, New York. Pages 175- 186

[8] Joachims, T., Freitag, D., Mitchell, T. WebWatcher: a tour guide for the World Wide Web. In: Georgeff, M.P., Pollack, E.M., eds. Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1997. 770~777.

[9] J. Kramer, S. Noronha and J. Vergo, "A User-Catered Design Approach to Personalization", Communications of the ACM, pp. 45-48, August 2000.

[10] B. P. S. Murthi, S. Sarkar, ""The role of the management sciences in research on personalization", Management Science, Vol. 49(10),pp.1344-1362, 2003

[11] Srivastava, J., Cooley, R., Deshpande, M., Web usage mining: discovery and applications of usage patterns from Web data, ACM SIGKDD Explorations Newsletter, Volume 1 , Issue 2, January 2002, pp. 12-23, ISSN:1931-0145, 2000.

[12] Mamcenko, J., Kulvietiene, R., Web Usage Mining of Distance Education Information System, WSEAS TRANSACTIONS on SYSTEMS, Issue 12, Volume 4, ISSN 1109-2777, December 2005.

[13] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization", ACM Transactions on Internet Technology, Vol. 3,No. 1, 2003, pp. 1-27.

[14] Nadkarni, P. M. (2000). Information retrieval in medicine. Overview and applications, Journal of Postgraduate Medicine,Vol. 46, pp.116-122.

[15] Rose, D.E. & Levinson, D. (2004). Understanding user goals in web search. Proceedings of the 13th International Conference on World Wide Web, pp. 13-19.

[16] Smyth, B. (2007). A Community-Based Approach to Personalizing Web Search, Cover Feature, IEEE Computer, 40(8),pp.42-50.

[17] Spink A., Yang Y., Jansen, J., Nykanen P., Lorence, D.P., Ozmutlu, S. & Ozmutlu, H.C. (2004). A Study of medical and health queries to web search engines, Health Information and Libraries Journal, Vol. 21, pp.44-51.

Terahertz Technology and Its Applications

Vidhi Sharma[1], Dwejendra Arya[2] and Megha Jhildiyal[2]

[1]Electronics and Communication Engineering, Gurgaon College of Engg. Gurgaon, India
[2]Electronics and Communication Engineering, I.E.T Alwar 301001 Rajasthan, India
[3]Electronics and Communication Engineering, Gurgaon College of Engg. Gurgaon, India
E-mail: vidhisharma777@gmail.com, d_arya2007@yahoo.co.in, megha.tuks007@gmail.com

Abstract
In this paper we review the recent progress of terahertz (THz) technology and its applications in various fields such as THz sensing for bio, medical, medicine, security and others.

Keywords: THz technology, THz spectroscopy, THz sciences.

Introduction

The term 'T-rays' was coined in the early 1990's by Bell Labs to describe the spectrum in the Terahertz range (1 THz = 1012). [1] The terahertz region of the spectrum lies on the border of where electronics and optics meet between the mm-wave and infrared bands (100 GHz-10 THz). THz wave could handle ultra broad band signals , have very large absorption due to water or water vapours and are transparent through many materials (e.g. plastics , papers, cloth and oil) that are opaque in visible and IR light. Many materials have a so called fingerprint spectrum in the spectrum range therefore terahertz wave are expected to be applied to ultrafast wireless communications, scanning systems of hazardous materials and assay devices for medical examinations. They are also expected to be applied to the multiresidue analysis for agrichemical, medical diagnostics, environmental assessment, process monitoring system for industrial products and biometric security. There is an atmospheric prevailing that terahertz technology represents the drawn of a new era.[3]

Applications

Overview: **-** The terahertz regime is sandwiched between the microwaves and the infra-red, bridging the gap between electronics and optics. Due to this exposed position in the electromagnetic spectrum, a plethora of metrological applications with high impact on a variety of industries exists. Even though the terahertz (THz) frequency range (0.1 THz to 10 THz) it has proven to be one of the most elusive parts of the electromagnetic spectrum.[2] For a long time, the THz regime was also referred to as the "THz gap" as neither optical nor microwave devices could fully conquer this shadowy domain with its many hidden scientific treasures. Little commercial emphasis was placed on the development of THz systems as the available sources, such as synchrotrons, backward wave oscillators, Smith-Pur-cell emitters or free electron lasers, were very costly components.[4] The numerous potential THz applications, ranging from medical imaging, security scanning and process control to ultra-fast communications, lead to a rapid development of THz systems and today first THz products are on the brink of large-scale market introduction. THz applications cover different industrial fields, illustrating the broad applicability of THz technology.

THz for the Polymer Industry

Despite the fact that polymers are a relatively young group of materials, they already managed to overhaul steel, the all-time favourite construction material, in terms of production volume. Many plastic materials are transparent to THz waves, so that they are commonly used as base material for THz optics, serving as inexpensive system components.[5] THz is returning the favour by providing an extensive set of quality inspection tools, often superior to existing measurement methods.

Dispersion Quality control for Polymeric compounds

Most technical plastics consist of a base polymer loaded with fine additive particles, which allow to custom tailor the functionality of the material. For example, the colour, the flammability or even the mechanical properties can be adjusted by com-pounding. However, in- or online monitoring of the dispersion quality achieved in such compounding processes is a challenging task. Here, THz technology can provide a valuable contribution.

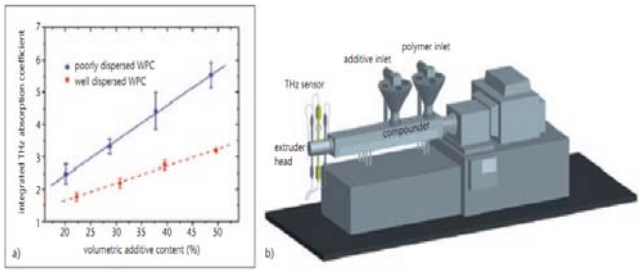


Figure 2.1 : Dispersion quality measurements on wood plastic composite samples

Figure 2.1 a) shows offline measurements of the absorption co-efficient of well and poorly dispersed wood plastic composite (WPC) samples, integrated in the frequency

range between 0.2 and 0.8 THz over the additive content. The absorption rises towards higher concentrations of the wood particles, as wood is less transparent to THz radiation than the base polymer (polypropylene), and due to the fact that the number of scattering centres increases.[7] Latter aspect is also responsible for the stronger absorption in case of the poorly dispersed WPC samples. In Figure 2.1 b), an application scenario is depicted. At the outlet of an extruder unit, a THz sensor head is installed, delivering the required data for the evaluation of the dispersion quality.

Plastic Weld Joint inspection

As plastic continuously replaces more expensive materials like metal or ceramics, the demands on the joining technology for plastic components are steadily increasing. Especially important is the plastic welding, which, if applied successfully, forms a stable physical bond. For example, high-density polyethylene pipes for gas and water transportation rely on this technology and any defect in the weld joint is associated with high costs and risks.[4] Non-destructive testing of plastic weld joints seemed to be a permanent obstacle. X-rays and ultrasonic waves were not able to satisfactory identify inclusions or delaminations and the only means of obtaining low failure rates was the careful monitoring of the welding process parameters.

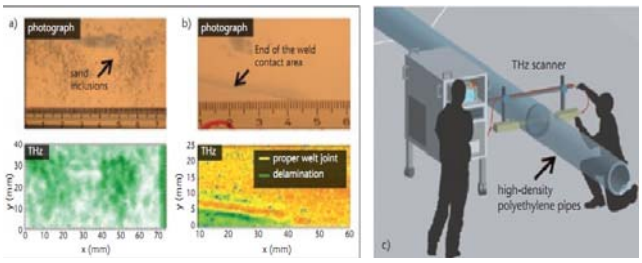


Figure 2.2 : Photographs and THz images of plastic weld joints

Initial measurement results for a weld joint with dielectric inclusions (sand) and a weld joint with a delamination are shown in Figure 2.2 a) and b) respectively. The THz waves clearly reveal both defect types – contact-free and non-destructive. The final system should be a mobile unit, which can directly be employed at the places where the weld joints are created. Figure 2 c) illustrates our vision for such a system.[3]

Fiber Reinforced Plastic composites

Another class of polymeric materials, which can benefit from THz technology, are fibre reinforced plastics. Here, the mechanical properties of the component are directly connected to the fibre orientation. Again, non-destructive testing methods are extremely rare and limited to specialized cases.

As the orientation of the fibers inside the host medium directly results in a birefringent behaviour of the composite material for sub-mm waves, THz technology can be employed to identify the fiber orientation.[2] While carbon fiber composites are strongly absorbing, so that the measurement options are limited to reflection geometries, glass fiber composites are fairly transparent to THz waves and can be analyzed in a transmission configuration.

Inspection of Safety critical components

Some polymeric components are employed in safety critical places. For example, plastic airbag caps commonly used in passenger cars. A thin grove serves as predetermined breaking line, where the cap is supposed to crack in case of an accident. However, the cap should withstand other non-critical physical impacts. Thus, the functionality of the cap depends on the correct grove depth of the predetermined breaking line. Controlling this thickness is not easily accomplished as ultrasonic measurements do not yield a resolution in the µm range. THz imaging can be employed to accurately determine the grove thickness by measuring the propagation delay of the THz pulse.

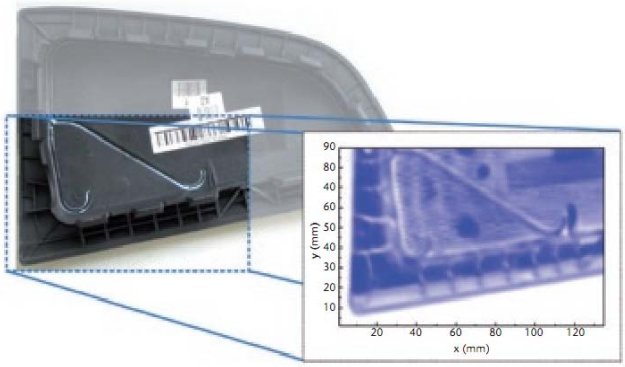


Figure 2.3 : THz imaging of security critical components

A THz image of an airbag cap, obtained in an inexpensive cw THz system, is shown in Figure 2.3. The thickness profile is clearly visualized, demonstrating the ability of cw THz systems for cost efficient quality inspection of polymeric components.[5]

Quality control for food Products

Apart from the polymer industry, another high potential application field for THz systems lies in the food industry quality control. While metallic contaminations (like screws) can easily be detected, non-metallic contaminations (such as stones, plastic parts or glass) seriously challenge conventional measurement systems. Failure in detection is not an option at any rate as it might lead to potentially fatal health risks.[7]

Conventional X-ray systems suffer from the low dielectric contrast between the food product and the contamination. Differentiation between a nut splinter and a sharp glass piece is close to impossible with x-ray devices. Ultra-sonic systems are not contact-free, so that integration in a process line is hard to achieve. Fortunately, THz time domain spectroscopy together with smart data processing algorithms allows for the secure distinction between desired ingredient and unwanted inclusion.

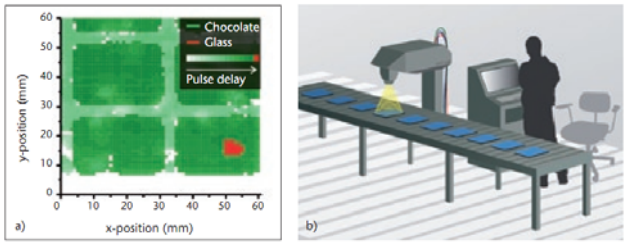


Figure 2.4 : Dielectric inclusions in food products

Figure 2.4 a) shows an image of a nut chocolate with an embedded glass piece. A smart algorithm corrects the error introduced by the uneven surface and the glass piece is clearly visible in the THz scan.[4] Current efforts in developing THz sensor arrays and line scan units will enable the direct integration into the production line as depicted in the schematic drawing in Figure 2.4 b). However, scan speed remains a critical issue for THz imaging applications and further innovations are required before such systems can be introduced into large scale production.

Security Scanning Applications

One of the most receptive fields in regards to THz technology is the security sector. THz systems have unique feature to reveal hidden objects beneath clothing, even if they are of non-metallic nature like ceramic weapons or explosives. Apart from revealing hidden objects, THz systems can also be employed for chemical recognition. Each material has its own spectroscopic fingerprint and the ability of THz systems to access both the refractive index and the absorption information leads to a high accuracy in the detection process.[6] Especially explosives are of interest and within this group foremost the liquid explosives receive much attention, as no other systems exist for their reliable recognition. The vision is a mobile, standoff detection system for check points in conflict areas or a handheld security scanner for the secure identification of liquids at airports.

Figure 2.5: terahertz image of men with hidden knife

Hydrology Monitoring of Plants for optimized irrigation Strategies

The steadily increasing water shortage, induced by desertification and global warming, requires intelligent irrigation strategies, especially for economic plants. The basis for developing such strategies is the understanding of the plant's hydrology.

Here, THz technology might play a key role: Due to the high absorption coefficient and the strong abnormal dispersion of water in this frequency region, the water content of a leaf considerably influences the THz wave propagation.[3] Therefore, THz systems can be employed to investigate this important quantity.

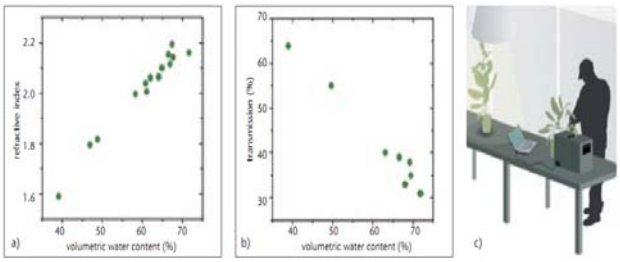


Figure 2.6 : THz technology contributing to the understanding of Plant Hydration

As shown in Fig. 2.6, both the leaf's refractive index and its transmission perceptibly depend on the water content. Thus, these measurements can be utilized to observe the water status.[1] The high spatial resolution achievable with THz systems due to the sub-mm wavelengths furthermore enables the potential of investigating the water transport and the water distribution within a leaf by cw or pulsed THz imaging.

Polar Molecules

With THz radiation one has the ability to detect and identify most polar molecules in the gas phase. This application requires THz radiation with broad bandwidth and relies on the fact that, as in the mid-infrared, many molecules have characteristic "fingerprint" absorption spectra in the terahertz region.

Material Characterization

For physics this new technology is very interesting because they can use it for characterization of materials like semiconductors and lightweight molecules.[7] The radiation can be used to determine the carrier concentration and mobility of semiconductors10, and in the superconductor research it can be used to determine the parameters of superconducting materials.

Study of Historical and Archaeological Work

Terahertz imaging methods work with extremely low powers, about 2 to 3 magnitudes of order lower than the blackbody radiation in this frequency range. The radiation is therefore non invasive to historical work, neither to paintings or paper in common nor to any kind of stone or metal. Thus this technique could be useful in history, archaeology etc.

Biomedical

In the area of biomedical diagnostics we can make use of THz tomographie. Although there is a limited penetration depth of

the radiation due to the strong water absorption, which excludes the use of THz radiation in most biomedical research areas, it can be used to examine tissue near the surface, in particular skin and teeth. On the other hand, the sensitive to water enables the investigation of tissue hydration. This opens a range of applications including analysis of burn depth and severity, and detection of skin cancer and caries.[2] A reliable non-invasive probe of burn depth would be of great value to physicians, who currently have no such technology. The detection of skin cancer works very well. Breast cancer detection could be an application as well; because of the lower water content of the tissue there. The detection of caries is another possible application.

Communication

Potential uses exist in high-altitude telecommunications, above altitudes where water vapour causes signal absorption: aircraft to satellite, or satellite to satellite.

Conclusion

Numerous applications for THz technology exist and many industrial branches can benefit from its unique capabilities. Further improvements in terms of measuring speed, system robustness and cost efficiency have to be implemented to make THz systems even more competitive. Fortunately, a plethora of research projects are currently pursued, so that THz systems rapidly approach large-scale market introduction

References

[1] D. Mitlleman, Ed. Sensing with terahertz radiation. Berlin, Heidelberg: Springer (2003)

[2] R. Wilk, F. Breitfeld, M. Mikulics, and M. Koch, Continuous wave terahertz spectrometer as a noncontact thickness measuring device, Applied Optics, Vol. 47, Issue 16, pp. 3023-3026 (2008)

[3] S. Wietzke, C. Jansen, F. Rutz, D. M. Mittleman, and M. Koch, Determination of additive content in polymeric compounds with terahertz time-domain spectroscopy, Polymer Testing, vol. 26, no. 5, pp. 614–618 (2007)

[4] S. Wietzke, C. Jördens, N. Krumbholz, B. Baudrit, M. Bastian, and M. Koch, Terahertz imaging: a new non-destructive technique for the quality control of plastic weld joints, Journal of the European Optical Society – Rapid Publications, vol. 2, pp. 07013 (2007)

[5] U. Ewert et al., Non-Destructive Testing of Glass-Fibre Reinforced Polymers using Terahertz Spectroscopy, ECNDT (2006)

[6] C. Jördens and M. Koch, Detection of foreign bodies in chocolate with pulsed terahertz spectroscopy, Opt. Eng., Vol. 47, 037003 (2008)

[7] C.Jördens, M. Scheller, B. Breitenstein, D. Selmar, and M. Koch, Evaluation of the Leaf Water Status by means of the Permittivity at Terahertz Frequencies, submitted to Journal of Biological Physics (2008)

Practical Implementation of Faster Arithmetic Coding Using Total Frequency in Power of Two

[1]Jyotika Doshi and [2]Savita Gandhi

[1]GLS Institute of Computer Technology; Gujarat Technological University, Ahmedabad-380006, India
E-mail: jyotika_doshi@yahoo.com
[2]Rollwala Computer Centre; Gujarat University, Ahmedabad-380009, India
E-mail: drsavitagandhi@gmail.com

Abstract
Arithmetic coding method is based on the fact that the cumulative probability of a symbol sequence corresponds to a unique subinterval of the initial interval [0, 1). The process iterates for each symbol to be encoded/decoded. At each iteration, current interval is divided into subintervals as per probability of the symbols and a subinterval corresponding to the symbol to be encoded/decoded is selected as new interval. When subinterval becomes very narrow, most significant bit of low and high bounds of interval becomes equal and this bit is extracted as an output. While implementing with integer arithmetic, for computing subintervals, a ratio of cumulative frequency to total frequency of symbol is used as cumulative probability. If frequencies are converted so that total frequency results in power of two, one can use shift operations for division and multiplication by total frequency and achieve performance in execution. It is observed that compression is 30% faster and decompression is 10% faster using this approach.

Keywords: data compression, arithmetic coding, static model, byte stream model, encoding, decoding, scaling frequencies, total frequency in power of two, shift operations, faster execution, practical implementation

Introduction to Arithmetic coding

Arithmetic coding is an entropy coder providing lossless compression. It works with any sample space so it can be used for the coding of text in arbitrary character sets as well as binary files [10]. It encodes data using a variable number of bits. The number of bits used to encode each symbol varies according to the probability assigned to that symbol. The idea is to assign short codeword to more probable events and long codeword to less probable events [8].

This data compression technique encodes a stream of input symbols with a code string that is a single fractional value on the number line between 0 and 1. This single output number can be uniquely decoded to create the exact stream of input symbols.

The method is based on the fact that the cumulative probability of a symbol sequence corresponds to a unique subinterval of the initial interval [0, 1) [1]. Before starting encoding process, symbols are assigned segments on interval [0, 1) according to their cumulative probabilities. It doesn't matter which symbols are assigned which segment of the interval as long as it is done in the same manner by both the encoder and the decoder [4]. If $S = (S_1, S_2, \ldots, S_n)$ is the alphabet of a source having n symbols with an associated cumulative probability distribution $P = (P_1, P_2, \ldots , P_n)$, an initial interval [0, 1) can be divided into n subintervals as $[0, P_1)$, $[P_1, P_2)$, $[P_2, P_3)$, …,$[P_{n-1}, P_n)$ where P_i is the cumulative probability of symbol S_i. Each subinterval length is proportional to the probability of the symbols [2].

This coding algorithm is symbol-wise iterative; i.e., it operates upon and encodes (decodes) one data symbol per iteration. On each iteration, the algorithm successively partitions current interval using the assigned segments on [0,1) and retains one of the partitions as the new interval corresponding to the symbol to be encoded. For symbol S_i in an iteration, if current interval is [Low, High), new subinterval is computed as [Low + Range x P_{i-1}, Low + Range x P_i) where Range = High – Low and $[P_{i-1}, P_i)$ is a subinterval of symbol S_i in an interval [0,1) 1[1, 2, 3, 4].

At the end of encoding process, i.e. after processing all input symbols, a number X is chosen from the last resulting interval as a single output code number. Choose any value with the shortest representation in the final interval [5].

While decoding, decoder starts with an initial interval [0,1) and encoded output code number X. Decoder divides current interval into different segments (just like encoder) and determines a segment where code X lies. The symbol corresponding to this segment is decoded as an actual input symbol and this segment becomes new interval. The process is repeated till all input symbols are recovered from X.

Main advantages of arithmetic coding are its flexibility and optimality. The drawback of the arithmetic coding is its low speed due to several multiplications and divisions needed for in subdividing current interval [8]. Compression ratio that can be reached by any encoder under a given statistical model is actually bounded by the quality of that model. However one can optimize one's algorithms in at least two dimensions: memory usage and speed [11].

Practical Implementation

Practical methods implement arithmetic coding on computers with fixed-sized 16 or 32 bit integer arithmetic. Use of integer operations helps to reduce truncation/rounding errors and give accuracy. Accuracy is necessary for lossless techniques.

Arithmetic coding, implemented as a byte stream model, considers each symbol made up of single byte. Thus possible symbols are 256 (values 0 to 255). An additional symbol EOF (End Of File) is used as a special symbol occurring only once at the last as input terminator. [2, 4, 6]

Initial interval [low, high)

Initial interval [0,1) is a range with low as all zeros and high as all 1's. With implementation using 16-bit (or 32 bit) unsigned math, the initial value of high is taken as FFFFh (FFFFFFFFh for 32-bit) and low as 0. It can be considered as a number with assumed decimal point before first bit, i.e. [.0000000000000000, .1111111111111111) with high value quite nearer to 1 but not 1.

Frequency of Symbols

Knowledge of frequencies is required at decoding stage to compute subintervals using cumulative probabilities. So frequencies need to be stored with compressed data. To save space on compressed file, frequencies are scaled to 1 byte value in the range 0 to 255. Actual frequencies are divided by 256. While scaling frequencies, care is to be taken so that the frequency of symbol occurred in source file should not become zero, it should be at least 1. (ex. For a symbol that has occurred 200 times in source file, integer division 200/256 will result into 0).

Cumulative frequency of symbols

Using frequencies, cumulative frequencies F_i can be computed. Cumulative probability P_i is the ratio of cumulative frequency to total frequency, i.e. $P_i = F_i$ / total frequency.

In C programming, to store cumulative frequencies of various symbols in main memory, a 1-D array data structure is used with 258 elements as $\{F_0, F_1, ..., F_{257}\}$ where symbol 257 is for EOF. F_i denotes occurrence of symbols less than symbol i. Thus F_0 is 0 and F_{257} is total frequency. Thus cumulative frequency range of symbol i is $[F_i, F_{i+1})$.

When encoding/decoding symbol i, new subinterval is computed as [1, 2, 3, 4, 6, 7, 9, 11, 12]

High = Low + Range x P_{i+1}
= Low + Range x F_{i+1} / total_frequency
Low = Low + Range x P_i
= Low + Range x F_i / total_frequency

Instead of P_i, using F_i / total_frequency reduces the effect of propagation of truncation errors. To avoid overflow errors, cumulative frequencies are limited to 2^{14} = 16384. This restriction again needs scaling of frequencies [4].

Existing System implementations

- Compute frequencies of symbols
  - Scale to have maximum 255 (for storing with compressed file)
  - Scale so as to have maximum total frequency = 16834 (to reduce overflow errors)
- Compute cumulative frequencies
- During encoding:
  Subintervals are computed as many times as number of symbols (bytes) in source (input) file. If Si is a symbol to be encoded, total_freq is total frequency of symbols after scaling, $[F_i, F_{i+1})$ is a segment in interval [0, total_freq) according to cumulative frequency of symbol S_i, current interval is [low, high) in range [0,1) is computed as follows [4,6,7,9,11,12]:

range = (high-low) + 1;
high = low + (((range*F_{i+1})/ total_freq) - 1); --- (**3.1**)
low = low + ((range * F_i) / total_freq); --- (**3.2**)

When range becomes too narrow, most significant bit of high and low becomes equal and this bit is extracted and appended in an output encoded number, say code.

During decoding

An encoded fraction number, say code, is read from compressed file. This code is a fraction in an interval [0,1). Using this code, it computes count in an interval [0, total_freq) as (((code - low + 1) * total_frequency - 1) / range). A segment is determined where this count falls and a corresponding symbol is output as a decoded symbol. It computes new subinterval exactly like in encoding phase. When most significant bit becomes equal, it is simply extracted. A new bit is read and appended to code. The process repeats till all symbols are decoded.

C-code to figure out subinterval in which code lies:

range = (high - low) + 1;
count=(((code-low+1)*total_freq-1)/range); --- (**3.3**)

C-code to determine symbol corresponding to the interval where count lies: for (s=256 ; count< F[s] ; s--); Here one may use bisection search or other search methods. While implementing, bisection search is used here.

C-code (executed as many times as the number of recoverd symbols, i.e. symbols in source file) to compute new interval after after decoding symbol s, say S_i with its cumulative frequency segment $[F_i, F_{i+1})$ in interval [0,total_freq)

range = (high-low) + 1;
high = low+((range*F_{i+1}) / total_freq - 1); --- (**3.4**)
low = low+((range * F_i) / total_freq); --- (**3.5**)

Proposed System with total frequency in power of two

As said before, the drawback of the arithmetic coding is its low speed due to several multiplications and divisions needed for each symbol being encoded or decoded. Proposed system improves over execution time.

In arithmetic coding, normalizing the interval requires several multiplication and division operations. Due to integer implementations, it uses cumulative frequency/total frequency instead of cumulative probability (as it will be fraction value). During encoding/decoding stage, division by total_freq is performed as many times as the number of symbols in source (input) file as shown in statements (3.1), (3.2), (3.4) and (3.5). Similarly during decoding stage, multiplication by total_freq (statement (3.3)) is performed as many times as the number of codes in compressed file.

In the proposed system, execution time of statements (3.1) to (3.5) can be reduced if total_freq is in power of two. Execution speed can be increased using bitwise shift

operations wherever it performs multiplication and division by total frequency. To have total frequency in power of two, frequencies need to be rescaled and adjusted.

For total frequency to be in power of two, say 2^k, k can be selected as the minimum integer value such that $2^k \geq$ original total frequency. (Refer Table I in section V). Now division and multiplication by total frequency can be performed using k times right shift and left shift operation respectively in statements (3.1) to (3.5).

Proposed System: Practical implementation

Proposed system needs to have total frequency in power of two, i.e. total frequency $=2^k$. First of all, to understand selection of k and normalization of frequencies, consider the example given in table I. Here only 10 symbols are considered instead of 256 (just due to space limitation in table on page) and maximum limit of frequency is considered as 20 instead of 255 (one byte value). As seen, total frequency is 87. Nearest number greater than 87 in power of 2 is $128 = 2^7$, so normalize these frequencies to be out of 128. Now total of frequencies should be 128, but it may not be so due to truncation/rounding errors. Here the values are shown as truncated due to integer operations. Other problem is that some frequencies (for example, 29 and 26) are crossing maximum limit of 20 and it should not be permitted. Thus total frequency is 107 only instead of 128. Remaining (128-107 = 21) frequencies can be distributed equally among the symbols that have occurred. So 21/10=2 is added to each frequency, maintaining maximum limit of 20. Still 5 is remained, which may be assigned to EOF symbol.

As seen in the table I, after converting frequencies from 2^k, new total may not be still 2^k. This is due to truncation/rounding errors and a limitation of maximum frequency. Such scenario needs to make some adjustments to distribute remaining counts evenly among symbols with non-zero frequency. This scaling and adjusting the frequencies may lead the probability distribution of symbols to deviate from its distribution based on actual frequencies. This may lead to little increase in compressed file size.

Steps to scale and then adjust the frequencies:

- Compute frequencies of symbols
- Scale to have maximum 255 (for storing with compressed file)
- Scale so as to have maximum total frequency = 16834 (to reduce overflow errors)
- Scale to have total frequency in power of two (variable pwr2 is used for new total frequency)
  - Compute total frequency, say total_old
  - Determine minimum k such that pwr2 = $2^k \geq$ total_old
  - Scale frequencies as freq = freq*pwr2/total_old.
    - If freq is beyond 255, keep it to maximum 255
    - If frequency of occurred symbol becomes zero, keep it 1
- Compute total of newly scaled frequencies. If total is still less than pwr2, then redistribute the remaining difference almost equally among symbols that have occurred in source file
  - If num_sym is the number of distinct symbols that have occurred in source file, distribute (pwr2-total)/num_sym to each occurred symbol. See that frequency remains in maximum limit.

Table I: Normalizing and Adjusting Frequencies to have Total Frequency in Power of Two

Symbols →	**S 0**	**S 1**	**S 2**	**S 3**	**S 4**	**S 5**	**S 6**	**S 7**	**S 8**	**S 9**	**Total frequency**
Frequencies after scaling to have maximum as 20	5	2	5	20	18	12	10	8	5	2	87
Freq. normalized to have from 128 (truncated)	7	2	7	29	26	17	14	11	7	2	122
Due to limitation of maximum frequency 20	7	2	7	20	20	17	14	11	7	2	107
Almost equal distribution of remaining	9	4	9	20	20	19	16	13	9	4	123

Now shift operations can be used in C-code as follows:

During encoding: Change statements (3.1) and (3.2) as

high = low + (((range * Fi+1) >> k) - 1); ---- (**5.1**)

low = low + ((range * Fi) >> k); --- (**5.2**)

During decoding: Change statements (3.3) to (3.5) as

count=((((code-low+1)<<total_freq)-1)/range); --- (**5.3**)

high = low + (((range * Fi+1) >> k) - 1); --- (**5.4**)

low = low + ((range * Fi) >> k); --- (**5.5**)

Implementation Results

Existing and proposed systems are implemented using 16 bit Turbo C compiler on Intel(R) Pentium (R) D, CPU 3.00 GHz, 1 GB RAM, Windows 95 OS. Execution time is measured in seconds for 17 files with varying sizes and file types. Some of the test files are selected from Calgary and Canterbury corpus, a widely used benchmark and also from web site *compression.ca/act/act_files.html* .

Table 2 lists files used for testing of both existing and proposed implementations. Table 3 and Table 4 include compression and decompression time (seconds) respectively and gain (in percentage) in speed using proposed system. Figure 1 and 2 shows comparison of execution time. Table 5 presents resulting compressed file size and Table 6 shows the compression rate under both implementations. Figure 3 and figure 4 shows comparison of compressed file size and compression rate.

Table II: Files used for testing.

SrNo	**File type (from where)**	**File name**	**File size (KB)**
1	text file (own)	test.txt	744
2	Image file (own)	shriji.jpg	4389
3	pdf file (own)	linux.pdf	7902
4	photograph (own)	family1.jpg	194
5	powerpoint file (own)	linuxfil.ppt	241
6	word document (own)	cycle.doc	1449

7	powerpoint show (own)	sadvchar.pps	1756
8	text file (act website)	world95.txt	2935
9	excel file (act website)	act2may2002.xls	1317
10	sound file (act website)	every.wav	6831
11	graphics file (act website)	lena3.tif	769
12	graphics file (act website)	monarch.tif	1153
13	executable file (act website)	pine.bin	1530
14	excel file (cantrbry corpus)	kennedy.xls	1006
15	file (cantrbry corpus)	ptt5	502
16	file (calgary corpus)	book2	597
17	file (calgary corpus)	obj2	242

Table III: Compression time and % gain in execution time with proposed system.

SrNo	Existing system Compression time (seconds)	Proposed system Compression time (seconds)	% gain in compression time
1	1.208	0.714	40.89
2	9.505	6.593	30.64
3	16.538	11.263	31.90
4	0.439	0.274	37.59
5	0.439	0.329	25.06
6	2.637	1.703	35.42
7	3.791	2.637	30.44
8	5.604	3.736	33.33
9	2.307	1.593	30.95
10	15.000	10.549	29.67
11	1.703	1.208	29.07
12	2.528	1.703	32.63
13	3.077	2.197	28.60
14	1.640	1.098	33.05
15	0.604	0.440	27.24
16	1.099	0.714	35.03
17	0.495	0.329	33.54
Total→	68.614	47.0797	31.38

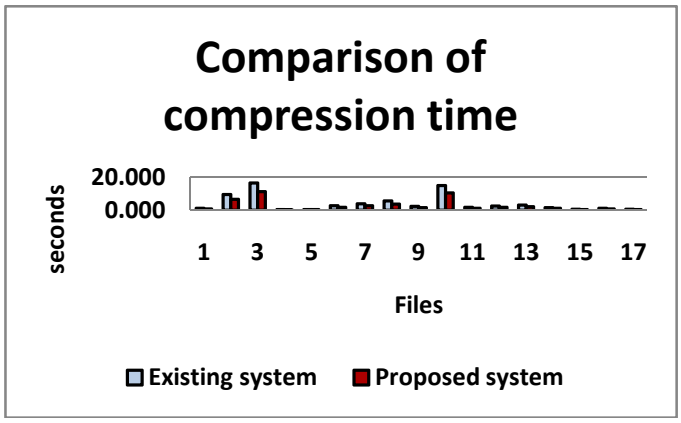


Figure I: Comparison of Compression time.

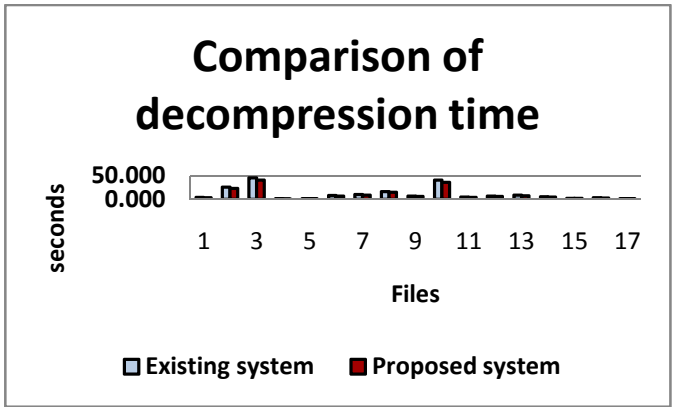


Figure II: Comparison of Decompression Time.

Table IV: Decompression time and gain in execution time with proposed system.

SrNo	Existing system Decompression time (seconds)	Proposed system Decompression time (seconds)	% gain in decompression time
1	3.956	3.461	12.51
2	26.099	23.187	11.16
3	45.879	40.824	11.02
4	1.154	0.989	14.30
5	1.310	1.153	11.98
6	7.912	7.033	11.11
7	10.384	9.230	11.11
8	16.648	14.835	10.89
9	6.868	6.428	6.41
10	40.769	36.483	10.51
11	4.560	4.065	10.86
12	6.868	6.153	10.41
13	8.791	7.857	10.62
14	5.270	4.78	9.30
15	2.362	2.142	9.31
16	3.351	3.021	9.85
17	1.374	1.263	8.08
Total→	193.5554	172.904	10.67

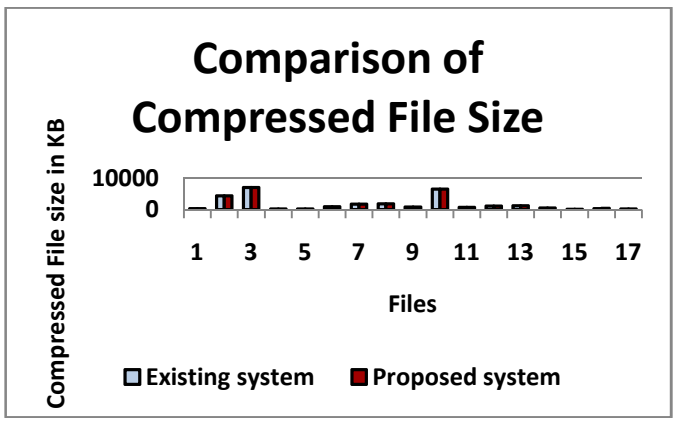


Figure III: Comparison of Compressed File Size.

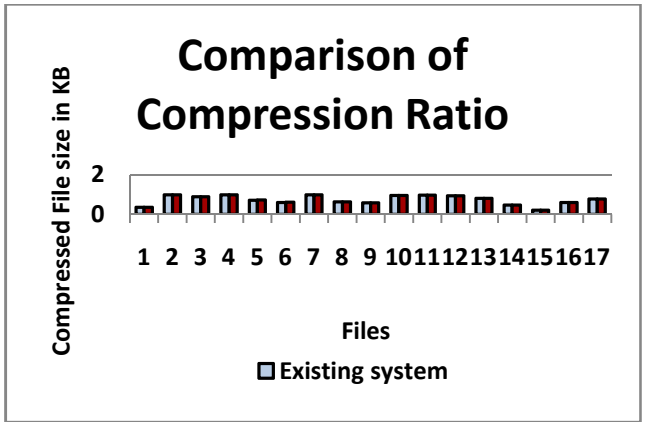


Figure IV: Comparison of Compressed Ratio.

Table V: Compressed file size and %loss in file size with proposed system.

SrNo	File size (KB)	Existing system	Proposed system	% loss in compressed file size
		Compressed file size (KB		
1	744	268	269	0.37
2	4389	4376	4376	0.00
3	7902	7032	7051	0.27
4	194	193	193	0.00
5	241	172	175	1.74
6	1449	871	889	2.07
7	1756	1730	1731	0.06
8	2935	1881	1885	0.21
9	1317	772	774	0.26
10	6831	6560	6560	0.00
11	769	745	745	0.00
12	1153	1080	1080	0.00
13	1530	1236	1236	0.00
14	1006	467	470	0.64
15	502	106	110	3.77
16	597	357	360	0.84
17	242	190	190	0.00
Total→		28036	28094	0.21

Table VI: Compression ratio.

SrNo	File size (KB)	Existing system Compression ratio	Proposed system Compression ratio
1	744	0.36021505	0.3615591
2	4389	0.99703805	0.9970380
3	7902	0.88990129	0.8923057
4	194	0.99484536	0.9948454
5	241	0.71369295	0.7261411
6	1449	0.60110421	0.6135266
7	1756	0.98519362	0.9857631
8	2935	0.64088586	0.6422487
9	1317	0.58618071	0.5876993
10	6831	0.96032792	0.9603279
11	769	0.96879064	0.9687906
12	1153	0.9366869	0.9366869
13	1530	0.80784314	0.8078431
14	1006	0.46421471	0.4671968
15	502	0.21115538	0.2191235
16	597	0.59798995	0.6030151
17	242	0.78512397	0.7851240

Observations

Significant reduction in execution time of compression and decompression is observed in proposed system. Overall reduction in execution time is 31.38% and 10.67% for encoding and decoding respectively as compared to existing implementations.

Overall loss in compressed file size is 0.21% with proposed system as compared to existing system. This loss is due to truncation/rounding errors and maximum frequency limit of 255 while rescaling frequencies to have total frequency in power of two. Due to rescaling/adjusting, probability distribution deviates from its distribution based on actual frequencies.

When data of existing and proposed system is analysed applying statistical test at 1% level of significance, it shows that there is no significant difference in mean compressed file size and mean compression ratio. But reduction in mean execution time for encoding and decoding is found to be significant.

Conclusion

There is a significant performance achievement in execution speed during encoding and decoding. Though there is little increase in compressed file size as compared to existing implementation, the proposed variation of arithmetic coding (with rescaled frequencies to make total frequency in powers of two) runs much faster than a usual implementation of arithmetic coding with actual frequencies.

Bibliography

[1] Ida Mengyi Pu, *Fundamental Data Compression*, Butterworth-Heinemann, 2006

[2] David Salomon, *Data Compression-The Complete Reference*, 3rd Edition, Springer, 2004

[3] Adam Drozdek, *Elements of data compression*, Brooks/Cole, 2002

[4] Mark Nelson and Jean-loup Gailly, ***The Data Compression Book*, 2nd edition,** M&T Books, New York, NY 1995

[5] Amir Said, "Introduction to Arithmetic Coding - Theory and Practice", available at *http://www.hpl.hp.com/techreports/2004/HPL-2004-76.pdf*

[6] Moffat, Neal, and Witten, *"Arithmetic Coding Revisited", ACM Transactions on Information Systems,* 16(3):256-294, July 1998, available at *http://www.cs.mu.oz.au/~alistair/abstracts/mnw98:acmtois.html*

[7] I.H. Witten, R.M. Neal, and J.G. Cleary, *"Arithmetic coding for data compression," Commun.ACM,* vol. 30, no. 6, pp. 520-540, June 1987.

[8] P.G. Howard and J.S. Vitter, *"Arithmetic coding for data compression", Proceedings of the IEEE,* 82(6):857-865, June 1994a.

[9] Paul G. Howard and Je_rey Scott Vitter, *"Practical Implementations of Arithmetic Coding",International Conference on Advances in Communication and Control (COMCON 3),*Victoria, British Columbia, Canada, October 16-18, 1991

[10] G.G. Langdon, *"An introduction to arithmetic coding", IBM Journal of Research and Development,* Vol. 28(2):135-149, March 1984.

[11] Eric Bodden, Malte Clasen, Joachim Kneis, *"Arithmetic Coding revealed-A guided tour from theory to praxis",* Sable Technical Report No. 2007-5, May 2007, available at *http://www.bodden.de/legacy/arithmetic-coding/*

[12] J. Carpinelli, A. Moffat, R. Neal, W. Salamonsen, L. Stuiver, A. Turpin and I. Witten,*"Word, character, integer, and bit based compression using arithmetic coding"*, Relevant software available at *http://www.cs.mu.oz.au/~alistair/arith_coder*

Factors Affecting the Design of Emotionally Engaging Games

Dr Clive Chandler

Faculty of Computing, Engineering and Technology, Staffordshire University, Staffordshire, England
E-mail: c.chandler@staffs.ac.uk

Abstract
As yet games have not achieved the engagement factor seen in the movie industry, the holy grail for any games designer is to engage, excite and attract their players without running into the downside of addiction

There are two major factors to achieving this goal that need to be considered:

1. Does the player "Believe" the game
2. The emotional "roller-coaster" ride experienced by the player

If these are achieved then the game becomes a XXX game. This paper seeks to investigate the factors affecting these two goals and offers guidelines in order to achieve a successful implementation, avoiding the obvious design pitfalls.

Index Terms: Games, Emotions, Design, Guidelines

As yet games have not achieved the engagement factor seen in the movie industry, the Holy Grail for any games designer is to engage, excite and attract their players without running into the downside of addiction

There are two major factors to achieving this goal that need to be considered:

1. Does the player "Believe" the game
2. The emotional "roller-coaster" ride experienced by the player

If these are achieved then the game becomes a XXX game. This paper seeks to investigate the factors affecting these two goals and offers guidelines in order to achieve a successful implementation, avoiding the obvious pitfalls in design.

Non Player Characters NPC
One of the important factors when considering the emotions of the player in any game is the role and behavior of the Non Player character (NPC) either the virtual world they inhabit or the way they interact affecting the players belief. What the designer seeks to achieve is a similar belief as that proposed by Alan Turing (Ref) in his test for AI.

Game Turing Tests
Turing argued that should a user be sat at a terminal in conversation via text typed on a keyboard with two participants in a different room, if after and indeterminate period of time, the user could not state which other participant was real and which was the computer generating the conversation then that computer was said to have achieved intelligence and passed the Turing test.

Similarly if the same criteria is applied to games and the player in a multiplayer game cannot tell if they are playing another player or an NPC generated by the game then the game has passed the Game Turing Test and as such has achieved engendering belief in the player.

In order to achieve such a position it is important therefore that the designer understands the fact that as humans we have emotions, and that we have differing perceptions and a developed belief system which all need to be taken into account such that the game design embraces these and doesn't seek to malign or break the system for a successful game to be achieved.

Perception
In games the player utilizes their perceptions to guide their gameplay and thus their emotions . But the player also brings with them their wealth of experiences , doubts, fears and inherent emotions which affect their perception of what they see, hear and encounter during the game. This is much akin to the argument of first person and second person readers as postulated by Eco (1979) and Barthez (1966) .

It is then the role of the game designer to understand this occurs and to attempt to enhance the players experience. However, because the player has preconceptions it is in the authors experience essential that they should ensure that the players mental model remains intact whilst engendering belief in the game world, their own character and other NPC's they encouneter.

Psychology
The role then of the designer is to use psychology in design to bolster the players belief system and to offer a "roller-coaster" ride of interest and excitement thus engaging the player. Unfortunately for the mere mortals this means delving into the murky world of psychology, in the authors opinion psychology appears to be invented in order to confuse the mortal man in that many of the perceptive and behavioural approaches are buried in psychological jargon, this paper hopes to distill some of the wisdom in a more palatable and understandable form.

Human Perception

Many of the players emotional responses and beliefs rely totally on their perceptions, however, perceptions can also get fooled for example Visual perception. There are several classic tests which can fool our small human brain either confusing it or because of the in built nature of the brain seeing things that aren't really there e.g.

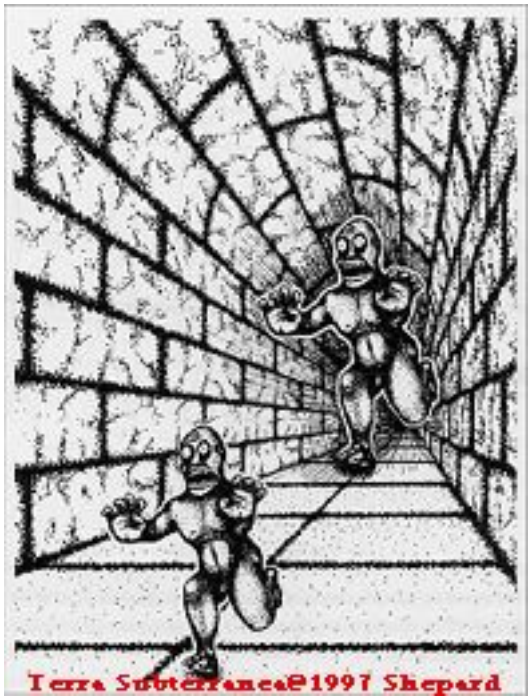


Figure 1: Depth/ Height perception.

Both characters appear different sizes but are in fact the same.

Figure 2: Brain interpretation.

Our brain tells us there is a dog and there is no dog actually there;

Similarly we can have a mental model broken, if words appear a different colour to the actual printing, i.e. imagine if the list below were in different colours than the word says;

Figure 3: Mental Model test.

We become confused, so just as visual perception can fool us into believing 3D for example it can also break our mental model.

But not only visually by sound a similar situation can occur a noise heard in the dark for example or by use of sound and action we can make it appear to rain indoors without anyone getting wet.

Theories of Emotions

As Sousa (2010) says:

"No aspect of our mental life is more important to the quality and meaning of our existence than emotions" (Sousa, 2010)

Human emotions have been the subject of research for many years, although interestingly enough psychologists are still arguing on what basic emotions exist and how they are formed (Scherer 2000; Ekman 1999, Fogel et al 1999, Gergen 1985, Plutchik 1980 and others). However what psychologists do agree on is that there are three components in the formation of emotions;

- Event
- Arousal
- Action

They argue that these three occur but disagree in which order, a similar argument to the chicken and egg. Plutchik (1980) proposed a basic set of emotions;

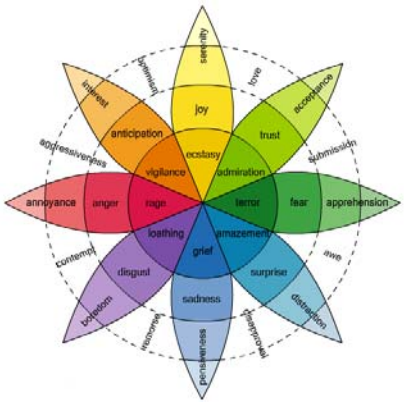


Figure 4: Plutchiks wheel of emotions from Plutchik 1980

Others argue some emotions are hardwired and others learnt i.e.

Table 1: A Selection of Lists of "Basic" Emotions (Ortony & Turner, 1990.

Reference	Fundamental Emotion	Basis for Inclusion
Arnold (1960)	Anger, aversion, courage, dejection, desire, fear, hate, hope, love, sadness.	Relation to action tendencies
Ekman, Friesen, & Ellsworth (1982)	Anger, disgust, fear, joy, sadness, surprise.	Universal facial expressions
Frijda (personal communication, Spetember 8, 1986)	Desire, happiness, interest, surprise, wonder, sorrow.	Forms of action readiness
Gray (1982)	Rage and terror, anxiety, joy.	Hardwired
Izard (1971)	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise.	Hardwired
James (1884)	Fear, grief, love, rage.	Bodily involvement
McDougall(1926)	Anger, disgust, elation, fear, subjection, tender-emotion, wonder.	Relation to instincts
Mowrer (1960)	Pain, pleasure.	Unlearned emotional states
Oatley & Johnson-Laird (1987)	Anger, disgust, anxiety, happiness, sadness.	Do not require propositional content
Panksepp (1982)	Expectancy, fear, rage, panic.	Hardwired
Plutchik (1980)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological processes
Tomkins (1984)	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise.	Density of neural firing
Watson (1930)	Fear, love, rage.	Hardwired
Weiner & Graham (1984)	Happiness, sadness.	Attribution independent

Emotion in games can be viewed from two perspectives i.e.

- The emotions we engender in our player
- The emotions of the NPC and the virtual world experience to make the game more believable

Behavioural Design

First introduced by John Hopson (2001) it uses the players natural behaviours as a guide to producing better games e.g.

- Where we place our save games
- Weapon Spawning
- Degree of challenge

Emotioneering

First coined by David Freeman (2003) this approach deals with how to engender belief and emotional response to NPC's in the game.

Freeman identified 32 categories with over 300 techniques to develop deeper emotions in games and developed a character "diamond" giving each NPC at least 4 different traits. Thus it enables the scriptwriter to develop NPC responses as the "traits" can be manifest in action, dialog or both.

Music in Games

Music can produce powerful emotional ties; everyone remembers where they were when a particular song was played. Equally certain songs are associated with certain times in history and their use can engender memories and ties in the players mind and foster beliefs in the time line of the game.

Accent

In 2006 Purkess et al highlighted the bias of interviewers when the applicant had an accent. BT when researching automated phone systems identified that customers trusted a "Scottish" accent more than any other.

If the NPC looks like a "ruffian" and has the accent of a "ruffian" they are perceived by players to be less trustworthy. In a war game the sergeant has to sound like a sergeant otherwise if he had an effeminate voice the trust is not as strong by the player. Similarly the way the characters dress influences belief.

Belief

The feeling of belief is achieved when the emotions of acceptance and trust are developed. Partly this is based on the player's first preconceptions i.e. *"First Impressions"* and partly it is developed during gameplay. Hence designers need to design for trust and scenarios which can foster or develop trust, although it could also be used deliberately in a game to develop false trust leading to a betrayal for storyline and effect.

Players develop trust in their own character 's abilities and any NPC's they encounter in a number of ways:

- Co-operation with NPC's if they are on the players "team" and they help, support, and importantly "share" the rewards – the player then begins to trust and belief is fostered – similarly distrust can be fostered if the NPC does not do this
- If the players character survives – they start believing
- If the players character beats a boss – they believe in their own characters abilities

Thus the designer can build scenarios to create or fracture trust and belief

Fun and Fear

Two aspects that are becoming more and more important in the emotional design of games are fun & fear

Fun

Fun can be defined as:

" a source of enjoyment" (Koster 2005)

It can happen from a number of sources i.e.

- Physical Stimuli

- Aesthetic appreciation
- Direct Chemical Manipulation

as Koster (2005) puts it:
"all about making our brains feeling good"

There are several factors which influence the feeling of fun in a game design too numerous and detailed for the scope of this paper however some factors according to Freeman (2003) include:

- Combat
  - Being part of a team
  - Melee battles
  - Aiming, targeting, shooting
  - Territorial acquisition
  - Defeating an enemy etc
- Travel
  - Walking, running, driving, flying etc
- Competition and solving puzzles
- Building and creating
- Narrative/story/drama
- Receiving rewards
- Superhuman abilities
- Networking
- Bartering
- Dancing
- Training

Not every game will involve all of these factors but whichever genre the player is using the designer needs to underpin these activities and not cause stumbling blocks in the design

All fun and no substance is as bad as all challenge and no fun. Like a roller-coaster with all ups and no spine-thrilling downs.

The name of the game is balance.

Fear

Many games incorporate fear as a means of excitement, many humans "like" being scared e.g. witness the popularity of the horror movie also the roller-coaster ride such as Oblivion at Alton Towers theme park or Splash Mountain at Walt Disney World. The so-called Adrenalin junkies thrive on these fearful experiences and so it is with games, however there is one difference the fear encountered on these rides is short lived much as a zebra escaping a chase by a lion. In games part of the problem is the difficulty level placed on the scenario causing fear, it can be prolonged if too difficult, this can cause problems. According to latest research the Adrenalin production turns unhealthy if prolonged and can cause problems.

As such it behooves the designer to be careful that such scenes do not precipitate ill health.

Conclusion

Given the foregoing discussions it is becoming increasingly important for the game designer to understand how emotions are formed and developed to ensure a balance in the game and to ensure that the players experience is a roller-coaster ride, not a merry-go-round and to promote the engagement of the player bye suitable design which fosters belief and trust within the player for his own character, the game environment and any NPC's he encounters

References

[1] Barthes R., 1966, Introduction to the structural analysis of the narrative. Communications, (8)

[2] BT, anecdotal research.

[3] Eco U., 1979, Lector in Fabula. Bompiani.

[4] Ekman, P., 1999. Chapter 3. In Dalgleish, T. & Power, M. *Handbook of Cognition and Emotion*. Sussex, Unoited Kingdom: John Wiley & Sons, Ltd.

[5] Fogel, A. et al., 1992. Social Process Theory of Emotion: A Dynamic Systems Approach. *Social Development*, 1(2), pp.122 - 142.

[6] Freeman, 2003, Creating emotion in games the art and craft of emotioneering, New Riders, Oct 2003

[7] Gergen, K.J., 1985. The Social Constructionist Movement in Modern Psychology. *American Psychologist*, 40(3), pp.266-75.

[8] Hopson J, 2001, Bahavioral design, Gammasutra article, http://www.gamasutra.com/features/20010427/hopson_01.htm (Accessed 15 July 2011)

[9] Koster R., 2005, A theory of fun for games design., Paraglyph Press

[10] Ortony, A. & Turner, T.J., 1990. What's Basic About Basic Emotions. *Psychological Review*, 97(3), pp.315-31.

[11] Plutchik, R., 1980. A General Psycholoevolutionary Theory of Emotion. In Plutchik, R. & Kellerman, H. *Emotion: Theory, Research, and Experience*. New York: Academic. pp.3-33.

[12] Purkess et al , 2006, Implicit souces of Bias in eemloyment interview discussions and decisions, Organisatinal behavior and human decision rocesses, 101(2), November 2006, Pages 152-167.

[13] Turing, A.M., 1950, Computing Machinery and Intelligence, Mind 54(236):433-460, October,

[14] Scherer, K.R., 2000. Emotion. In Hewstone, M. & Stroebe, W. *Introduction to Social Psychology: A European Perspective*. 3rd ed. Oxford: Blackwell. pp.151-91.

[15] Sousa, R.d., 2010. Emotion. In E.N. Zalta, ed. *The Standford Encyclopedoa of Philosophy*. 2010th ed.

From Grid Computing to Cloud Computing & Security Issues in Cloud Computing

Rajendra Kumar Dwivedi

Assistant Professor (Department of CSE), M.M.M. Engineering College, Gorakhpur (UP), India
E-mail: rajendra_bhilai@yahoo.com

Abstract
The cloud is a next generation platform that provides dynamic resource pools, virtualization and high availability. In starting the concept of the Distributed Computing and the Grid Computing is discussed. Then I have focused on the concept of Cloud Computing and its characteristics. This paper provides a brief introduction to the Cloud Computing platform and the services provided by it. This paper also contains some information regarding Cloud Storage. I have also given the comparison of the Cloud Computing with the Grid Computing. The prevalent problem associated with the Cloud Computing is the Cloud Security and the appropriate implementation of the Cloud over the network. I have discussed the policy, software and hardware security issues. Access Control and the use of the Digital Signature to enhance the data security of the Cloud is also discussed in this paper in short.

Keywords: Distributed Computing; Grid Computing; Cloud Computing; IaaS; Paas; SaaS; HaaS; StaaS; Virtualization; Security Issues

Introduction

Cloud Computing is the development of parallel computing, distributed computing and the Grid computing. It works on the idea to make many normal computers together to get a super computer which can do a lot of things. Cloud computing is a new mode of business computing and it will be widely used in near future. However, there still exist many problems in cloud computing today. Data security and privacy risks have become the primary concern for people to shift to cloud computing.

Grid Computing

It is a form of distributed computing and refers to the use of several computers to solve a single problem at the same time. It uses networked, loosely coupled computers acting simultaneously to perform very large tasks. In Grid computing, a program is generally divided into many parts and these parts are allocated to several computers, often up to many thousands. Grid computing is used to solve scientific and technical problems which require a large amount of computing or access to large amount of data. Grid computing can be thought of as a distributed parallel processing system. Functionally, grids can be classified into computational grids and data grids. Computational Grids focus primarily on computationally intensive operations and Data Grids control sharing and management of large amount of distributed data.

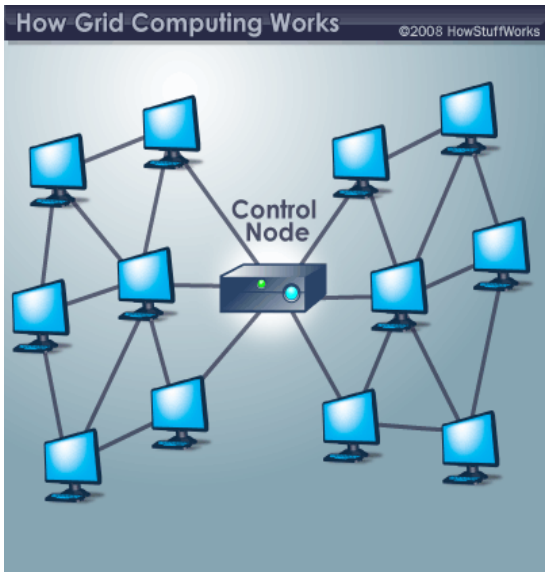


Figure 1: Grid Computing.

Characteristics of Grid Computing

1. Grid is formed using Heterogeneous computers (different O.S. and hardware)
2. Grid is formed using Loosely coupled machines (grids are distributed in nature over a network)
3. Grid uses geographically scattered machines (not in a single location)
4. Resource handling is done by resource manager at each node as an independent unit

Cloud Computing

Cloud is a virtualized pool of computing resources. It can rapidly deploy and increase workload by speedy providing physical machines or virtual machines. It is used for delivering hosted services over the internet. It is a style of computing where dynamically scalable and often virtualized computing resources are provided as a service over the internet. In many cases, cloud computing services provide common business applications online that can be accessed using a web browser while storing software and data on the servers.

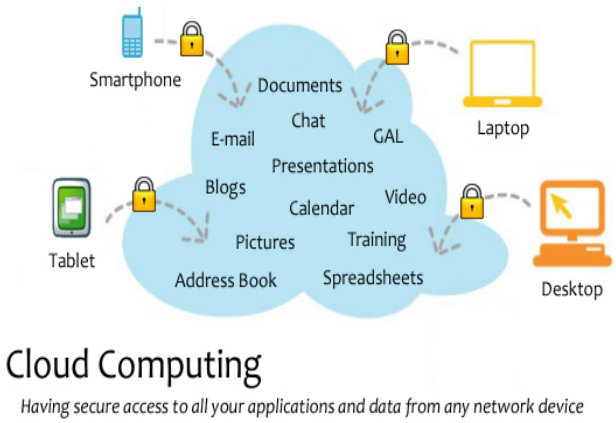


Figure 2: Cloud Computing.

Types of Cloud

1. Public (External) Cloud
2. Private (Internal) Cloud
3. Hybrid (Combined) Cloud

Characteristics of Cloud Computing

1. Dynamically scalable and elastic in nature result of continuous evolution of data management technology
2. Users pay only for the required capacity provides secure and dependable data storage center, so user need not to think about storing data and killing the viruses
3. It does not need user's high level equipments, so reduces user's cost.
4. Fully managed by service providers, so users need not know how the cloud runs.
5. High reliability
6. High extendibility
7. Extremely inexpensive
8. On demand service
9. Versatility
10. Virtualization

Cloud Computing Examples

1. Amazon EC2(Elastic Compute Cloud)
2. GoogleApps
3. IBM's Blue Cloud
4. Yahoo
5. Microsoft
6. Zoho
7. Mosso
8. Salesforce
9. GoGrid
10. ElasticHosts

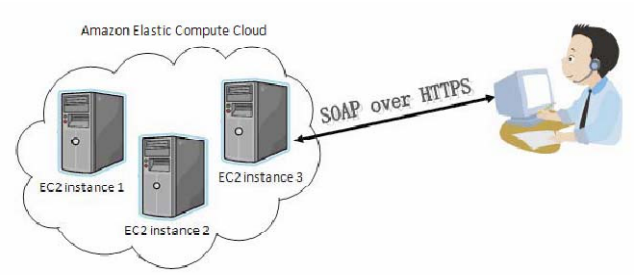


Figure 3: Usage of Amazon Elastic Compute Cloud.

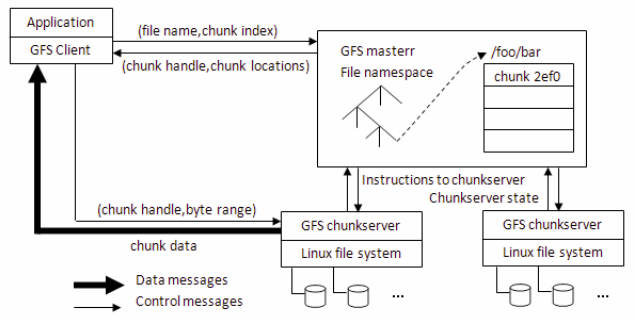


Figure 4: Google File System architecture.

We can understand the technologies used in these examples, if we refer their architectures. See the above figures of Amazon EC2 and Google File System Architecture.

Cloud Storage

It is a model of networked computer data storage where data is stored on multiple virtual servers, in general hosted by third parties, rather than being hosted on dedicated servers.

Cloud storage system Architecture

Look at the following fig to understand a typical cloud storage system Architecture

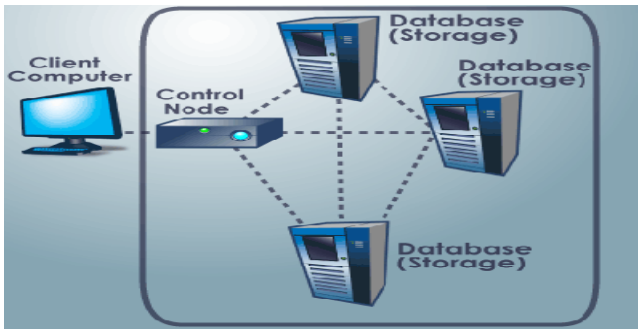


Figure 5: A typical cloud storage system Architecture.

Evolution of cloud storage

To understand this evolution look at the following figure:

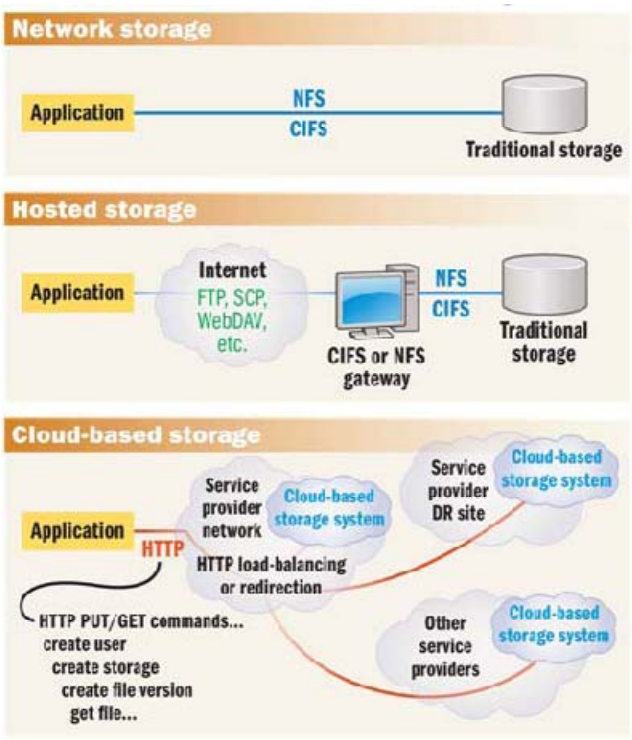


Figure 6: evolution of cloud storage.

Services & Layers

Services

There are maily 3 services as mentioned below :

1. *SaaS (software as a service)* - Service user can use the service from anywhere. Ex- Yahoo
2. *PaaS (platform as a service)* –Developers can create applications on the service provider's platform. Ex- Amazon web service
3. *IaaS (infrastructure as a service)* – It provides virtual server instance and block of storage on demand. Ex- GoogleApps.

Some authors termed 2 more services as mentioned below:

1. *HaaS (Hardware as a service)*
2. *StaaS (Storage as a service)*

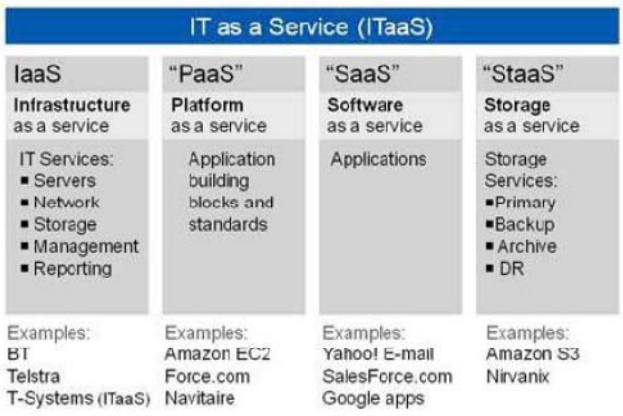


Figure 7: Cloud Computing Services.

Layers

Client
Application
Platform
Infrastructure
Server

Figure 8: Layers in Cloud Computing Model.

There are 5 layers in Cloud Computing Model as discussed below:

1. *Client* : consists of software and hardware that relies on cloud computing for application delivery
2. *Application* : deliver software as a service (SaaS) over the internet.
3. *Platform* : deliver platform as a service (PaaS) over the internet.
4. *Infrastructure* : deliver infrastructure as a service (IaaS) over the internet.
5. Server : consists of software and hardware that are specifically designed for the delivery of cloud services.

Cloud Computing Platforms

I am discussing the following 4 Cloud Computing Platforms and the comparison of these platforms for several characteristics –

1. Abicloud
2. EUCALYPTUS (Elastic Utility Computing Architecture for Linking Your Programs To Useful Systems)
3. Nimbus
4. Open Nebula

Table 1: The comparison of several cloud computing platforms

	Abicloud	EUCALYPTUS	Nimbus	Open Nebula
Cloud Character	Public/ Private	Public	Public	Private
Scalability	Scalable	Scalable	Scalable	Scalable
Cloud Form	IaaS	IaaS	IaaS	IaaS
Compatibility	Can manage EC2	Support EC2	Support EC2	Open, Multiplatform
Deployment	Pack and redeploy	Dynamical deployment	Dynamical deployment	Dynamic deployment
Deployment manner	Web interface drag	command line	command line	command line
Transplantability	Easy	common	common	common

VM Support	Xen, VMware	Xen, VMware	Xen	Xen, VMware
Web Interface	Libvirt	Web service	EC2 WSDL, WSRF	Libvirt, EC2, OCCI API
Structure	Open platform encapsulate core	Module	Light weight components	Module
Reliability	–	–	–	Rollback host and VM
O.S. Support	Linux	Linux	Linux	Linux
Development Language	C++, Python	JAVA	JAVA, Python	JAVA

Comparison of Cloud Computing & Grid Computing

These two computings can be compared on the basis of their charactersics.The following table shows the comparison of these two computing.

Table 2: The comparison of Cloud and Grid Computing

Characteristic	Cloud Computing	Grid Computing
Service Oriented	Yes	Yes
Loose Coupling	Yes	Half
Strong fault tolerant	Yes	Half
Business model	Yes	No
Ease of Use	Yes	Half
TCP/IP Based	Yes	Half
High security	Half	Half
Virtualization	Yes	Half

Security Issues

Cloud Computing Issues

There are several issues in Cloud Computing. The main issue is the "Security Issue in cloud computing". My main focus is on Security Issues. First of all, let us know that what are the common issues in cloud computing. Some of these cloud computing issues are given below-

1. Security
2. Privacy
3. Reliability
4. Legal Issues
5. Open source
6. Open standard
7. Compliance
8. Freedom
9. Long term viability
10. Availability and Performance
11. Data Usage
12. Sustainability and siting
13. Use by crackers

Security Issues in Cloud Computing

There are several security issues in Cloud Computing. Some are given below-

Policies

1. *Inside Threats*- good supervision should be done for having trusted employees
2. *Access Control*-Digital signature can be implemented for access control.
3. *System Portability*- The problem of vendor lock-in should be handled . (**vendor lock-in:** If a company is dissatisfied with one cloud computing service- or if the vendor goes out of business- the firm can not easily and inexpensively transfer these services to another provider or bring it back in-house.)

Software Security

1. *Virtualization technology*-up to date version of virtualization product should be installed for the security reasons.
2. *Host Operating System*- should be up to date and secure from hackers
3. *Guest Operating System*- should be up to date and secure from hackers
4. *Data Encryption*-should be done on all the data for its safety.

Physical security

1. *Backup*- Either a backup plan should be provided automatically for each customer, or they can use the plans provided elsewhere in the cloud.
2. *Server Location*-It should be at appropriate place. Room should have adequate space and isolated. A Cooling System and Fire Suppression System should be installed there.
3. *Firewall*- Cloud Computing service providers should provide a complete firewall solution to their clients.

Conclusions & Future Work

Cloud Computing announced a low-cost super computing services to provide the possibility, while there are a large number of manufacturers behind, there is no doubt that cloud computing has a bright future. In future we can work on many issues like-

1. Security of cloud platform and data in transmission
2. Interoperation and standardization
3. Consistency guaranty
4. Continuously high availability
5. Dealt mechanisms of cluster failure in cloud environment
6. Synchronization in different clusters in cloud platform

References

[1] Shufen Zhang,Shuai Zhang,Xuebin chen, Shangzhuo Wu,"Analysis and Research of Cloud computing system Instance", 2010 second international Conference on Future Networks, IEEE.

[2] Junjie,xuejun,Zhou "Comparison of several cloud computing platforms",second International Symposium on Information science and engineering,

2009 IEEE

[3] Jiyi WU,Lingdi,Xiaoping GE,Ya Wang,Jianqing, "Cloud storage as the infrastructure of cloud computing", 2010 International Conference on intelligent Computing and Cognitive Informatics, IEEE

[4] Shuai,Shufen,Xuebin,Xiuzhen, "The Comparison Between Cloud Computing and Grid Computing", 2010 International conference on Computer application and system modeling (ICCASM2010), IEEE

[5] Eystein Mathisen, "Secuity challenges and solutions in cloud computing", 5th IEEE International conference on Digital Ecosystem and technologies (IEEE DEST 2011), 31May-3June2011,Daejeon,Korea

[6] Jianfeng,Zhibin, "Cloud computing Research and security issues",2010 IEEE

[7] S. Ghemawat, H. Gobioff and S. Leung. The Google file system. In Proceedings of the 19th ACM Symposium on Operating Systems Principles, pages 29-43,2003.

[8] Aymerich, F.M. Fenu, G.Surcis, S. An approach to a Cloud Computing network. Applications of Digital Information and Web Technolgies, 2008

[9] K. Keahey and T. Freeman, "Science Clouds: Early Experiences inCloud Computing for Scientific Applications," in proceedings of Cloud Computing and Its Applications 2008, Chicago, IL. 2008.

[10] D. Nurmi, R. Wolski, etc., "The Eucalyptus Open-source Cloud computing System," in Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, Shanghai, 2009, 124-131.

[11] B. Sotomayor, K. Keahey, I. Foster. Combining Batch execution and Leasing Using Virtual Machines, HPDC 2008, Boston, MA, 2008, 1-9

[12] Hayes, B.: Cloud computing. Communications of the ACM. 51 (7) (2008).

[13] Vouk M.A.: Cloud Computing - issues, Research and Implementation Journal of Computing and Information Technology. CIT 16, 4 (2008) 235–246.

[14] Tharam Dillon, Chen Wu, Elizabeth Chang, 2010, 24th IEEE International Conference on Advanced Information Networking and Applications ,"Cloud computing: issues and challenges".

[15] Elinor Mills, January 27,2009. "Cloud computing security forecast: clear skies".

[16] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, [2005] "Live migration of virtual machines" In Proc. Of NSDI'05, pages 273-286, Berkeley CA, USA, 2005. USENIX Assiciation.

Comparative Study of Pattern Recognition: Neural Networks Approach V/S Algorithmic Approach

Namrata Aneja

Dyal Singh College, Karnal-132001, India

Abstract
There is a great scope of expansion in the field of Neural Network, as it can be viewed as massively parallel computing systems consisting of an extremely large number of simple processors with many interconnections. NN models attempt to use some organizational principles in a weighted directed graphs in which nodes are artificial neurons and directed edges are connections between neuron outputs and neuron inputs. The main characteristic of neural network is that they have the ability to learn complex non- linear input output relationships. A single artificial neuron is a simulation of a neuron (basic human brain cell) and scientists have tried to emulate the neuron in a form of artificial neuron called perceptron. Pattern recognition is one of the areas where the neural approach has been successfully tried. This study is concerned to see the journey of pattern recognition from algorithmic approach to neural network approach.

Introduction
Pattern recognition is a field in computer studies that focuses on how machines interact with the environment, discover the different patterns according to a given rule and provide the desired output regarding the patters (Jain et al, 2000). A pattern in this case is a vaguely defined entity that bears a name such as a human face, or the image of a fingerprint. Pattern recognition may be either supervised or unsupervised. In supervised classification, the input pattern forms part of an already defined class while in unsupervised classification the pattern are assigned to an undefined category. Pattern recognition has a wide range of application including financial planning, data organization and recovery in multimedia databases and personal identification in biometrics (Murty & Devi, 2011).

Three important aspects of pattern recognition that must be considered in designing pattern recognition systems include data input and preprocessing, representation of data and decision-making (Jain et al, 2000). The specific approaches and models for these aspects depend on the nature of the problem and how properly the problem is defined. When the problem is precisely defined, the recognition patter is likely to be more compact, resulting in a less sophisticated decision making technique. There are several methods applied in pattern recognition including neural networks, syntactic recognition, statistical recognition, and template matching (Jain et al, 2000,). Apart from the neural network approach the other approaches are algorithmic, meaning they rely on programmed instructions to carry out their tasks (Liu et al, 2007,). This paper will compare and contrast the algorithmic and neural network approaches to pattern recognition.

Algorithmic Approach Versus Neural Network Approach
To start with here is an Bruti Force Algorithm which is first Pattern matching algorithm , Here we compare a given pattern with the given substring

text S by trying each position one at a time.
Wk= Substring(S, k, Length (P))
Here S is a string
W is a substring
P as length of substring
K is beginning with kth character.
Minimum value of k=1.
Maximum value of k=Length(S)-Length(p)+1

First find W and compare S character by character. If all the characters are the same pattern matching is complete

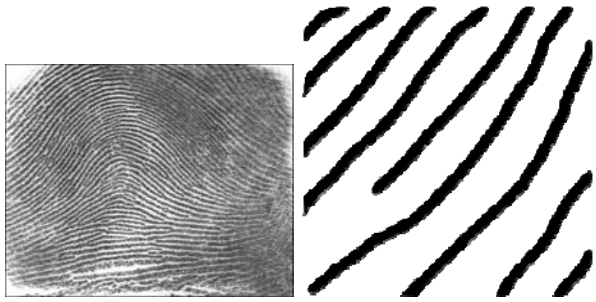


Figure 1

Similarly, in case of finger tips as given in fig.1 one print is stored in the memory in the form of 0 and 1. If the pixel is black then it is 1 else it is 0. When the new finger tip comes across it is matched pixel by pixel on the basis of two dimensional matrix algorithms. If the pixel is black it gives 1 else 0. Then both the 0's and 1's tables are matched .if they are similar then finger tips are matched thus pattern recognition successful else unsuccessful.

Pattern recognition using neural networks via Boltzmann Machine(BM).The BM is a hetroassociative, pattern matcher that stores arbitrary spatial patterns (Ak,Ck),k=1,2,….,m,

Here the kth pattern pair is represented by the vectors Ak= (a1k,…ank) and ck=(c1k…,cqk).

The BM is represented by the 3 layer topology.

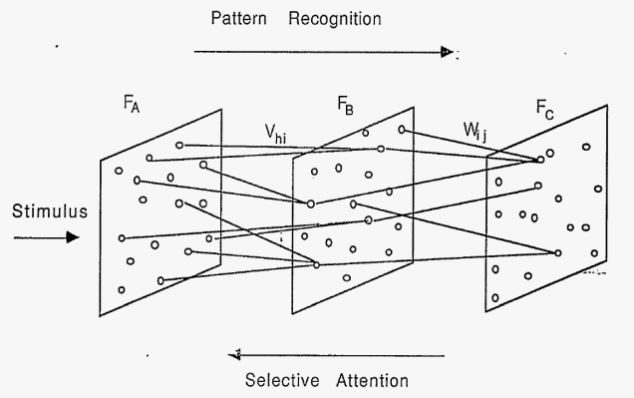


Figure 2

Shown in fig 2. Where the Fa cells correspond to Ak's components and fc corresponds to ck components. The FA to FB interlayer connections are represented by Vhi, and all the FB to Fc inter-layer connections are indicated with wij.

Let F be the set of all input patterns, f, e.g. f = Ak. Let T be the parameter space of a family of BM, e.g. t = T, where t = p(Vhi, Wij). Therefore, given an input pattern f =F, one can find a BM such that the equilibrium distribution of this BM is the given pattern f. Therefore, the input pattern, f, is encoded into a specification of a BM, t. This mapping from F (pattern space) to T (parameter space of BM) defines the BM-transformation. BM-transformation encodes an input pattern into a relatively small vector which catches the characteristics of the input pattern. The internal space T is the parameter space of BM. The internal change rule of our system is defined by a formula.

$\Delta Ei=\sum vhibi+\sum wjibi$

A stimulus pattern is presented to the first stage of the forward paths, the input layer FA, which consists of a two-dimensional array of receptor cells. The second stage of forward paths is called the hidden layer Fb The third stage of the forward paths is the recognition layer Fc. After the process of learning ends, the final result of the pattern recognition shows- in the response of the cells of FA. In other words, cells of the recognition layer Fc work as diagnostic cells; usually one cell is activated, corresponding to the category of the specific stimulus pattern. Cells in BM are featuring-extracting cells. Connections converging to these cells are variable and reinforced by learning (or training). After finishing the learning, those cells can extract features from the input patterns. In other words, a cell is activated only when a particular feature is presented at a certain position in the input layer. The features extracted by the cells are determined by the learning process. Generally speaking, local features, such as a line at a particular orientation, are extracted in the lower stages. More global features, such as part of a training pattern, are extracted in higher stages. Finally ,each cell of the recognition layer at the highest stage integrates all the information of the input pattern; each cell responds to only one specific pattern. In other words, only one cell, corresponding to the category of the input pattern, is activated. Other cells respond to the patterns of other categories.

To better understand artificial neural computing it is important to know first how a conventional 'serial' computer and it's software process information. A serial computer has a central processor that can address an array of memory locations where data and instructions are stored. Computations are made by the processor reading an instruction as well as any data the instruction requires from memory addresses, the instruction is then executed and the results are saved in a specified memory location as required. In a serial system (and a standard parallel one as well) the computational steps are deterministic, sequential and logical, and the state of a given variable can be tracked from one operation to another. In comparison, ANNs are not sequential or necessarily deterministic. There are no complex central processors, rather there are many simple ones which generally do nothing more than take the weighted sum of their inputs from other processors. ANNs do not execute programmed instructions; they respond in parallel (either simulated or actual) to the pattern of inputs presented to it. There are also no separate memory addresses for storing data. Instead, information is contained in the overall activation 'state' of the network. 'Knowledge' is thus represented by the network itself, which is quite literally more than the sum of its individual components.

The main difference between the neural network and conventional approaches lies in the way they tackle the problem of pattern recognition (Minin, 2006,). The algorithmic approach used by conventional computers involves a host of instructions, translated into digital signals, which the computer follows to arrive at the solution. As such, the ability of the computer to solve a problem is limited to the programmer's understanding of the problem as well as their capability to solve it. This is one of the limitations of algorithmic approaches since they cannot solve problems that the programmer vaguely understands (Minin, 2006,).

Neural networks on the other hand, are capable of using analog signals and imitate the human brain in information processing (Ripley, 2007). The functional elements of a neural network are small processors called neurons that work hand in hand to solve a given problem (Ripley, 2007,). Neural networks are similar to the brain because they acquire knowledge to solve problems through learning as opposed to algorithmic approaches, which rely on programming. Another similarity of neural networks with the human brain is that information is stored in synaptic weights achieved through interconnection of neurons (Ripley, 2007,). The ability to learn from examples is one of the key features that make neural network approach unique. Hence, the examples chosen must be appropriate for the specific problem to be solved to minimize time wastage and network malfunction. This is particularly important since neural networks are not programmable. Since the network handles the problem independently, it is impossible to predict the outcome (Minin, 2006,).

On the contrary, the operation of conventional computers using algorithmic approaches is predictable. This is because algorithmic approach is based on a set of precise instructions, written in machine language that the computer can interpret and follow the instructions to arrive at the desired solution (Stergiou & Siganos, 2011). Hence, the user can predict how the computer will deal with the problem. Neural networks are

extremely useful in pattern recognition because of their ability to discover trends and patterns that algorithmic approaches cannot detect. When carefully trained, neuron networks can be regarded as experts in their specific areas of application hence can be relied on to predict future trends and answer complex logical questions (Minin, 2006).

Some of the strengths that neural networks possess include ability to retain the knowledge obtained through experience and adapt it to solve similar problems in future, excellent information organization skills, and real time operation achieved through parallel information processing. Additionally, the networks can still function even after partial destruction (Minin, 2006).

Despite these differences, both algorithmic and neural network approaches are complementary to each rather than being competitors. This is because some problems such as those involving calculations are best done using algorithmic approaches while others require neural networks (Stergiou & Siganos, 2011). In addition, both approaches overlap especially in the underlying theory. Some of the procedures used in neural networks were derived from statistical pattern recognition technique. Largely both methods are branches of statistics (Petridis & Kehagias, 1998).

Different categories of neural networks are available. The feed -forward and the Radial-Basis Function networks are the most common in pattern classification. The networks are arranged in layers with one-way connections between them (Jain et al, 200). Other networks in this category include the Kohonen-Network and the self –Organizing Map. Neural networks have gained popularity in pattern recognition because they do not depend much on the availability of specific knowledge for solving specific problems in pattern recognition as other methods do. Instead, the networks modify the knowledge learnt to suit the problem at hand and make decisions accordingly (Jain et al, 2000,).

The conventional approaches to pattern recognition are either parametric or non-parametric, in which the models applied are either fixed with few parameters or flexible respectively. Neural networks, on the other hand, insist on moderation of the two extremes, preferably models that are highly flexible but with limits and having a substantial number of parameter (Ripley, 2007).

There are many advantages and limitations to neural network analysis and to discuss this subject properly we would have to look at each individual type of network, which isn't necessary for this general discussion. In reference to backpropagational networks however, there are some specific issues potential users should be aware of…….??? Depending on the nature of the application and the strength of the internal data patterns you can generally expect a network to train quite well. This applies to problems where the relationships may be quite dynamic or non-linear. ANNs provide an analytical alternative to conventional techniques which are often limited by strict assumptions of normality, linearity, variable independence etc. Because an ANN can capture many kinds of relationships it allows the user to quickly and relatively easily model phenomena which otherwise may have been very difficult or impossible to explain otherwise.

Conclusions

Neural network approach to pattern recognition is the new area of research in computer science due to its potential to make the computer more useful in solving complex problems in pattern recognition. The main difference between this method and the conventional algorithmic approaches such as the syntactic and statistical methods is its ability to process information in a manner similar to the human brain. While algorithmic approaches require programming, neural networks operate without human intervention once the learning process is complete. The ability of neural networks to learn from examples and adapt the knowledge to solving numerous problems give this approach an edge over the algorithmic methods in discovering patterns that humans may not notice. Despite these differences, the two approaches complement each other. While algorithmic approaches are more useful in detecting arithmetic relationships, neural networks are more applicable in solving less analytical problems in pattern recognition.

Acknowledgment

The literary help and support rendered by Sh. Girish Aneja CA …… and Dr R K Bhardwaj Associate Professor Dyal Singh College Karnal is fully acknowledged

References

[1] Jain, K. A., Duin R. P. W. & Mao J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(1), 4-37.

[2] Liu, D., Fei S., Hou Z., Zhang H. & Sun C. (2007). *Advances in neural networks-ISNN: 4th international symposium on neural networks, ISNN 2007 Nanjing, China, June 3-7 proceedings, part 3.* New York, NY: Springer

[3] Minin, A. (2006). The neural network analysis. Retrieved July 13, 2011 http://www14.informatik.tu-muenchen.de/konferenzen/Jass06/courses/2/presentations/minin_handout.pdf

[4] Murty, N. & Devi S. (2011). *Pattern recognition: An algorithmic approach*. New York: Springer.

[5] Petridis, V., & Kehagias, A. (1998). *Predictive modular neural networks: Applications to time series*. Boston: Kluwer Academic Publishers.

[6] Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

[7] Stergiou, C. & Signos D. (n.d). Neural networks. Retrieved July 13, 2011 http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#Neural%20networks%20versus%20conventional%20computers

Security Aspects in Multimedia

Srawan Nath[1], Richa Rawal[2], Ruchi Dave[3] and Naveen Hemrajani[4]

[1]M.Tech(SE) Student, SGVU, Jaipur, India
E-mail: nath.srawan@gmail.com
[2]M.Tech(CS) Student, SKIT, Jaipur, India
E-mail: richarawal_23@yahoo.co.in
[3]Associate Professor, SGVU, Jaipur, India
E-mail: ruchi.davey@gmail.com
[4]Associate Professor, SGVU, Jaipur, India
E-mail: naven_h@yahoo.com

Abstract
In the today's world, various media types such as text, audio, image and video entered in network arena. Reasons for this is sharing and acquiring the data in a quicker period of time and this is amply supported by the considerable cheap availability of resources like high bandwidth. This step raised many challenges opposing the former's success. One of the most challenging issues is the lack of security of the data. Many techniques in digital watermarking were developed, to counter the challenges, but there is none of the single approach that solves the security problems for all the media types. There are some of the methods are suggested which meets the above mentioned demand. They are spread spectrum watermarking, non-repudiation oblivious watermarking and attack characterization. The first technique is about embedding the watermark cannot be removed imperceptibly, whereas the second one proposes watermarking schemes for data distribution and the last one improves the robustness of the watermark by characterizing the attack using reference watermark.

Keywords: Digital Watermarking, Spread Spectrum Watermarking, Security, Non-repudiation Oblivious Watermarking

Introduction

The well acceptance of multimedia communications has brought about a number of new concerns to IT managers and system administrators, ranging from network performance aspects, copy-right and intellectual property aspects and security concerns. As a result concerns regarding security, scalability and manageability of existing systems become more acute as current solutions may not satisfy the demands of multimedia communications. Information has become the source of billion dollar investment and multibillion dollar income. Enforcing proper security measures on the information has therefore become vital .it is following this idea that many security algorithms has mushroomed to the occasion in a considerable quicker period of time.

Recently in the field of multimedia there is an increased requirement for security, due to various threats which includes replication of digital data without any information loss, and manipulation of the same without any detection .as the utility and usage of the internet has grown considerably. Multimedia documents are easily copied throughout authorized and illegal channels, it is following this reason that the authors of the work hesitate to publish their work electronically, fearing the threat that it can be pirated easily. Once the information or the content is acquired by miscreants, it can be easily modified by employing some specific software tools and further, can be claimed by them as their own work which is definitely an unbearable act. For this requirement of security solutions needed. Security solutions are especially of interest for such fields as distributed production processes and electronic commerce, since their producers provide only access control mechanisms to prevent misuse and theft of material. This paper includes security criteria with in different classifications regarding the following three basic threats: threat of confidentiality, integrity, availability.

Requirements and Measures

The basic requirement for the security of a given multimedia system must be derived. Security requirements are met by security measures, which generally consist of several security mechanisms. Security services can be implemented by security mechanisms. Overall, security requirements are described within a security policy. The security policy defines also which measures are implemented to realize these requirements.

The security requirements can be met by the following security measures:

Confidentiality: Cipher systems are used to keep information secret from unauthorized entities.

Data Integrity: The alteration of data can be detected by means of one-way hash functions, message authentication codes, digital signatures, fragile digital watermarking, and robust digital watermarking.

Data origin authenticity: Message authentication codes, digital signatures, fragile digital watermarking and robust

digital watermarking enable the proof of origin.

Entity authenticity: Entities taking part in a communication can be proven by authentication protocols. These protocols ensure that an entity is the one it claims to be.

Non-repudiation: Non-repudiation mechanisms prove to involved parties and third parties whether or not a particular event occurred or a particular action happened.

The event or action can be the generation of a message, the sending of a message, the receipt of a message and the submission or transport of a message. Non-repudiation certificates, non-repudiation tokens, and protocols establish the accountability of information. These mechanisms are based on message authentication codes or digital signatures combined with notary services, time stamping services and evidence recording.

Security Practices

Security measures are based on modern cryptographic mechanisms as well as on security infrastructures. The way in which the discussed security mechanisms can be applied to multimedia resource is difficult to analyze, due to the complexity of multimedia data. Cryptography has been widely used for access control mechanism s to prevent misuse and theft of the material. Though the theft can be prevented the ease of with which perfect copies can be made without any information loss facilities unauthorized copying of multimedia documents like music and film in a large scale . The situation demanded for the need of a better technique that would prevent copyright violation after authentication of the legitimate customer. Digital watermarking flourished for meeting the above stated demands.

Cryptography

Cryptography is a practice of converting the data that is to be transmitted over the network medium, into an un-recognizable format. This is usually achieved by employing computation of functions ranging in various levels of complexities. Apart from that, many of the modern cryptographic mechanisms rely on few unproven assumptions too. The complexity of the crypt-function, determines the security level and the reliability. The trade-off for the complexity of the function is the time period, i.e. more complex the procedure, more time it takes in encrypting and decrypting the data. Though the resources available in modern system are sufficiently large, it cannot resolve the hard computations in an acceptable period of time. Therefore, though solution for achieving thorough security is possible, it is hardly reinforced owing to the reasons stated above.

The Cryptosystems by themselves are further subdivided into private-key and public-key cryptosystems. The private-key cryptosystems are the ones in which the communicating entities share a common secret key K, which is supposed to be kept secret. The encryption uses this key to encode the message before distributing it, while the same can be decoded using the same key, available at the receiver end. The length of the key is directly proportional to its chances of not being cracked. This key need not necessarily be a secret one. The other entity who knows the trapdoor information, can decipher the code using his private secret key. It is analyzed that the computation of Private Key from Public Key is computationally not viable.

For the media data in particular, most of the time it is required that each bit of data of the original signal be taken as the input for the cryptosystem in calculating the encrypted data as output. But when these data are distributed over the network, due to errors in the channel, the data bits may get modified or garbled, though slightly. But this modification is far more than sufficient in upsetting the cryptosystem. It is therefore essential that such errors do not affect the cryptosystem as such, and for that to happen, the method such as the one explained above, should be avoided though it is efficient.

To overcome all of the above stated challenges and yet to maintain maximum confidentiality, one can use digital watermarking in combination with the cryptosystems, which can be used as a supplementary system to the former. Digital watermarking systems as such are very robust and nearly fool-proof.

Digital Watermarking

Digital watermarking techniques based on steganographic systems offer the possibility to embed information directly into the media data. Besides cryptographic mechanisms, watermarking represents an efficient technology to ensure both data integrity and data origin authenticity. Watermarking techniques usually used for digital imagery and now also used for audio and 3D-models, are relatively young and their amount is growing at an exponential rate. Copyright, customer or integrity information is embedded, using a secret key, into the media data as transparent patterns. Because the security information is integrated into the media data, one cannot ensure confidentiality of the media data itself, only the security information using a secret key.

Based on application areas for digital watermarking, the following five watermarking classes are defined: authentication watermarks fingerprint watermarks, copy control watermarks, annotation watermarks, and integrity watermarks. The most important properties of digital watermarking techniques are robustness, security, imperceptibility / transparency, complexity, capacity, and the possible verification procedure.

Techniques

Text Watermarking

Watermarking of text documents provides a means of tracing documents that have been illegally copied, distributed, altered, or forged. Brassil*et al* have investigated text watermarking and proposed a variety of methods for embedding hidden messages in PostScript documents. The work of Brassil*et al.* currently does not use SS embedding, but it could be added to the system to strengthen robustness and security. Line shifting moves entire lines of text up or down by a small amount, typically 1/150 or 1/300 inch (0.170 or 0.085 mm). Similarly,

word shifting may horizontally shift individual words or blocks of words; words at the ends of a line are not shifted to preserve justification. Recovery of the message from a printed or photocopied document requires a number of post-processing steps (scanning, skew correction, and noise removal). After post-processing, the message receiver automatically measures line shifts, word shifts, and/or feature alterations to detect the message.

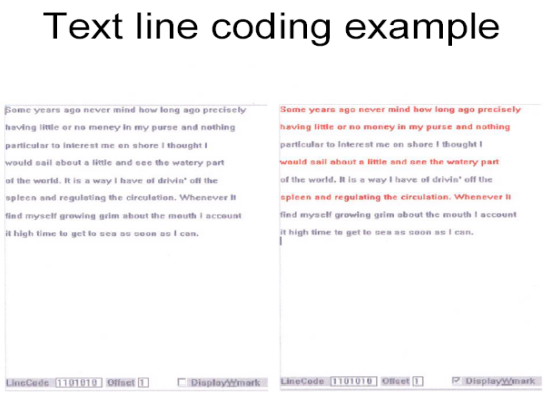


Figure1.0: Text line coding example.

Digital Audio Watermarking

Digital audio watermarking involves the concealment of data within a discrete audio file. To combat online music piracy, a digital watermark could be added to all recording prior to release, signifying not only the author of the work, but the user who has purchased a legitimate copy. Newer operating systems equipped with digital rights management software (DRM) will extract the watermark from audio files prior to playing them on the system. The DRM software will ensure that the user has paid for the song by comparing the watermark to the existing purchased licenses on the system. Watermarking could be used in voice conferencing systems to indicate to others which party is currently speaking. Also, watermarking technology include embedding auxiliary information which is related to a particular song, like lyrics, album information, or a small web page, etc.

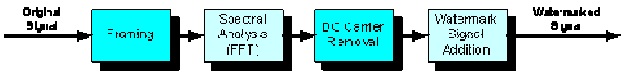


Figure 2.0: Watermarking Insertion Process.

Modern Approaches

Spread Spectrum Approach

For robust watermarking, it is essential that the distortions and tampering of the marked signal doesn't really effect changes in the embedded watermark. One of the effective methods for achieving this demand is by watermarking in the particular frequency domain of the signal. This method is called as frequency-based method of watermarking and is achieved using spread spectrum techniques. The signal when traveled through any medium undergoes few changes. The first and foremost one of them is the due to the effect of noise on the channel. Further, channel codes and standard encryption procedures may slightly modify the contents of the data, leading to partial degradation of the content, though they are generally information lossless techniques. The other sources of partial or complete data destruction are lossy compression, affine transformations, and other common signal distortions like re-sampling, conversion, re-quantization, etc.

If one wish to add resilience to the watermark against any of such attacks or operations, the watermark must be placed in the perceptually significant regions of the signal, i.e. it must not be placed in regions of least importance of the signal, as these are the components which are most likely to get eliminated by the lossy compression algorithms and other geometric processes. Further, it should not be placed in regions of high frequency owing to probable data loss among these components due to affine transformations like affine scaling and cropping. Though such affine transformations results in irreversible data loss in most of the spatial watermarking techniques, its impact is hardly felt in frequency-based schemes.

Non-Repudiation Oblivious Watermarking

Digital watermarking method finds its usage in audio, image, video and multimedia data for the purpose of ensuring security by maintaining the confidentiality and data integrity of the contents. It is also required that non-repudiation in the media document is effected during the distribution process, so that the end users who have copied or acquired the document illegally can be identified and proven. To achieve this, a technique that comes to the fore is Non-repudiation Oblivious watermarking, whose usage can be highly felt in the distributed environment for multimedia data.

Non-repudiation oblivious watermarking is a technique for creating undeniable watermarks. The owner of the data distributes his contents to N number of distributors, who act as agent's in supplying the content to the clients after suitably watermarking them using their own watermark key. The content provider or the owner will also be able to identify which distribution agent watermarked the content. In this system, it is made sure that a particular distributor does not watermark the content that would appear to have been watermarked by another distributor using the same watermark key, for the same content. By this way, any locations where the pirated copies of the media are released can be detected.

Future Enhancements

The large number of network multimedia systems dictates the need for copyright protection of digital property. To conclude, any successful watermarking algorithm would have to exploit properties of the human visual system and combine these with effective modulation and channel coding. Future work will concentrate on producing watermarks that are robust to filtering, lossy image compression, noise corruption and changes in contrast. In addition these algorithms must anticipate possible attacks on the integrity and security of the watermark and to devise suitable countermeasures. An increase in high speed internet connections has also prompted the motion picture industry to take note of possible revenue

losses due to unauthorized movie distribution via the internet. Microsoft is currently developing new watermark technologies and is in the process of testing future operating systems equipped with DRM for all media types.

Conclusion

This paper includes the present security mechanisms which can be used easily for multimedia systems for each kind of multimedia data and multimedia applications separately. Also discussed, the usages of cryptographic schemes along with few of the digital watermarking techniques are given. Watermarking is not a new concept and has been a while in the field of security in distributable digital data; but its robustness, reliability and imperceptibility for all types of media has remained un-explored and un-addressed in the past. This condition has to be explored, where a single watermarking algorithm is employed for all possible multimedia signals. Further, it has been proposed that crypto mechanisms can be used as an additional security layer in digital watermarking, which in turn combines the effectiveness of three different approaches. All these, when incorporated into a single procedure, will prove to be a robust and healthy approach in resisting any type of attacks.

Acknowledgement

We extend our gratitude to Mr. Naveen Hemrajani, Principal, Suresh Gyan Vihar School of Engineering, Jaipur, Mrs. Savita Shiwani, Head of the Department of Computer Science, Suresh Gyan Vihar School of Engineering, Jaipur and our guide Mrs. Ruchi Dave, Associate Professor, Suresh Gyan Vihar School of Engineering, Jaipur.

References

[1] 'Approaches to Multimedia and Security' by Klara Nahrstedt, Jana Dittmann, Petra Wohlmacher
[2] 'A Robust Image Authentication Method Surviving JPEG Lossy Compression' by Chang
[3] 'Digital Watermarking and Multimedia Security' by Jeng-Shang Pan, Peng Shi
[4] 'Multimedia Security in Communication' by Shiguo Lian, Yan Zhang, Yu Chen

Intelligent Fuzzy Hybrid PID Controller for Temperature control in Process Industry

Er. Rakesh Kumar (Assistant Prof.), Er. H.S Dhaliwal (Assistant Prof.),
Er. Ram Singh (Assistant Prof.), Er. Mandeep Sharma (Lecturer)

Department of Electrical Engineering B.H.S.B.I.E.T. Lehragaga,
Punjab Technical University Jalandhar, Punjab, India
E-mail: raj5sept@rediffmail.com

Abstract
This paper presents a systematic approach for the design and implementation of temperature controller using Intelligent Fuzzy Hybrid PID Controller for Temperature control in Process Industry. The proposed approach employs PID based intelligent fuzzy-controller for determination of the optimal results than PID controller parameters for a previously identified process plant. Results indicate that the proposed algorithm significantly improves the performance of the chemical plant. It is anticipated that designing of PID based fuzzy controller using proposed intelligent techniques would dramatically improves the speed of response of the system, Rise time and settling time would be reduced in magnitude in the intelligent scheme as compared with conventional PID controller. ..

Keywords: Process plants, Steam temperature control, Industrial system, Multiobjective control; Optimal-tuning; PID control Fuzzy logic control, genetic algorithms, nonlinear control, optimal control, PID control

Introduction

Well-known proportional-integral-derivative PID controller is the most widely used in industrial application because of its simple structure. On the other hand conventional PID controllers with fixed gains do not yield reasonable performance over a wide range of operating conditions and systems (time-delayed systems, nonlinear systems, etc.). Control techniques which based on fuzzy logic and modified PID controllers are alternatives to conventional control method. Fuzzy logic control (FLC) technique has found many successful industrial applications and demonstrated significant performance improvements. However, fuzzy controller design remains a fuzzy process due to the fact that there is insufficient analytical design technique in contrast with the well-developed linear control theories. A wide variety of fuzzy PID-like controllers have been developed. In most cases, fuzzy controller design is accomplished by trial-and-error methods using computer simulations. Significant studies based on the closed-form analysis of fuzzy PID-like controllers started with the work of Ying, Siler, and Buckley , where they have used a simple four-rule controller similar to that of Murakami and Maeda ,More analytical work in this regard was subsequently reported for the four-rule controllers, and linear-like fuzzy controllers. Palm has analytically demonstrated the equivalence between the fuzzy controller and sliding-mode controllers.

Fuzzy logic can handle imprecise data and can effectively used in controller.intelligent fuzzy control system is as shown in fig 1.

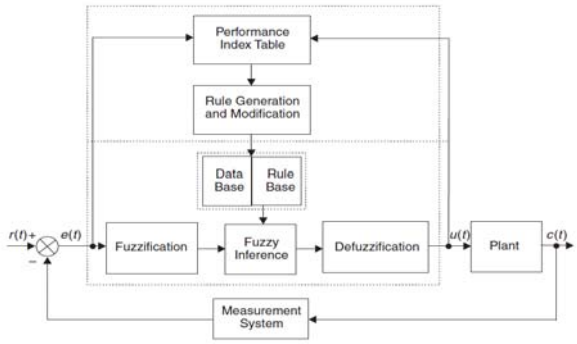


Fig 1. FUZZY CONTROL SYSTEM

This Paper is an attempt to undertake the development of a new analytical approach to the optimal design of fuzzy controllers. We propose a new methodology for the optimal design of fuzzy PID controllers. In the proposed work the main objective of the investigator is to compare the performances of conventional PID controllers and the intelligent fuzzy logic controller. For this comparison, two parameters needs to be evaluated i.e. Overshoot and settling time. This paper suggests a fuzzy logic based controller which acts with the help of artificial intelligence techniques. There are many artificial intelligence techniques and fuzzy logic is one of them.

Methodology

A new methodology is proposed for the analytical design of a fuzzy PID controller. There are various steps for the design of the Fuzzy PID Controller.

In Step 1, the structure of a fuzzy PID controller is designed and the structural parameters are set for the preliminary design. The tuning parameters are identified

Step 2, while in Step 3 an analytical fuzzy calculation is performed, which produces a closed-form relationship between the design parameters and control action for the fuzzy inference.

Step 4, numerical simulation (or control theory) is used to obtain the control performance data.

Step 5, genetic-based optimizations are carried out to produce optimal design parameters. This also provides useful information for the redesign of the original system.

Finally, if necessary, redesign is undertaken using the designer's expertise for further improvement to the control system. Note that the theoretical study in Step 3 makes the fuzzy controller transparent. This step is important since it will establish a close link between fuzzy control design technique and classical/modern control theory.

Simplicity is a key principle of this design methodology

The reason is obvious if we see that fuzzy logic controllers are systems which simulate human control exercise. For many everyday control tasks, people initially try to apply simple rules. Three rules used in this work are very common in a feedback set-point control problem. If a satisfactory control process can be achieved by applying simple rules, the use of complex rules, which is often associated with a higher cost of computation, becomes unnecessary. Simplicity is the best and direct way to maintain a clearly physical insight into the control laws. It also makes high-dimensional fuzzy systems tractable for using simple mathematical expressions for describing functionality between design parameters and nonlinearity.

Simulation Results

Oil Tank Temperature Controller

This temperature controller is used to control the temperature of raw oil in oil tank. In this the set temperature is 255^{o}C and PID temperature controller reaches set temperature in five hours and ten minutes.

Fuzzy model was developed using error, change in error and fuzzy output to improve the settling time.

Table 1: Fuzzy system, for oil tank temperature controller

(a) Membership functions of Error input.

Membership function for Error			
Linguistic variable	Initial value	Peak value	Final value
Very Small (VS)	-2	20	50
Small (S)	20	50	90
Medium (M)	60	100	140
High (H)	100	140	190
Very High (VH)	160	190	230

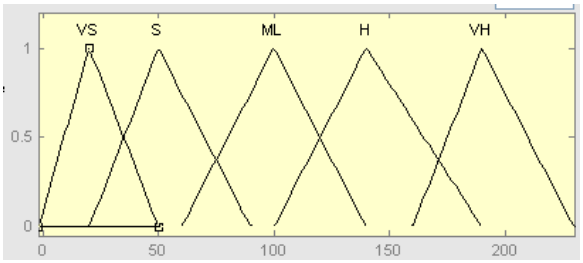
(b) Membership functions of Change in Error input.

Membership function for Change in Error			
Linguistic variable	Initial value	Peak value	Final value
Very Small (VS)	-10	-6	-2
Small (SM)	-6	0	6
Medium (MD)	0	8	15
Large (L)	9	17	25

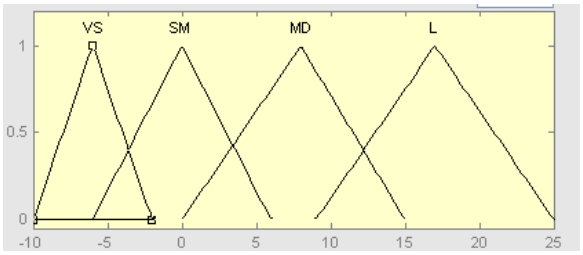
(c) Membership functions of Fuzzy output.

Membership function for Fuzzy Output			
Linguistic variable	Initial value	Peak value	Final value
Very Small (VS)	28	39	50
Small (S)	38	70	90
Medium (M)	70	110	140
Large (L)	120	165	210
Very Large (VL)	180	220	255

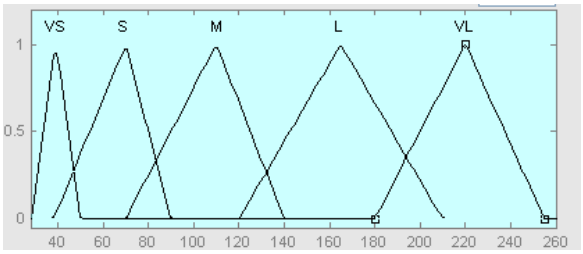
To develop Fuzzy controller, firstly error signal(e) is calculated by subtracting output of PID temperature controller from set temperature then change in error(Δe) was calculated by subtracting previous error from current error. Considering error and change in error as input and fuzzified output as output function membership functions are created for each input and output. Membership functions for these quantities are defined as in above Table 1. The membership functions are shown in schematic form in Fig. 2.



(a) Membership functions of Error input.



(b) Membership functions of Change in Error input.



(c) Membership functions of Fuzzy output.

Figure 2: Fuzzy system, for oil tank temperature controller

A rule base was developed for the fuzzy model using simple IF-THEN rules.

The rule base is summarized as in Table 2

Table 2: Rule base

Fuzzy output (Fz)	Change in Error (Δe)			
Error(e)	VS	SM	MD	L
VS	VL	VL	VL	-
S	-	L	VL	-
M	-	L	L	-
H	S	M	M	L
VH	VS	VS	S	S

On the basis of this rule base a fuzzified output is calculated. This Fuzzy model is simulated in MATLAB fuzzy logic toolbox GUI, and results are obtained. Then results are plotted along with the actual temperature and set temperature obtained from the process, are plotted in Fig.3

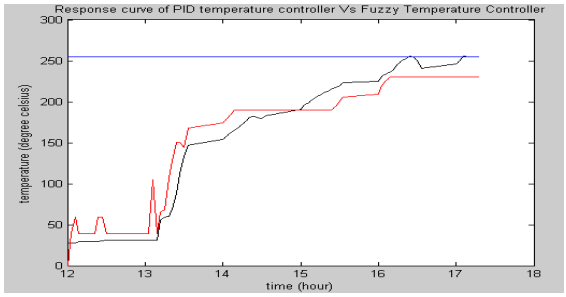


Figure 3: Response curve of PID temperature controller Vs Fuzzy temperature controller in oil tank temperature controller

Red graph shows the fuzzy output of fuzzy model of oil tank temperature controller, black line represent the output of PID temperature controller and blue line represent the set temperature enter in PID temperature controller. Fuzzy output has some oscillations in rising time and a steady state error of 35°C. To improve this fuzzy response the membership functions of all the input and output are increased. Membership functions of error and change in error inputs have been increased to six and that of fuzzy output has been increased to seven.

The rule base is also revised as shown in Table 3. By using this rule base, oscillations in fuzzy response decreases and steady state error was also reduced than the last fuzzy model.

Table 3 Improved Rule base.

Fuzzy output (Fz)	Change in Error (Δe)					
Error (e)	N	NS	SM	M	L	VL
VS	EL	EL	EL	EL	-	-
S	-	VL	VL	VL	-	-
M	-	L	L	L	-	-
H	-	ML	L	L	-	-
VH	S	M	M	ML	ML	-
EH	VS	VS	S	M	M	M

The MATLAB simulation results are plotted along with the actual temperature and set temperature obtained from the process, are plotted in Fig.4.

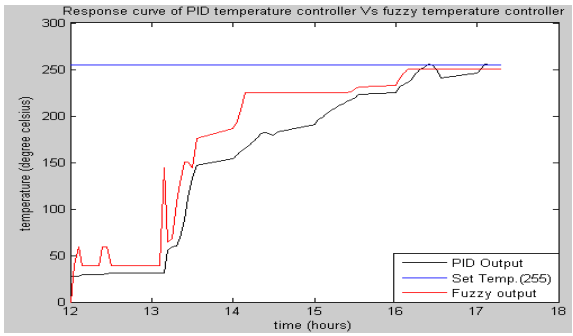


Fig.4 Improved Response curve of PID temperature controller Vs Fuzzy temperature controller in oil tank temperature controller

Here the steady state error is decreased to 5°C, and settling time also improved. Further improvements in Fuzzy output and rule base have been made. To achieve this requirement, the range of last two membership function of fuzzy output has been changed. The new range is VL – 200-220-255 and EL – 250-255-260 and rule base is shown in Table 4.

Table 4 Improved Rule base

Fuzzy output (Fz)	Change in Error (Δe)					
Error (e)	N	NS	SM	M	L	VL
VS	EL	EL	EL	EL	-	-
S	-	EL	EL	EL	-	-
M	-	VL	VL	VL	-	-
H	-	ML	L	L	-	-
VH	-	M	M	ML	ML	-
EH	VS	VS	S	M	M	M

The MATLAB simulation results are plotted along with the actual temperature and set temperature obtained from the process, are plotted in Fig.5.

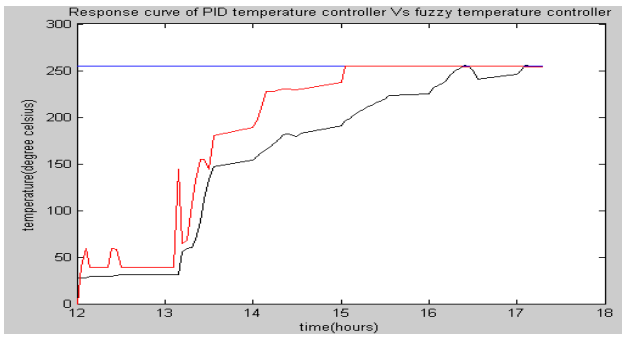


Fig.5. Improved response curve of PID temperature controller Vs Fuzzy temperature controller in oil tank temperature controller

Here the steady state error and settling time both have been improved. Steady state error is decreased to zero and settling time is reduced by 2 hours and 10 minutes. For Further improvements, Fuzzy output and rule base have been revised. Revised Fuzzy output and rule base are shown in Table 5.

Table 5. Improved Rule base

Fuzzy output (Fz)	Change in Error (Δe)					
Error (e)	N	NS	SM	M	L	VL
VS	EL	EL	EL	EL	-	-
S	-	EL	EL	EL	-	-
M	-	VL	VL	VL	-	-
H	-	L	VL	VL	-	-
VH	-	ML	ML	L	L	-
EH	VS	S	VS	M	M	-

The MATLAB simulation results are plotted along with the actual temperature and set temperature obtained from the process, are plotted in Fig.6.

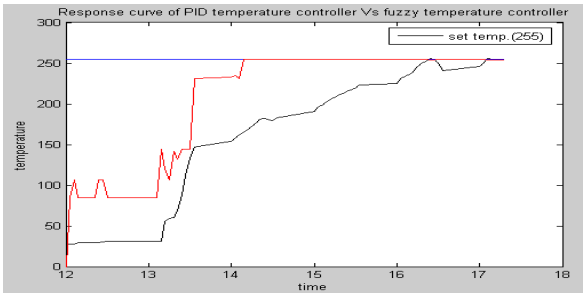


Fig.6. Improved response curve of PID temperature controller Vs Fuzzy temperature controller in oil tank temperature controller

Here fuzzy output response contains lesser oscillations. Settling time also reduce by 45 minutes from last fuzzy model. Total time reduced is 2 hours and 55 minutes. To further improve fuzzy response, fuzzy system is revised. Membership functions of error input and fuzzy output are increased. Membership functions of error input have been increased to seven and that of fuzzy output has been increased to nine. Improved error input, fuzzy output and rule base are shown in Table 6

Table 6. Improved Rule base

Fuzzy output (Fz)	Change in Error (Δe)					
Error (e)	N	NS	SM	M	L	VL
VS	EL	EL	EL	EL	-	-
S	EL	EL	EL	EL	-	-
ML	EL	EL	EL	EL	-	-
MH	-	VL	VL	-	-	-
H	-	ML	ML	L	L	L
VH	-	HS	HS	LM	HM	HM
EH	VS	VS	LS	-	-	LS

The MATLAB simulation results are plotted along with the actual temperature and set temperature obtained from the process, are plotted in Fig.7.

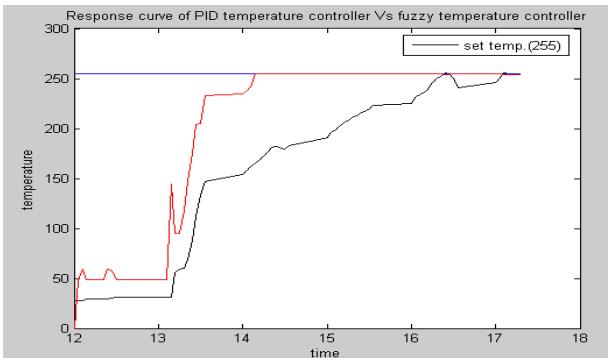


Fig.7. Improved response curve of PID temperature controller Vs Fuzzy temperature controller in oil tank temperature controller

Here fuzzy output response contains lesser oscillations than the last fuzzy output response. Settling time is also reduced by 20 minutes, i.e. settling time is 2 hours and 15 minutes. But one spick is produced during the rise time that can be removed by changing the rule base as in Table 8

Table 8 Improved Rule base

Fuzzy output (Fz)	Change in Error (Δe)					
Error (e)	N	NS	SM	M	L	VL
VS	EL	EL	EL	EL	-	-
S	EL	EL	EL	EL	-	-
ML	EL	EL	EL	EL	-	-
MH	-	VL	EL	-	-	-
H	-	ML	ML	L	L	L
VH	-	HM	HM	HM	HM	HM
EH	VS	VS	LS	-	-	VS

The MATLAB simulation results are plotted along with the actual temperature and set temperature obtained from the process, are plotted in Fig.8.

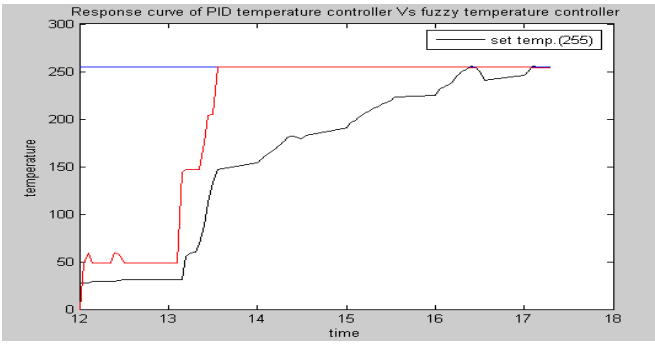


Fig.8. Improved response curve of PID temperature controller Vs Fuzzy temperature controller in oil tank temperature controller

Here settling time reduced by 10 minutes from last fuzzy model. Finally fuzzy model give fuzzy output response with lesser oscillations. This fuzzy model reduces the settling time by 3 hours and 15 minutes.

Comparing first fuzzy model in Fig. 3 and last fuzzy model in Fig 8 developed for oil tank temperature controller, an analysis is made that by increasing the number of membership functions from 5 to 7 for error input. and from 4 to 6 membership functions for change in error input, from 5 to 9 membership functions of fuzzy output, a response curve has been obtained that has a settling time of 1hour 55 minutes, and oscillations in response curve are all most removed. This fuzzy model reduces the settling time by 3 hours and 15 minutes.

Conclusions

Aiming at characteristic of agro plants and control requirement, a Fuzzy-PID hybrid controller with advantages of both fuzzy controller and PID controller integrated is presented in this paper. The available field application shows Fuzzy-PID hybrid controller can not only restrain the large fluctuation to temperature effectively, but also has excellent static performance. Fuzzy-PID hybrid controller has decisive effect on keeping stable temperature of agro and provides powerful support for smooth production process. Owing to improving production and super quality product by application of Fuzzy-PID hybrid controller, considerable economy benefit is brought to the enterprise.

References & Bibliography

[1] Erdal Kayacan and Okyay kaynak, "An Adaptive Grey Fuzzy PID Controller With Variable Prediction Horizon," Tokyo, Japan, pp 760-765, September 20-24, 2006.

[2] B.G. Hu, G.K.I Mann and R.G Gosine, "New methodology for analytical and optimal design of fuzzy PID controllers," IEEE Transactions on Fuzzy Systems, Vol. 7, no. 5, pp. 521-539, 1999.

[3] Awang N.I. Wardana, "PID-Fuzzy Controller for Grate Cooler in Cement Plant," IEEE Transactions on Fuzzy System, Vol. 32, no.7, pp.1345-1351,2005.

[4] Han-Xiong Li,Lei Zhang, Kai-Yuan Cai, And Guanrong Chen," An Improved Robust Fuzzy-PID Controller With Optimal Fuzzy Reasoning," IEEE Transactions on Systems, Vol. 35, no. 6, 1283-1292, December 2005.

[5] Is in Erenoglu, Ibrahim Eksin, Engin Yesil and Mujde Guzelkaya, "An intelligent hybrid fuzzy PID controller," Proceedings of 20th European Conference on Modeling and Simulation, 2006.

[6] Leehter Yao and Chin-Chin Lin, "Design of Gain Scheduled Fuzzy PID Controller," World Academy of Science, Engineering and Technology, pp.152-1561, 2005.

[7] Zhen-Yu Zhao, Masayoshi Tomizuka, Satoru Isaka, "Fuzzy gain scheduling of PID controllers," IEEE Transactions on Systems, man and cybernetics, Vol. 23, no. 5, September/October 1993, pp. 1392-1398.

[8] B. Nagaraj, S. Subha, B. Rampriya, "Tuning Algorithms for PID Controller Using Soft Computing Techniques," IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.4, April,2008, pp. 278-281

Impact of Parallel Computing on Bioinformatics Algorithms

O.P. Gupta*, Sita Rani and Dhruv Chander Pant*****

**Associate Professor, School of Information Technology, PAU Ludhiana, Punjab, India*
E-mail: opgupta@pau.edu
***Associate Professor, RIMT – IET, Mandi Gobindgarh, Punjab, India*
E-mail: sitasaini80@yahoo.in
***Ph.D. Scholar, PTU Jalandhar, Punjab, India*
E-mail: dpant9@gmail.com

Abstract
The biggest challenge bioinformatics facing today is to manage, analyze and process volumous genome data. Such analysis and processing is very impractical with the help of uniprocessor computers, so the need of parallel computing in bioinformatics arises. Now distributed computers, cloud computers and multicore processors are available at very low cost to deal with bulk amount of genome data. Along with these technological developments in distributed computing, many efforts are being done by the scientists and bioinformaticians to parallelize and implement the algorithms to take the maximum advantage of the additional computational power.

In this paper various parallel computing architectures and parallel implementation of the bioinformatics algorithms are discussed. The performance analysis of the parallelized algorithms is also analyzed.

Keywords: Algorithms, Bioinformatics, Genome data, Parallel Computing,

Introduction

Bioinformatics

Bioinformatics is the applications of computer science to store, mange, analyze and process biological data [1], [2]. Bioinformatics is applied in various areas like molecular medicine, personalized medicine, preventative medicine, gene therapy, drug development, waste cleanup, climate change studies, alternative energy sources, biotechnology, antibiotic resistance, forensic analysis of microbes, bio-weapon creation, crop improvement, insect resistance, vetinary sciences etc. [3]. In all these application areas bioinformatics algorithms deal with bulk amount of genome data. Generally the bioinformatics applications face the following challenges [4]:

- To manage and process the bulk amount of genome data.
- To reduce data analysis time.

So break through technological development was needed to solve many critical problems of bioinformatics [5]. Such data management is impractical with the help of uniprocessor computers. So the use of parallel computing in bioinformatics applications is important. Now to deal with bulk amount of genome data distributed computers, cloud computers and multicore processors are also available at very low cost.

Parallel Computing Architectures

In parallel computing a problem is broken into discrete parts and instructions of different parts run on different CPUs concurrently as shown in Fig. 1.

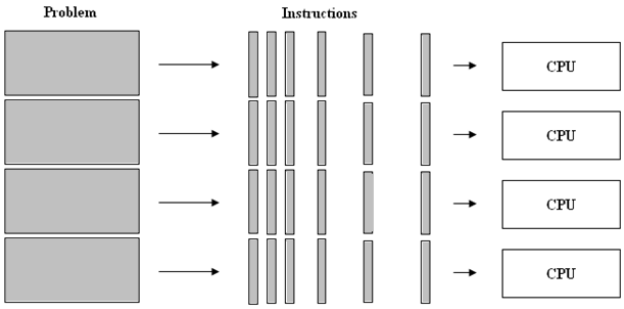


Fig. 1 Concept of Parallel Computation

In parallel computation a computing resource may be a single computer with multiprocessors, different number of computers connected by a network, multi core processors or the combination. And the problem should be able to be broken into different parts that can run simultaneously [6]. The various advantages of using parallel computing are:

- Save time and/or money
- Solve larger problems
- Provide concurrency
- Use of non local resources

Parallel computer systems can be classified into two main models: Single Instruction Multiple Data (SIMD) Systems and Multiple Instructions and Multiple Data (MIMD) Systems Fig.2. A SIMD system consists of multiple simple processors with small local memory. These processors use explicit communication to transfer data to each other. All the different processors should be strongly synchronized. Because of the complexity and inflexibility, SIMD systems are not used for very advanced applications.

MIMD systems are more suitable to bioinformatics applications. In MIMD machines each process executes

completely independent of the other process asynchronously. MIMD systems are further classified on the bases of shared and distributed memory. A process running in the shared memory system can access any local or remote memory of the system whereas a process running in distributed memory cannot.

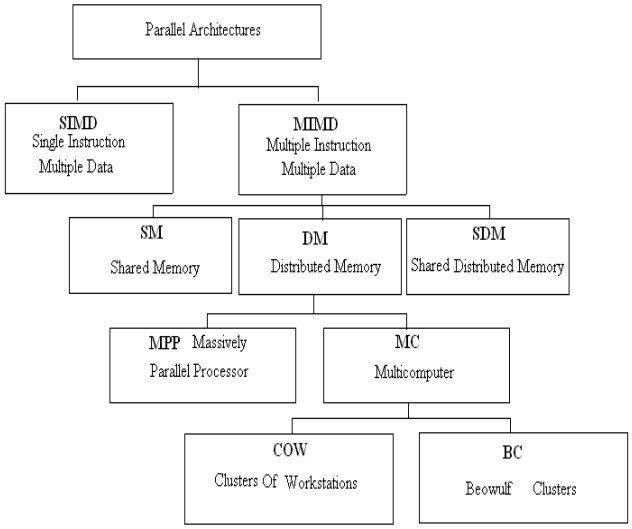


Fig. 2 Summarized Parallel Computer Architectures.

Shared memory systems have many advantages for bioinformatics applications. Design of parallel programs is simplified with a single address map. Different processes can also communicate without any time loss, because every CPU has direct access to memory. Whereas in distributed memory systems a time penalty is incurred for intercrosses communication because of the lack of a single address map for the memory.

Current trends in multiprocessor design try to utilize the positive factors of both the architectures. Each CPU has some local memory attached to it and hardware creates an illusion of common memory shared by the whole system. So the memory installed in any node may be accessed by any other node with very low less time penalty.

But as now very fast processors are available in the work stations, so microcomputers are connected with the help of Local Area Network. In this way virtual parallel computers are developed. These computers are also called Multi- computers which are constructed with the help of Cluster of Workstations (COWs). One more architecture of multi – computers is Beowulf – clusters which consist of very simple hardware components like ordinary PCs. In this architecture a public domain server controls the whole cluster.

Parallelized Implementation Of Different Bioinformatics Algorithms

Cluster Implementation of Sequence Alignment Algorithms

Smith waterman algorithm is used for local alignment between two sequences [7]. The algorithm is based on dynamic programming technique. If the two sequences of size *n* are to be matched then algorithm takes time *O (n*n).* As the value of *n* increases the time required becomes significantly high. Thus the need of parallel implementation of Smith waterman algorithm arises [8]. In smith waterman algorithm is implemented on clusters. The results of this cluster implementation are shown in Table 1.

Table 1 Parallel Implementation of Sequence Alignment Algorithms on Clusters

Sequence length	Sequential Algorithm	Parallel algorithm, np Processor					
		1	2	4	8	16	32
500	0.24	0.3	0.2	0.1	0.1	0.5	1.3
1000	1.7	2.7	1.5	0.9	0.6	1.1	1.5
1500	5.9	8.8	4.8	2.9	1.8	1.7	2.1
2000	13.9	20.3	10	6.3	3.7	3.2	3.4
2500	26.2	39.5	21	11.6	6.9	5.1	4.5
3000	45.5	67.2	35.4	19.5	11.4	8.1	6.5
3500	71.6	106	55	30.1	17	11	8.9
4000	107.2	158	82	44.2	25	16	12
4500	152	225	118	62.4	34	21	15
5000	208	310	158	86.4	46	28	20

In parallelization of the algorithm pipelining is used. In the score matrix, each row is computed sequentially and is blocked till the required cells in the above row are computed. When 32 processors are used with a sequence of 5000 characters long, the implementation showed an improvement up to 10.30 times. Smith waterman algorithm is also parallelized by its implementation on cell broadband engine [9]. In this implementation a static load balancing strategy is used. Under this strategy, work load at the beginning is divided equally among all the processors and processes. In the first step, algorithm reads the input dataset. In the next step the input sequences are processed by processing units to acquire the respective sequence parts in their local memories. For a sequence of 2048 characters long with this algorithm a speed up of 6.5 times is obtained. For multiple sequence matching multiple sequence alignment algorithms are used [10]. If there are n sequences, *n*(n-1)/2* pair-wise alignments need to be calculated. As the number of sequences increase, number of pair wise alignments also increase and the complexity of the algorithm also.

Once the distance matrix is calculated, in the next phase of the algorithm phylogenetic tree is produced. And in the final phase of the algorithm, previously generated phylogenetic tree is used to determine the order of the alignment. Experiments were performed with a number of techniques and concluded that to distribute all the n sequences to each processor was a better method. In this technique each of the *P* processors performs exactly *n*(n-1)/2P* alignments. Although this method has very high communication cost, even then it showed maximum speed up. For n = 500 sequences, where each sequence had 200 characters this technique showed a speedup of 5.81 times [11]. The result of implementation on cluster is shown in Fig. 3.

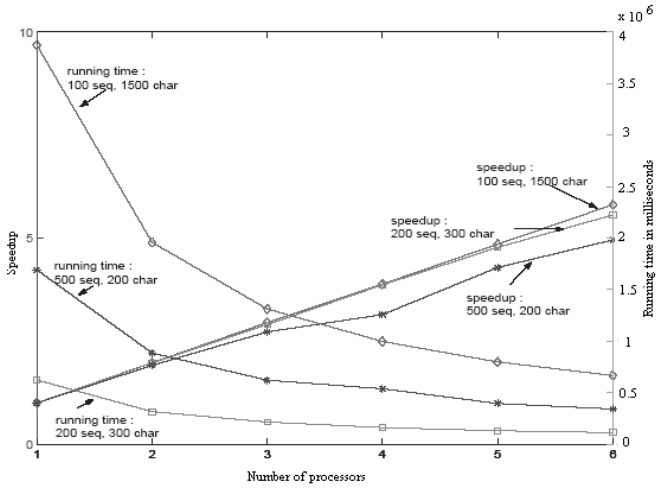


Fig. 3 Running Time and Speed ups for Parallel Implementation of Clustlaw.

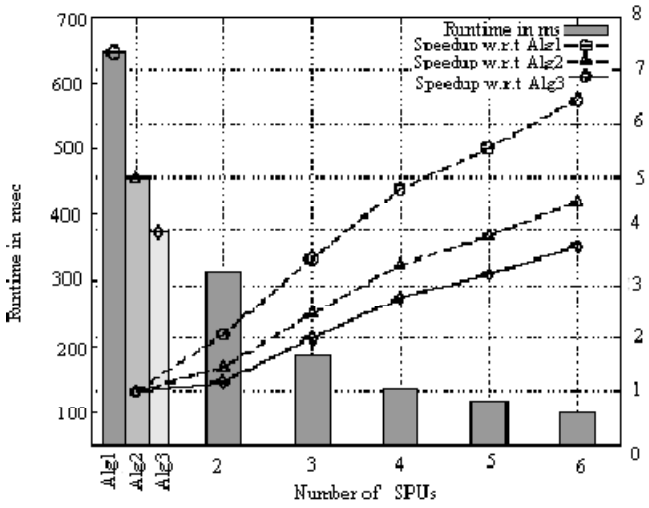


Fig. 5 Global Alignment for input size of 2048X2048

Parallel multiple sequence alignment was also performed on the cell broadband engine [12] where the parallel portions of the code were executed on synergistic processing units whereas sequential code on power processing units. For n= 8 pair of sequences where each sequence had 2048 characters showed a speed up of 46.37x times which is shown in Fig. 4..

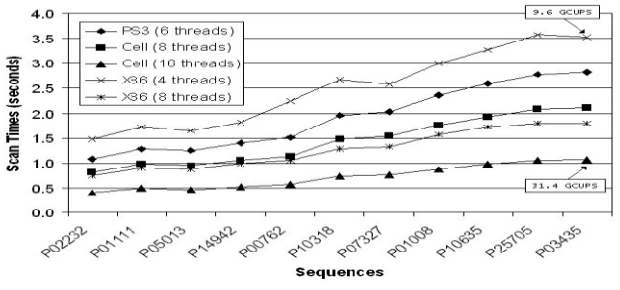


Fig. 4 Performance of sequence Search algorithms on Cell

Cell Implementation of Sequence Alignment Algorithms

Cell broadband engine based implementation for global alignment was performed on IBM Cell SDK 3.0 to obtain the results the implementation was executed on Sony Play Station 3 (SP3) and was compiled with optimized level -O3. The performance of this implementation was studied on different number of SPUs. The result of the implementation is shown in Fig. 5 by using up to 6 synergistic processing units (SPUs). When this implementation is compared with a sequential implementation on a desktop with 3.2 GHz Pentium 4 Processor, a speed up of 6.5 xs is obtained. When this implementation was compared with best sequential algorithm with single SPU and a Pentium 4 Processor the speed ups were 4.5x and 3.5x respectively [11].

FASTA is a multiple sequence alignment algorithm. It is used to compute pair of match and mismatch between the sequences. Then this computation is used to detect the similarity between the sequences.

Cell Broadband Engine Implementation of FASTA for Multiple Sequence Alignment

AltiVec application-programming interfaces, already there in the Fasta package, was converted to the synergistic processing unit application programming interfaces to run Fasta on cell broadband engine. The interfaces which are not converted are implemented with the help of multiple instructions i.e. Vec max and vec subs [11]. Vec max application programming interfaces is used to determine the maximum of two vectors. Result is stored in the output vector. From the stored result, synergistic processing unit find the greater vector. Vec subs application programming interfaces are used to perform saturated subtraction. In saturated subtraction any element with negative value is set to zero. In smith waterman a positive value of each cell is needed, so this application programming interface is helpful in the execution of smith water man.

Once the alignment scores are calculated with the help of power processing units, scores, query and library sequences are delivered to synergistic processing unit to execute the smith waterman kernel. But this cell implementation is limited by the size of the sequence. A sequence of more than 2048 characters cannot be compared in this implementation because of the size of the synergistic processing unit local memory. This problem can be rectified with the help of pipeline approach. Once smith waterman is implemented on the cell, then it can be used in FASTA package. In FASTA each query sequence is compared with every sequence in the database. Hence balancing load between each pair of sequences is evaluated.

Protein Structure Prediction Algorithms

The most important application of protein structure prediction is drug design. In protein structure prediction tertiary structure of the proteins is predicted from its amino acid sequences. On the bases of physical properties many protein structures are possible. So it is very difficult to understand the stability of a structure.

Genetic algorithm is used to implement protein structure prediction on computational grid [13]. Cell broadband engine

implementation of protein structure prediction is also done [14]. In this implementation sequences are shifted from database to synergistic processing units with this implementation a speed up between 3.2x and 3.6 x was achieved.

Conclusions

It is concluded that Parallel Computing is having very good impact on computational and data intensive applications. The processing time of bioinformatics algorithms can be improved by parallelization. The sections of the algorithms which take more time can be divided in to subprograms to execute concurrently.

References

[1] D. Jawadat , " Era of Bioinformatics", in Proceedings of 2nd IEEE international conference on Information and Communication Technologies: From Theory to Applications, 2006, pp 18060-1865.

[2] R. Hughey and K. Karplus, "Bioinformatics: A New Field in Engineering Education" in Proceedings of 31st ASEE/IEEE Frontiers in Education Conference, 2001, pp 15-17.

[3] O.P.Gupta and S. Rani, "Bioinformatics applications and Tools: An Overview", CiiT- International Journal of Biometrics and bioinformatics, vol 3, no 3, pp 107-110, 2010.

[4] I. Gorton, P. Greenfield, A. Sazalay and R. Williams, " Data Intensive Computations in 21st Century", in Computer Magazine of IEEE Computer Society, vol 41, no 4, pp 30 -32, 2008.

[5] C. Mueller, M. Dalkilic and A. Lumsdaine , " Implementing Data Parallel algorithms for Bioinformatics", in proceedings of SIAM Conference on Computational Science and Engineering", 2005, pp 226-232.

[6] K. Hwang and Z. Xu, " Scalable Parallel Computing: Technology, Architecture and Computing" Mc-GrawHill Series in Computer Engineering, 1998.

[7] T.F. Smith and M.S. Waterman , " Identification of Common Molecular Subsequences", Journal of Molecular Bilogy, vol 147, no 1,pp 195-197, 1981.

[8] Y. Chen, S. Yu and M. Leng , " Parallel Sequence Alignment Algorithms for Clustering System", International Federation for Information Processing, vol 207,pp 311-321, 2006.

[9] A. Wirawan, K.C. Keong and B. Schmidt, " Parallel DNA Sequence Alignment on Cell Broadband Engine", Springer – Verlag Berlin Heidelber, pp 1249 – 1256, 2008.

[10] J. Ebedes and A. Datta, " Multiple Sequence Alignment in Parallel on a Workstation Cluster", Oxford University Press, vol 20,no 77,pp1193-1195,2004.

[11] B.K. Pandey, S.K. Pandey and D. Pandey, " A Survey of Bioinformatics Applications on Parallel Architectures", International journal of Computer Applications, vol 23, no 4, pp 21 – 25, 2011.

[12] V. Sachdeva, M. Kistler, E. Speight and T.H.K. Tzeng, "Exploring the Viability of Cell Broadband Engine for Bioinformatics applications", in Proceedings of IEEE International Parallel and Distributed Processing Symposium, 2007, pp 1-8.

[13] G. Minervini, G.L. Rocca, P.L. Luisi and F. Polticelli, " High Throughput Protein Structure Prediction in a Grid Enviorment", Journal of Bio- Algorithms and Med System, vol 3, no 5, pp 39-43, 2007.

[14] H. Zhang, B. Schmidt and W.M. Witting, " Accelrating BLASTP on the Cell Broadband Engine", in Proceedings of the 3rd International Conference on Pattern Recognition in Bioinformatics, 2008, pp 46 – 470.

Image Retrieval using Dual Tree Complex Wavelet Transform

Sanjay Patil[#] and Sanjay Talbar[$]

[#]Associate Professor, Jaywant College of Engg. and Management, K.M. Gad, Maharashtra, India
E-mail: sanjayashri@rediffmail.com
[$]Professor, E & TC Dept, SGGS College of Engineering &Technology, Nanded, Maharashtra, India
E-mail: sntalbar@yahoo.com

Abstract
This paper demonstrates a novel approach for shift invariant image retrieval using set of dual-tree discrete wavelet transform (DT-DWT) and dual-tree complex wavelet transform (DT-CWT). The DT-CWT is relatively recent enhancement to the DT-DWT. It is nearly shift invariant and directionally selective in two and higher dimensions. The two dimensional DT-CWT is nonseparable, but it is based on a computationally efficient, separable filter banks (FB). The magnitude and phase of CWT coefficients can be exploited in the development of the new efficient and effective wavelet based algorithms where DWT is inefficient. In this paper 1000 images database of 10 different classes is used. Experimental results indicate that the proposed method gives excellent retrieval accuracy of 95% for Dinosaur class of images. The Roses class gives 93.78% retrieval accuracy. Also, for other classes retrieval accuracy is good. There is improvement in retrieval accuracy using DT-CWT than using DT-DWT filters.

Index Terms: Dual-tree discrete wavelet transform (DT-DWT), dual tree complex wavelet transform (DT-CWT), filter bank (FB), content based image retrieval, similarity measures.

Introduction

N present scenario, due to internet and availability of large data storage facility, huge numbers of images have been produced and available for active research. Traditionally, retrieval of the images was text based. In this method the images or scene is described by some text annotation and images are searched by using keyword based searching methods. This task is very time consuming and it is very difficult to describe color, texture, shape and object within the image [1].

To avoid the limitation of text based searching, new techniques are developed and images are searched based on their visual content like color, texture or shape which is known as Content Based Image Retrieval (CBIR) Systems. CBIR technique uses low level features to represent the images relevant to the query image from the database. The figure 1 represents block diagram of typical CBIR System.

For the given image database, first extract features of database images. The features can be visual features like color, texture, shape, region or spatial features or some compressed domain features. The extracted features are described by feature vectors. These feature vectors are then stored to form Image feature database.

For a given query image, similarly extract its features and form a feature vector. This feature vector is matched with the already stored vectors in image feature database. The distance between the feature vector of the query image and those of the images in the database are then calculated. Obviously the distance of a query image with itself is zero if it is in database. The distances are then stored in increasing order and retrieval is performed with the help of indexing scheme.

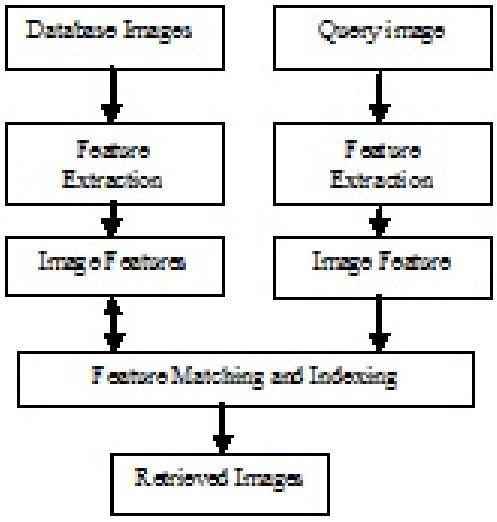


Figure 1: Block Diagram of typical CBIR System.

In this paper a new method based on Dual Tree Discrete Wavelet Transform and Dual Tree Complex Wavelet Transform (DT- CWT) is compared. The technique makes use of complex wavelet transform which represents the latest research result on multi resolution analysis. [2]

DWT Implementation and Limitations

Filter Bank (FB) for 1-D DWT

It is proved that using multi resolution analysis; DWT can be expressed in terms of FIR filters using *Mallat's Pyramid Algorithm* [3,4]. The input signal is filtered in parallel by a low-pass filter and a high-pass filter to give approximation (coarser) and details of the input signal as shown in Figure 2.

The approximation part can be further decomposed up to level *j*. This forms the analysis *FB* and for reconstruction synthesis FB is used.

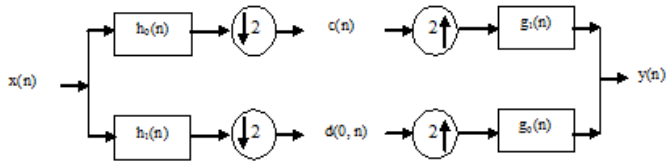


Fig. 2: Analysis and Synthesis FB's for implementing DWT.

Troubles with Real Wavelets

Shift Invariance: Real DWT are very sensitive to shifts. Small shifts in the input signal can cause major variations in the distribution of energy between DWT coefficients.

Poor Directional Selectivity: DWT coefficients reveal only three spatial orientations (horizontal, vertical and diagonal).

Lack of Phase: DWT analysis of real signals lacks the phase information that accurately describes non-stationary signal behavior.

Oscillations: As wavelets are band pass functions, the wavelet coefficients tend to oscillate positive and negative around singularities. Due to wavelet overlapping, it is possible to have a small or even zero wavelet coefficient.[5]

Importance of Phase

Figure 3 shows importance of phase information in implementation of DWT. For Lena and Mandrill images, if we take Fourier transform of both the images and take their inverse by exchanging their phases, we can see that the Lena image appears like a Mandrill image and vice-versa. So, the phase information is equally important in description of the image.

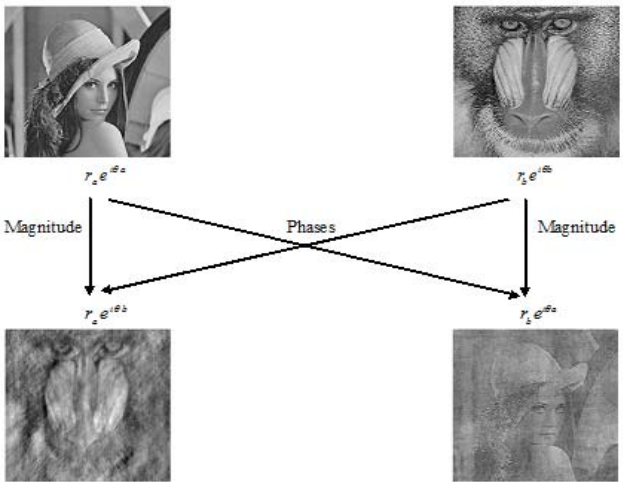


Fig.3: Importance of phase.

Complex Wavelet Transform (CWT) Solution

The limitations of real DWT can be overcome by using complex DWT. The features of complex DWT are as follows,

- Approximate shift invariance.
- Good directional selectivity with Gabor-like features.
- Perfect reconstruction with linear-phase.
- Limited redundancy independent of number of scales (2^m for *m*-D).
- Can be implemented using existing efficient DWT software and hardware.

Analytic Signal and Hilbert Transform

A real sinusoid $A\cos(\omega t+\theta)$ can be converted to a complex sinusoid by generating a phase-quadrature component $A\sin(\omega t+\theta)$ to give an *imaginary part.* This phase-quadrature component can be generated from the in-phase component by a simple $+\pi/2$ time shift. When a real signal $x(t)$ and its Hilbert transform $y(t)$ are used to form a new complex signal $z(t)=x(t)+jy(t)$, the signal $z(t)$ is called the (complex) *analytic signal* corresponding to the real signal $x(t)$. This analytic signal does not have any negative frequency components, i.e. the spectrum of analytic signal is one sided.

Consider a complex scaling function and a complex wavelet,

$$\begin{aligned}\phi_c(t)&=\phi_r(t)+j\phi_i(t)\\ \psi_c(t)&=\psi_r(t)+j\psi_i(t)\end{aligned} \tag{1}$$

Where, the subscript *r* and *i* indicates real and imaginary components of corresponding scaling and wavelet functions such that

$$\begin{aligned}\phi_i(t)&\approx H\{\phi_r(t)\}\\ \psi_i(t)&\approx H\{\psi_r(t)\}\end{aligned} \tag{2}$$

Here, $H(\cdot)$ indicates Hilbert transform. The theory of these complex wavelets can be broadly classified into two approaches. First, non redundant approach which prevents the resulting complex DWT from overcoming the problems in real DWT, while the second redundant approach overcomes the problems of real DWT. An efficient way to implement an (complex) analytic wavelet was given by *Kingsburry* [6, 7], namely *Dual-tree Complex Wavelet Transform* (DT- CWT).

The Dual-TREE COMPLEX Wavelets APPROACH

This approach employs two real wavelet trees; the upper tree gives the real part while the lower tree gives the imaginary part of the CWT as shown in Figure 4. These trees are themselves real and use two different sets of perfect reconstruction (PR) filters. But they are designed such that the overall transform is analytic.

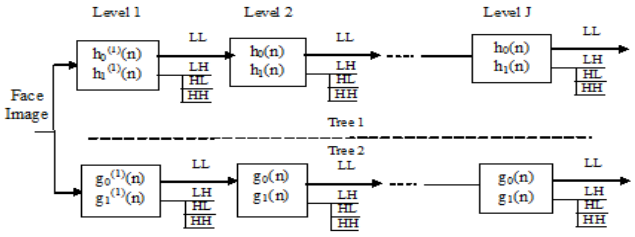


Figure 4. Analysis filter bank for Complex DT-DWT

Here $h_0(n)$ and $h_1(n)$ are the low-pass and high-pass filter pairs for the upper filter bank and $g_0(n)$ and $g_1(n)$ are low-pass and high-pass filter pairs for the lower filter bank. The filters used for first stage should be different from the remaining stages to have the frequency response to be one sided (analytic) [8]. In this paper for the implementation, *Farras* filters [9] are used for the first stage and *Kingsburry's* Q-shift filters [10] are used for the remaining stages.

2-D Complex Dual Tree DWT

The wavelets associated with conventional DWT are given by,

$$\psi_1(x,y) = \phi(x)\psi(y) \quad \text{(LH wavelet)}$$
$$\psi_2(x,y) = \psi(x)\phi(y) \quad \text{(HL wavelet)}$$
$$\psi_3(x,y) = \psi(x)\psi(y) \quad \text{(HH wavelet)}$$

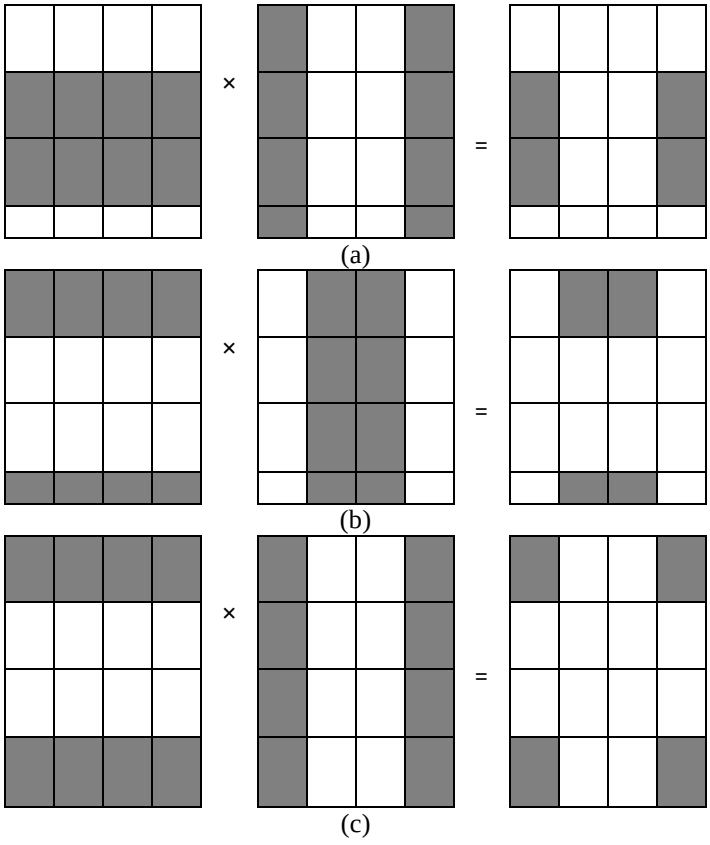


Fig 5. Idealized support of the Fourier spectrum for LH, HL and HH wavelets.

The LH is the product of the low pass function $\phi(\cdot)$ along the first and $\psi(\cdot)$ along the second which is shown in figure 5(a). Similarly, referring to figure 5(b), (c), the HL and HH wavelet the product can be interpreted. The LH and HL wavelets are oriented vertically and horizontally respectively. But the HH wavelet is not oriented diagonally +45° and -45° orientations and produces a checkerboard pattern as shown in the Figure 5(c).

Real-Dual-Tree Complex Wavelet Transform (R-DT-CWT)

The real 2-D dual-tree discrete wavelet transform (R-DT-DWT) can be implemented using $\{h_0(n), h_1(n)\}$ to implement one separable 2-D wavelet transform and $\{g_0(n), g_1(n)\}$ to another. Applying both separable transforms to the same 2-D data gives a total of six sub bands: two HL, two LH, and two HH sub bands. Take the sum and difference of each pair of sub bands to get a transform which is two-times expansive and free of the checkerboard artifact as shown in Figure 6.

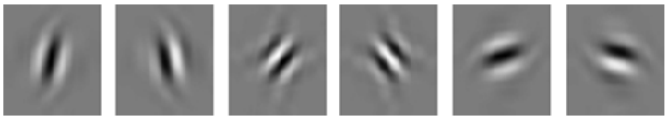


Fig 6: Impulse responses for 2-D R-DT-DWT.

If we define the separable 2-D wavelet bases as,

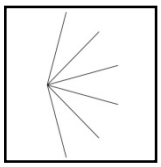
$$\begin{aligned}
\psi_{1,1}(x,y) &= \phi_h(x)\psi_h(y), & \psi_{2,1}(x,y) &= \phi_g(x)\psi_g(y) \\
\psi_{1,2}(x,y) &= \psi_h(x)\phi_h(y), & \psi_{2,2}(x,y) &= \psi_g(x)\phi_g(y) \\
\psi_{1,3}(x,y) &= \psi_h(x)\psi_h(y), & \psi_{2,3}(x,y) &= \psi_g(x)\psi_g(y)
\end{aligned} \quad (3)$$

Then the wavelets are defined by,

$$\psi_i(x,y) = \frac{1}{\sqrt{2}}(\psi_{1,i}(x,y) + \psi_{2,i}(x,y))$$
$$\psi_{i+3}(x,y) = \frac{1}{\sqrt{2}}(\psi_{1,i}(x,y) - \psi_{2,i}(x,y)) \quad (4)$$

for i = 1, 2, 3. The normalization $1/\sqrt{2}$ is used so that sum/difference operation constitutes an orthonormal operation. These wavelets are strongly oriented at an angles ± 15°, ± 45°, and ± 75°.

An illustration of R-DT-DWT operated on a synthetically generated image is as shown in the Figure 7. The input images have spokes in ± 15°, ± 45°, and ± 75° directions. Figure 7 shows that the R-DT-DWT sub bands are oriented at ± 15°, ± 45°, and ± 75° angles [14].



(a)

Fig7: Synthetically generated image, Complex-Dual-Tree Complex Wavelet Transform (C-DT-CWT)

Similarly, if we consider only the complex part of the complex wavelet, the 2-D frequency plane is the same as the spectrum of the real part with the separable 2-D wavelet bases given by,

$$\begin{aligned} \psi_{3,1}(x,y)&=\phi_g(x)\psi_h(y), & \psi_{4,1}(x,y)&=\phi_h(x)\psi_g(y) \\ \psi_{3,2}(x,y)&=\psi_g(x)\phi_h(y), & \psi_{4,2}(x,y)&=\psi_h(x)\phi_g(y) \\ \psi_{3,3}(x,y)&=\psi_g(x)\psi_h(y), & \psi_{4,3}(x,y)&=\psi_h(x)\psi_g(y) \end{aligned} \quad (5)$$

and the associated wavelets are given by,

$$\begin{aligned} \psi_i(x,y) &= \frac{1}{\sqrt{2}}(\psi_{3,i}(x,y)+\psi_{4,i}(x,y)) \\ \psi_{i+3}(x,y) &= \frac{1}{\sqrt{2}}(\psi_{3,i}(x,y)-\psi_{4,i}(x,y)) \end{aligned} \quad (6)$$

for i = 1, 2, 3.

When we combine the two parts of the dual tree into a complex basis function (and it's conjugate) then we also separate positive frequencies from negative frequencies. The real and imaginary parts of each complex wavelet are oriented at the same angle, and the magnitude of each complex wavelet is an approximately circular bell-shaped function [11]. A set of six complex wavelets can be formed by using wavelets as the real parts and wavelets as the imaginary parts. Figure 8 illustrates a set of six oriented complex wavelets obtained in this way.

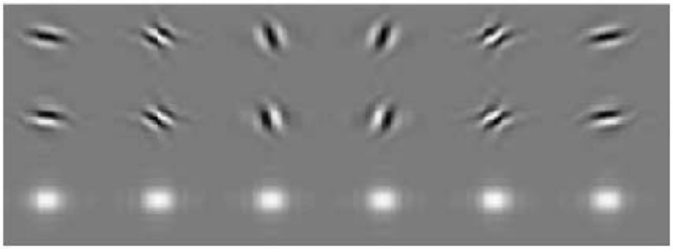


Fig 8: Impulse responses for 2-D DT-CWT: First row is interpreted as the real part and the second row as imaginary part of the complex wavelet. The third row shows the magnitude response (same for both, real and complex wavelets).

Implimentation and Results

Image Database

The Image Database of 1000 color images of 10 different subjects and variable sized 100 images of each subject like sunflowers, dinosaurs, elephants, beaches, mountains, roses, buses, horses etc is used. Some of the sample database images shown in figure 9.

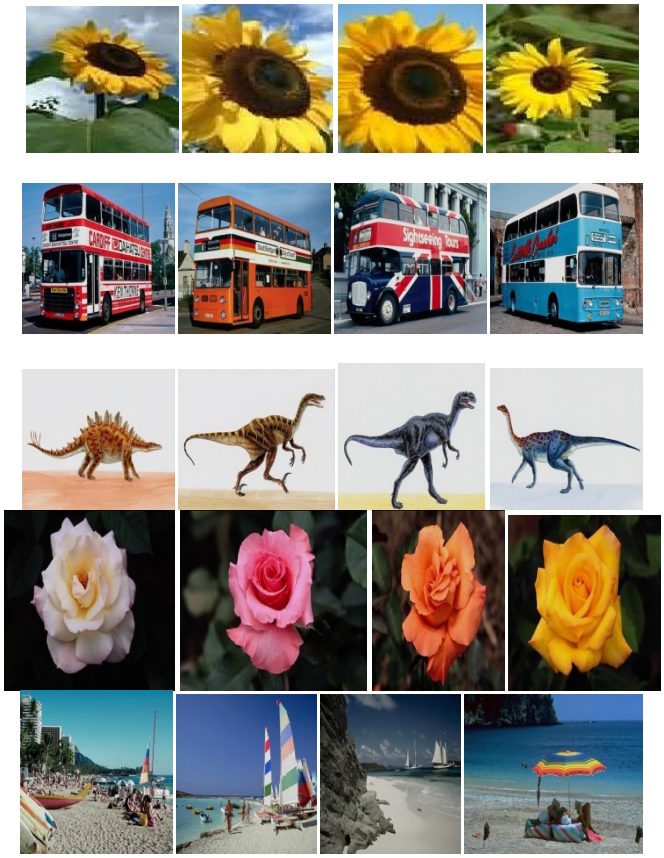


Figure 9: Sample Database images.

Similarity Measures

Similarity measures also termed as distance metric plays important role in Content based image retrieval. Content-based image retrieval calculates visual similarities between a query image and images in a database. Therefore, the retrieval result is not a single image but a number of images ranked by their similarities with the query image. Different *similarity measures* will affect retrieval performances of an image retrieval system significantly so, it is important to find best distance metric for CBIR system. The query image will be more similar to the database images if the distance is smaller. If x and y are two d-dimensional feature vectors of database image and query image respectively, then the distance metrics are given by[12],

i. Minkowski distance,

$$d_{\min}(x,y) = \sqrt[p]{(x_i - y_i)^{1/p}} \quad (7)$$

ii .Cityblock distance,

$$d_{MAN}(x,y) = \sum_{i=1}^{d} |x_i - y_i| \quad (8)$$

The Cityblock Distance was proposed in [13] for computing the dissimilarity scores between color images.

iii. Canberra Distance,

$$d_C(x,y) = \sum_{i=1}^{d} \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (9)$$

Image Features

Each image from database was decomposed using DT-DWT and DT-CWT. The analysis was done upto fourth level of decomposition and two different feature sets are computed using DT-DWT and DT-CWT respectively. The query image feature and database image features are compared with Minkowski, City block and Canberra distance metrics.

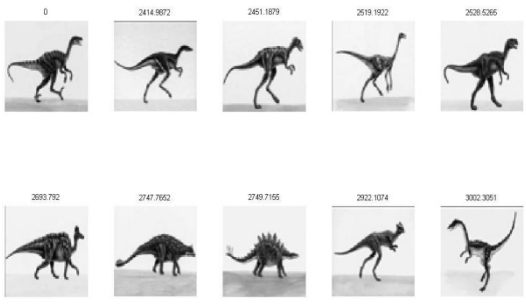
DT-DWT based Feature Extraction

All the variable sized database images are resized to 64x64 to reduce the size of feature vector. The input image is filtered through two trees, tree 1 and tree 2. For both the trees the filters used for first stage are different than the remaining stages. In this implementation Farras filters are used for first stage and Kingbury's quadrature shift filters are used for the remaining stages of tree. Finally, all the sub bands of both the trees at Fourth level (J=4) are concatenated to form a feature vector. Retrieval of similar images was done by comparing database feature vector with feature vector of query image using similarity measures.

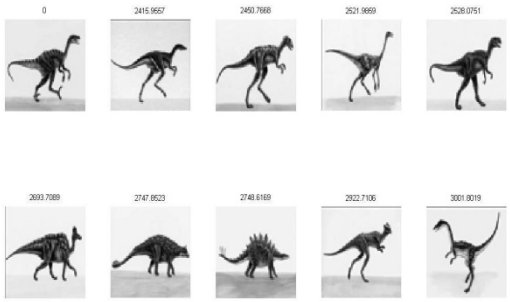
DT-CWT based Feature Extraction

The feature extraction based on DT-CWT is similar to that of DT-DWT. However, in this case, the feature vector size is doubled. Initially, the input image is filtered using the same set of filters defined for tree 1 and tree2 in DT-DWT based decomposition. The input image is further filtered using same set of filters, with the low pass and high pass filters of both the trees exchanged. Finally, all the sub bands of both the trees at level *J* are concatenated to form a feature vector of the image. The table I gives the figures of feature vector size for DT-DWT and DT-CWT for different decomposition levels.

(a)



(b)



(c)

Fig.10: (a) Query Image, (b) Top 10 matches using DT-CWT, (c) Top 10 matches using DT-DWT.

Table I

Level of Decomposition		First	Second	Third	Fourth
Feature Size	DT-DWT	8192	2048	512	128
	DT-CWT	16384	4096	1024	256

Table II

Image Subject Name	Percentage Average Retrieval Accuracy	
	DT-DWT	DT-CWT
Sunflower	68.34	71.43
Bus	52.12	53.26
Dinosaur	93.67	95.00
Elephant	46.57	47.04
Horse	49.96	51.23
Rose	92.09	93.78
Mountain	50.37	52.42
Beach	50.84	53.04
Dishes	40.85	41.74
Historic Building	43.41	43.90

The average retrieval accuracy for various image classes is shown in Table II. It was observed that the retrieval accuracy is excellent for Dinosaurs and Roses class of the images and it is good for Sunflower and average for remaining classes of the images. Also the performance of retrieval is better for DT- CWT than the DT-DWT technique. But retrieval time is little more for DT-CWT due to double feature vector size.

Conclusion

We have presented the novel approach for shift invariant image retrieval using set of dual-tree discrete wavelet transform (DT-DWT) and dual-tree complex wavelet transform (DT-CWT) to evaluate performance of content based image retrieval system. We have used 1000 images database of 10 different classes. The query image feature and database image features are compared with Minkowski,

Cityblock and Canberra distance metrics. It was observed that similarity measure not affect much on retrieval accuracy, but Canberra distance metric gives better results as compared to Minkowski or Cityblock metrics. Experimental results indicate that the proposed method gives excellent retrieval accuracy of 95% for Dinosaur class of images. The Roses class gives 93.78% retrieval accuracy. Also, for other classes retrieval accuracy is good. The proposed method achieves improvement in retrieval accuracy than using DT-DWT filters.

References

[1] Y. Rui and T. S. Huang, "Image retrieval: Current techniques, promising

[2] directions and open issues," *J. Vis. Commun. Image Represent.*, vol. 10,

[3] no. 4, pp. 39–62, Apr. 1999.

[4] Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.

[5] S. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7, pp. 674-693, July 1989.

[6] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, CBMS series, Philadelphia, 1992.

[7] Nick Kingsbury, "Complex Wavelet for Shift Invariant Analysis and Filtering of Signals", Journal on Applied and Computational Harmonic Analysis 10, pp 234-253, 2001.

[8] N G Kingsbury, "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters", Proc. 8th IEEE DSP Workshop, Bryce Canyon, Aug 1998.

[9] N G Kingsbury: "The dual-tree complex wavelet transform: a new efficient tool for image restoration and enhancement", Proc. EUSIPCO 98, Rhodes, Sept 1998.

[10] I. W. Selesnick, R. G. Baraniuk, and N. Kingsbury. The dual-tree complex wavelet transforms - A coherent framework for multiscale signal and image processing. *IEEE Signal Processing Magazine*, 22(6):123-151, November 2005.

[11] A. F. Abdelnour and I. W. Selesnick. Nearly symmetric orthogonal wavelet bases. *In Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, May 2001.

[12] N. G. Kingsbury. Image processing with complex wavelets. *Phil. Trans. Royal Society London A*, September 1999.

[13] T. Celik, H. *Ö*zkaramanli, and H. Demirel, "Facial feature extraction using complex dual-tree wavelet transform," *Computer Vision and Image Understanding*, vol. 111, no. 2, pp. 229–246, 2008.

[14] M. Swain and .D. Ballard, "Color indexing," lnternatiunal Journal of Compurer Vision, Vo1.7.no.1, 11-32, (1991).

[15] Sharmin Siddique, "A Wavelet Based Technique for Analysis and Classification of Texture Images," Carleton University, Ottawa, Canada, Proj. Rep. 70.593, (2002).

[16] M.Kokare, P.K.Biswas, and B.N. Chatterji, "Texture image retrieval using new rotated complex wavelet filters," IEEE Trans. Syst., Man, Cybern. vol. 36, No. 6, pp. 1273-1282, Dec. 2006.

Sanjay Patil received B.E. and M.E. from Shivaji University, Kolhapur, and Maharashtra, India in 1992 and 2003 respectively. He is doing Ph.D. at Shri Ramanand Tirth Marathawada University, Nanded, and Maharashtra, India. From 1994 to 2010 he worked as faculty member in Department of Electronics Engineering in Engineering colleges affiliated to University of Mumbai. Since 2010 he is working as Associate Professor at Jaywant College of Engg and Management, Killemachhindragad, Tal- Walwa, Dist-Sangli, Maharashtra. His field of interest is Digital Signal and Image Processing.
Email Id: sanjayashri@rediffmail.com

Sanjay Talbar received B.E., M.E. and Ph.D. from Shri Ramanand Tirth Marathawada University, Nanded, and Maharashtra, India in 1986, 1990 and 2000 respectively. Since 1987 he is working as faculty member at SGGS college of Engg and Tech., Nanded, Maharashtra, India. Presently he is working as Professor at SGGS college of Engg and Tech., Nanded, Maharashtra, India. He has published no of papers in National and International Conferences and Journals. His field of interest is Digital Image Processing and Embedded Systems.
Email id: sntalbar@yahoo.com

Location Based Information Delivery in Tourism

Jitendra Sharma[#1], Sunil Pratap Singh[#2] and Preetvanti Singh[#3]

[#]Department of Physics & Computer Science, Dayalbagh Educational Institute, Dayalbagh, Agra, India
E-mail: [1]jksharmamca@gmail.com, [2]sunil_pratap@rediffmail.com, [3]preetvantisingh@gmail.com

Abstract
Context in Mobile tourist information systems is captured as the current location of the user. Traveling to a country every one has a number of locations where they want to travel. Therefore a traveling plan is necessary to select nearby point of interests like accommodation places, tourist places, transportation facilities, medical facilities.

In this paper a Location based information delivery system is designed. The influence of such a richer context model on the user interaction for both the capturing of context and context-aware user/device interaction is discussed. The basics of context in this work are location, time of day, personal preferences and device type describing how these basics are leveraged to become habituated web-based information that is delivered to mobile tourists.

Keywords: Context Awareness, Mobile Information System, Delhi Tourism, User Interface Design.

Introduction

Mobile phone and the Internet have revolutionized the communication and with it the life style of people. An ever-increasing number of mobile phone and Personal Digital Assistants (PDA) allow people to access the Internet where ever they are and when ever the want. From the Internet they can obtain on one hand information on events (cinema, concerts, and parties) and on the other hand information on places (city maps, restaurants, museums, hospitals).

Context-awareness is very critical for mobile users. These users have at their disposal devices that are not very advantageous to interactivity (e.g., small screens, space constrained keyboards). Context-awareness can help to improve user interaction by knowing a priori the user's situation, personal preferences, information interests, and environment conditions, so that the user doesn't have to specify these constraints, and information delivery is automatically adapted to his state of affairs.

Tourism Context Challenges

We consider three facets of context: (1) the concepts of context pertinent to a mobile tourist application; (2) the issues one faces when managing context data; and (3) the usage of the context data. We will discuss how each of these challenges affects the interface and interaction design.

Concepts of Context

For a mobile information system, several aspects of context can be considered, such as the characteristics of the particular mobile device (storage and screen size) and network (bandwidth and peers), context of the application (requirements in storage download and display capability), context of the user of the system (e.g., time, location, and interests), context of information objects (e.g., location).

Context Management

For the management of context we distinguish four tasks:
Modelling: A mobile information system needs an open hierarchical approach to context modeling, that is, context should be explicitly modeled on several levels to support further change. Here we consider only the application aspect in detail.

Observation: The application context is assumed to be relatively static. New services might be on offer depending on the location of the user. For these, an infrastructure must be provided to dynamically integrate or release services. Most general object contexts can be pre-captured. The observation of the user context and the evaluation of the current context of the objects are more demanding.

Storage: Data about the user context may have to be available on the mobile device as well as on the server. The data needs to be stored and distributed in an efficient way.

Access: For access to (context) data, the same issues arise as for storage: efficiency and privacy. Context-awareness can be used to reduce the amount of data to be accessed and distributed by pre-selecting the pertinent data.

Usage of context

Usage and benefit of context information depends on the quality that the system designer aims for:
Effectiveness: Effectiveness in the system design of a mobile information system means that the pertinent (i.e., right) information is delivered to the user in a way that they are satisfied with the service.

Efficiency: The system's interaction with the user should not be impaired by data storage and exchange. This affects the implementation of the context-dependent selection of the communication model, communication partners, and local storage on the mobile system as well as context-aware pre-caching strategies and display options.

Related Work

With the evolution of Internet technologies web-based tools for tour planning can now be easily made available, implemented and become a valuable resource for the traveling community and tourists. Ando and Mimura et al [13] developed a travel time information system for the road users and by using the historical data of Toyota City and analyzed the effects of the factors such as what day of the week, what time period of the day, weather. Berger et al [6] described an e-Tourism environment based on a community-driven approach to foster a lively society of travellers who exchange travel experiences, and recommend tourism destinations. Chiu et al [4] proposed a Collaborative Travel Agent System (CTAS) based on a scalable, flexible, and intelligent Multi-Agent Information System (MAIS) architecture for proactive aids to Internet and mobile users to provide tourist information and services such as airlines, hotels, tour operators effectively during trips or even in the planning stage.

Idris and Yahaya [1] discussed the design and implementation of a prototype web information system that used web aggregation as the core engine. Rao et al [11] outlined the design and development of a prototype web-GIS Decision Support System (DSS), CRP-DSS, for use in resource management and assessment of environmental quality.

Schernthanner and Asche [7] provided a basis for discussion of a generic approach to housing market analysis based on free open source geo-information systems. Park et al [3] developed a smart context-aware self guided tour assistant as a context aware real world application. Loh et al [15] presented a recommender system that helps travel agents in discovering options for customers, especially those who do not know where to go and what to do. Kumar et al [12] presented a GIS-based Advanced Traveler Information System (ATIS) for Hyderabad City, India.

O'Grady and O'Hare [9] presented a vision of how an electronic context-aware tourist guide might operate. Dunstall et al [14] described a prototype travel recommender system called the Electronic Travel Planner (ETP), which prepares travel itineraries for tourists. G.M.P. O'Hare, and M.J. O'Grady [5] introduced the design of Gulliver's Genie a context-aware tourist guide that assists roaming tourists.

Raento et al [10] developed ContextPhone, using an iterative, human-centered design strategy. Nie et al [16] designed and implemented the tourist route planning and navigation system (TRPNS) based on Location Based Services (LBS).

J. Kjeldskov and J. Paay [8] presented the design of a context-aware mobile information system prototype facilitating sociality in public places: Just-for-Us. Vasiljevic et [2] al presented web-based dynamic maps from geographic sources promoted on the internet that could be later integrated into official websites.

Proposed System Architecture

The System architecture that supports a tourist information service is shown in Figure 1. This architecture is Web service-based and the major elements include:

- A thin client device that hosts a Web browser.
- A web application server that delivers web content customized to the user's context.
- A UDDI (Universal Description, Discovery and Integration) services directory that provides users with a centralized registry of tourist information services (e.g., a restaurant finder service).

A context manager that keeps track of the user's dynamic context e.g., location, wireless device features) as well as the user's preferences.

A collection of web services that deliver tourist content (e.g., landmark information, restaurant locations, etc). Each Web service has a WSDL (Web Services Description Language) document in XML format that describes the Web service's interface and gives a concrete binding to a network address.

When a new user registers himself in the system, he logs on to the application server and enters his preferences (e.g., restaurant and accommodation preferences). These preferences are then forwarded for storage to the context manager. If the client device possesses a GPS receiver then the client sends location updates to the context manager at regular intervals. If the user doesn't have a GPS receiver then at the point of the information inquiry he enters his location manually in the form of a city name or zip code.

In our implementation we used Microsoft's .NET Framework, a native XML Web services platform. Microsoft .NET includes ASP.NET, a framework for creating application servers that support dynamic web pages, standard Web service technologies like SOAP and WSDL, as well as a multi-language development environment (including C#). One of our Web services was implemented in C#.

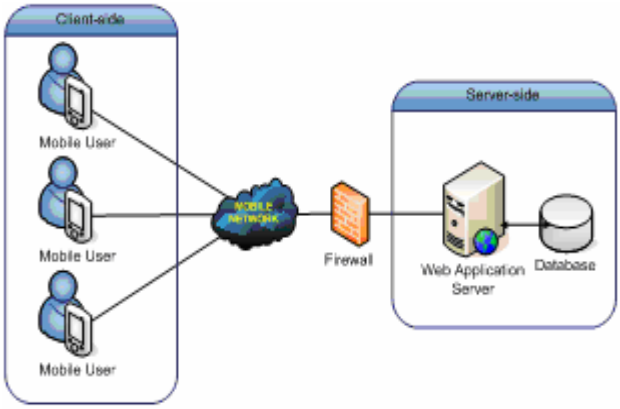


Fig.1 Architecture of Mobile Web-based System

Modelling the User

A great challenge in delivering well-targeted information to the user is to make a high-loyalty profile of the user's interests and preference. However, users switch roles and conditions swiftly-varying mode of transport, role and interest. For example, a traveler may meet up with a friend in a city, and unpredictably have access to a car and an addition set of interests and constraints. On the other hand, one more user may be primarily visiting a city for business purpose, but a Saturday evening free for leisure normally they would inquire

about a club, but due to their work commitment the next day, they pursue a less exhausting alternative. Thus one of the great questions is how to capture changes in the user's job, Such as business context or personal context.

Mobile User Interface

The mobile user interface design makes the user's interaction as simple and efficient, in terms of accomplishing user goals.

For the developed system, it has been designed using Microsoft Visual Web Developer 2008 for presenting and viewing the tourism data on the Web. Figure 2 and Figure 3 present screen shoots of user interface which allows the users to select and input the query criteria in order to view the tourism data they want.

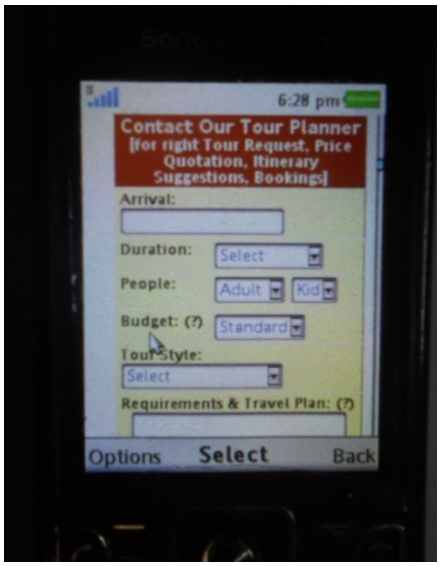


Fig. 2 User Interface Screen-1

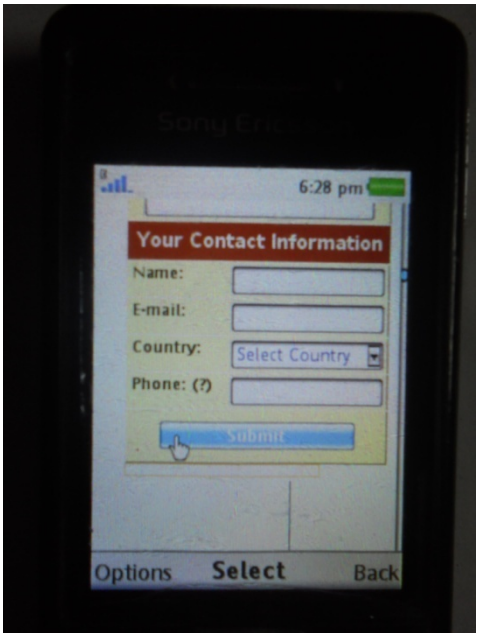


Fig.3 User Interface Screen-2

Personal Adaptation

An important aspect of context-awareness is the system's ability to deliver information that is personalized, i.e., custom-made to user's specific needs. The personalization aspect includes system awareness of the user's information interests and service preferences. A tailored list of restaurants that meets the user preferences is generated as shown in figure. Upon receiving the list of restaurants from the web services, the application server can modify directly this list only restaurants that meet user pre-specified preferences. Initially, the list of restaurant is modified to include preference levels as indicated in user profiles and distance from the user's current location. The result of this step is a modified list with Restaurant elements that include preference level and distance sub-elements.

```
<Restaurant>
    <Name>Ramji Bhai Restaurant</Name>
    <Street>7404 Shivaji Park</Street>
    <City>New Delhi</City>
    <State>NCT of Delhi</State>
    <Country>India</Country>
    <PINCode>110032</PINCode>
    <Phone>(011)22110889</Phone>
    <Cuisine>South Indian</Cuisine>
    <PriceHigh>70</PriceHigh>
    <PriceLow>40</PriceLow>
    <Latitude>00.0000</Latitude>
    <Longitude>00.0000</Longitude>
    <OpenTiming>08.00</OpenTiming>
    <CloseTiming>11.00<CloseTiming>
    <Distance>1.786</Distance>
    <PrefenceLevel>2</PreferenceLevel>
</Restaurant>
```

Advantages of System

One of the main advantages of Internet is its ability to provide almost unlimited access to information to anybody and anywhere, who has technical possibilities to connect with the Web. The developed system will provide the tourists to answer the fundamental questions such as near-by facilities, finding route, searching tourist places of interest etc. Using this kind of system increases convenience and efficiency in tourism activities by providing location information in order to save money, manpower and time.

Conclusion

In this paper; System deliberate to provide tourism information for tourists visiting to Delhi has been developed. The development of System has followed all essential and mandatory steps from capturing data to publishing on the web. The data is stored in a data base and contain historical, cultural, geographical, administrative and hospitality related information in order to be accessed by the tourists through mobile Internet to improve the convenience, wellbeing and efficacy of their travel.

References

[1] A.Z. Idris, and N.A.Yahaya, "Design and Implementation of an Aggregation-based Tourism Web Information System", IJCSNS International Journal of Computer Science and Network Security, vol. 9(12), pp. 143-148, 2009.

[2] D. Vasiljevic, S.B. Markovic, T. A. Hose, B.Basarin, L. Lazic, V. Stojanovic, T. Lukic, N. Vidic, G. Jovic, S. Janicevic, and D. Samardžija "The Use of Web-Based Dynamic Maps in the Promotion of the Titel Loess Plateau (Vojvodina, Serbia), a Potential Geotourism Destination" GEOGRAPHICA PANNONICA Vol.13(3), 78-84, 2009.

[3] D.J. Park, S.H. Hwang, A.R. Kim and B.M. Chang "A Context-Aware Smart Tourist Guide Application for an Old Palace" International Conference on Convergence Information Technology, 2007.

[4] D.K.W. Chiu, Y.T.F. Yueh, H.F. Leung, and P.C.K. Hung, "Towards ubiquitous tourist service coordination and process integration: A collaborative travel agent system architecture with semantic web services", Inf Syst Front, vol. 11, pp. 241-256, 2009.

[5] G.M.P. O'Hare, and M.J. O'Grady "Gulliver's Genie: a multi-agent system for ubiquitous and intelligent content delivery" Computer Communications, Vol. 26, pp.1177–1187, 2003.

[6] H. Berger, M. Dittenbach, D. Merkl, A. Bogdanovych, S. Simoff, and C. Sierra, "Opening new dimensions for e-Tourism", Virtual Reality, vol. 11, pp.75-87, 2007.

[7] H. Schernthanner, and H. Asche. "The Potsdam Housing Market: A GIS-based Spatial Analysis using FOS", REAL CORP 2010 Proceedings, Tagungsband Vienna, 18-20 May, 2010.

[8] J. Kjeldskov and J. Paay "Just-for-Us: A Context-Aware Mobile Information System Facilitating Sociality" Mobile HCI'05, September 19–22, Salzburg, Austria, 2005.

[9] M. O'Grady and G.M. P. O'Hare "Accessing Cultural Tourist Information Via A Context-Sensitive Tourist Guide" Information Technology & Tourism, Vol. 5 pp. 35–47, 2002.

[10] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen "ContextPhone A Prototyping Platform for Context-Aware Mobile Applications" IEEE CS and IEEE ComSoc,pp.51-59, 2005.

[11] M. Rao, G. Fan , J. Thomas , G. Cherian , V. Chudiwale, and M. Awawdeh "A web-based GIS Decision Support System for managing and planning USDA's Conservation Reserve Program (CRP)" Environmental Modelling & Software Vol. 22 pp. 1270-1280, 2007.

[12] P. Kumar, V. Singh, and D. Reddy "Advanced Traveler Information System for Hyderabad City" IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 6(1), 2005.

[13] R. Ando, and Y. Mimura, "A Study to Develop an Information Providing System on Travel Time", Int. J. ITS Res., vol. 8, pp. 77-84, 2010.

[14] S. Dunstall, M. E. T. Horn, P. Kilby, M. Krishnamoorthy,B. Owens, D. Sier, and S. Thiebaux "An Automated Itinerary Planning System For Holiday Travel" Information Technology & Tourism, Vol. 6, pp. 195–210, 2004.

[15] S. Loh, F. Lorenzi, R. Saldana, and D. Licthnow "A Tourism Recommender System Based on Collaboration And Text Analysis" Information Technology & Tourism, Vol. 6, pp. 157–165, 2004.

[16] Y.F. Nie , X. Fu , and J.X. Zeng "A Tourist Route Planning and Navigation System Based on LBS" Proceedings of the International Symposium on Web Information Systems and Applications (WISA'09), pp. 288-290, 2009.

Knowledge Management Application of Internet of Things in Construction Waste Logistics with RFID Technology

Lizong Zhang[1], Anthony S. Atkins[2] and Hongnian Yu[3]

Faculty of Computing, Engineering and Technology, Staffordshire University,
Octagon, Beaconside, Stafford ST18 0AD, United Kingdom
E-mail: [1]l.zhang@staffs.ac.uk, [2]a.s.atkins@staffs.ac.uk, [3]h.yu@staffs.ac.uk

Abstract
The Internet of Things (IoT) is an emerging concept and that is currently under creation and development. However, it has already made an impact on many research domains by providing new solutions and ideas particularly in waste management and recycling. The IoT concept has provided a new research path that is conducive with public awareness of environmental protection considerations and improving recycling rates. This paper focuses on plasterboard waste as an example to propose a smart waste management framework. The 3 layers of the IoT model has been extended to 4-layers by splitting the application layer into knowledge management and visualization layer respectively. A smart waste management application has been developed, based on a case study of a local SME waste recycling company. This smart waste management system uses a service science approach, and it not only provides full logistical records for waste transportation but also provides waste collection arrangements and incident handling guidance to both management and operational staff.

Keywords: Knowledge Management Systems, RFID Technology, Plasterboard Waste, Smart Waste Management, Internet of Things, Service Science

Introduction

The Internet of Things (IoT) is a emerge concept that was first proposed by Kevin Ashton in 1999 to describe an emerging global, Internet-based information service architecture [1]. This concept is currently under development, and there is no generally accepted definition, but IoT is receiving considerable attention and influences virtually all areas of academic and business, and waste disposal is no exception to the Internet of Things.

Waste management and recycling has become an issue, which is being addressed by both developed and developing countries. Inappropriate handling may result in serious environment concerns, or even disasters. Some waste which appears to be safe may release harmful components if not treated correctly, and plasterboard waste is an example as mixing it with domestic waste in landfill sites can result in H_2S gas being emitted.

Regulation introduced in November 2008 prevents plasterboard waste being landfill with normal waste. It has to be treated as 'high gypsum waste' and disposed of in special mono cell designed landfill sites, which results in a significantly higher disposal cost. Currently DEFRA (Department for Environment Food and Rural Affairs) indicates that less than 10% of plasterboard waste is being recycled in the UK[2]. A target has been set by the Waste Framework Directive that aims to recover at least 70% of construction and demolition waste by 2020 [3]. Currently, there are only 4 recycling facilities available, which results in transportation issues and this is a barrier for increasing recycling. Consequently a solution for improving recycling rates is currently being sought by both government and environmental pressure groups.

An achievable solution to waste disposal is provided by the concept of 'Internet of Things'. In generally, the basic idea of IoT is enabling smart environments to recognize and identify objects, and retrieve information from the Internet to facilitate their adaptive functionality [1]. Based on the understanding of IoT and plasterboard waste management issues, a solution is proposed in this paper for auditing and tracking plasterboard waste and providing real-time instruction and job scheduling for operating staff using RFID technology and rule-based knowledge management technology. This 'Smart Waste Management System' undoubtedly could help to prevent fly-tipping and illegal disposal of waste, and result in improved recycling rates in the UK.

Internet of Things and Smart Waste Management

There are many different definitions of the 'Internet of Things' proposed by different organisations and countries. The EU, China, Korea, USA and Japan are the main supporters and promoters of IoT, but there are many different names given to this concept, for example: 'Sensation China'- China, 'Smart Earth'- USA, 'U-Korean'- Korea and "Ubiquitous Computing" – Japan, etc. All of these research entities propose their own definition of IoT, and also the architectures in their project, which governments are funding in their respective countries. Consequently, the definition and architectures they propose are mostly 'localized' to match the demands of their respective societies and investment expenditures.

In Europe, there are about 8 different research groups on the IoT. All of them are integrated under the umbrella of the European Research Cluster on the Internet of Things (IERC) funded by the European 'Seventh Framework Programme'. These groups are working together and focus on different

aspects of IoT, with the aim of building an acceptable IoT prototype by 2020 [4].

A definition of 'Internet of Things' proposed by CASAGRAS2, a research group under the IERC is '*a global network infrastructure, linking physical and virtual objects through the exploitation of data capture and communication capabilities. This infrastructure includes existing and evolving Internet and network developments. It will offer specific object-identification, sensor and connection capability as the basis for the development of independent federated services and applications. These will be characterised by a high degree of autonomous data capture, event transfer, network connectivity and interoperability'*. [5].

Comparison of definitions from other regions of the world shows the concept is similar but is expressed in different ways. For example, in China, the IoT is receiving considerable attention compared with other countries because the Chinese Premier Jiabao Wen visited the Wuxi IoT centre in August 2009 and initiated the launch of a new project 'Sensing China'[6]. The Chinese government's official definition of the 'Internet of Things' is '*expanding application and network extension of communication network and internet, it uses sensor technology and intelligence device in perception of the physical world, and communication, computing, process mining and knowledge via network, realising the information exchange and seamless connection between 'people and things' or 'things and things', and result in real-time control, management and decision making to the physical world.*'[7]

The two definitions mentioned concern the key concepts in IoT: identifying objects, data capture/mining, network and applications. Consequently, the general idea is similar and it can be concluded as: making the physical world object recognisable by the network and enabling exchange of information itself. The network could provide some function to control or manage the physical world. In addition, the architectures from the different research group are similar, even if these are different in detail, and the outlined structure can be described in a three layer model: the sensing layer, network layer and application layer respectively [8].

Plasterboard waste management is an application area that could adopt the IoT to improve the recycling rate. Based on the understanding of IoT and the waste management issues, a 'Smart Waste Management System' with a 4-layer structure is proposed, which follows the IoT three model but is extended to adopt knowledge management technology and visualisation 'localised' for the construction industry.

The 4-Layers System Architecture

The 4-layer structure is an extension of the three layer model of 'Internet of Things', which is illustrated in Figure 1. This structure consists of 'Data Acquiring' layer, 'Data Integration' layer, 'Knowledge Management' layer and 'Visualization' layer respectively. The two main technologies adopted in this framework are the RFID technology and knowledge management technology[9].

The first layer is referred to as the 'Data Acquiring' layer which represents the sensing layer in IoT three layer model. This layer is the route for acquiring the data and information about the physical world; it mainly relies on the RFID technology. In the waste management scenario, the 'objects' referred to the IoT model are the waste or vehicles with the identifying function attachment. In this framework, RFID is an ideal solution because it can provide a low cost tracking and tracing system for contaminated waste in the construction industry. The proposed use of passive RFID tags provides adequate read range using low price (<5p) passive tags for monitoring waste container movement. As the vehicle or waste containers are RFID tagged, they then became the 'Objects' in IoT that could be recognized by the system[10].

The data captured in the Data Acquiring layer is not only from the RFID technology, but also from other sources, such as human input data and online weather reports etc. This 'raw' data only has meaning in its domain, for example, RFID data by itself is meaningless unless it is classified and linked to some 'object' and the information contextualised such as a tagged vehicle enter a recycling site at specific date and time etc. The lowest layer is only responsible for collecting data and the different sources in this layer are virtually independent.

Integrating data from the 'Data Acquiring' layer to provide information is the aspect of 'Data Integrating', which is located in the second layer. This layer can be referred to as the 'Network' layer in IoT model which has more functionality. It is not only responsible for 'network' but also for 'data storage'. The data from the different sources are stored in the database within organised positions where the next layer could request this information[11].

The 'Data Integrating' layer stores the data individually from the lowest layer and then links them by extracting the important parts to store in the central database. The information stored in the central database is a combination of all data sources and will be used as 'fact' for the next layer of knowledge management. For example, waste ID, waste type, weather conditions, time packed and comments from operational staff, will be stored together.

The third layer is the 'Knowledge Management' layer, which is responsible for generating logistical and tracking support including collection arrangements, incident solutions and also providing guidance for operational staff. This feature is supported by the adopted rule-based knowledge management technology, based on information and data from the lower layer. It processes this information and sends the reasoning result to the upper layer[11].

The Rule-based Knowledge Management system is the major part of this layer. This layer contains two main components: the knowledge base, which is used for storing domain knowledge in production rule format; and the fact base, which stores the current situation that equates to the central database in the lower layer. The core of this layer is a reasoning engine that can apply the domain knowledge to the current situation and then generate an acceptable result.

The results need to be translated to human readable text or diagrams, this is the requirement of the highest 'Visualization' layer[11]. The first important feature is the explanation, provided by the 'explanation mechanism' of the Knowledge Management System. This is responsible for explaining the result to the user and also the reasoning procedure required. In general, the features of this layer are translating the results and displaying on a terminal. However, the type of terminal can

vary and could include computer, mobile phone, PDA or special designed terminal. The display media could be short message, web pages and/or client program. Consequently the display on the target client/terminal device is another feather of this layer. The third feature of this layer is the communication function between the users and system, with all users' commands and operations of the system going through this layer.

The Knowledge Management Layer and Visualization Layer correspond to the Application Layer in the IoT model, where it is the main feature to achieve the system functionality. The lowest two layers are only for data collection, storage and management. Therefore, in this 4-layer framework, the upper two layers can be changed or modified to match different application and domain areas.

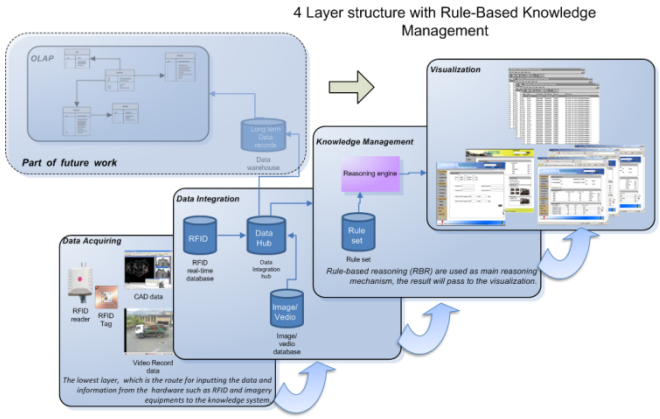


Figure 1: The 4-layer Structure of Waste Management System[11]

Structure of Smart Plasterboard Waste Management System

The data collection technology mainly relies on RFID technology, which is an automatic identification technology and the successor to barcode systems for providing identity of objects. It can be applied in asset tracking, access control and security aspects etc.

In the smart plasterboard waste management system, RFID plays two key roles: firstly, the object identification, and secondly, location information provider. The reason for choosing RFID as the data collection technology is mainly based on low price and the application environment. The passive RFID system that is used in the prototype provides a cheap tag costing about less than 5p and it could be cheaper if larger volume (>1000) are purchased. The low price enables the disposable feature of the tags, which can be applied to individual waste bags for either recycling or disposal of the waste. The RFID system can overcome the difficulties of barcode and/or written identification systems as it is less prone to contamination from the waste and/or weather conditions.

The system structure is illustrated in Figure 2, which is developed for a refurbishment and/or demolition site scenario. The RIFD tags could be attached to each bag of waste and the key 'gates' are installed with RFID readers. The system can then collect the movement data of any waste container and a central server would store the data for use in the knowledge management system.

The Knowledge Management System

There are two main features of the smart waste management system: firstly it generates the waste collection arrangement which is based on the company's contracts and previous collection records. The second is to help the operating staff deal with the waste including normal guidance and incident handling. To achieve these features, knowledge management technology has to be adopted into the system.

Rule-base Knowledge Management technology is introduced as the key part for applying the intelligence functions to the system by reasoning the fact data along with the rules. Figure 3 is a technology view of the system, which illustrates and focuses on the knowledge management system.

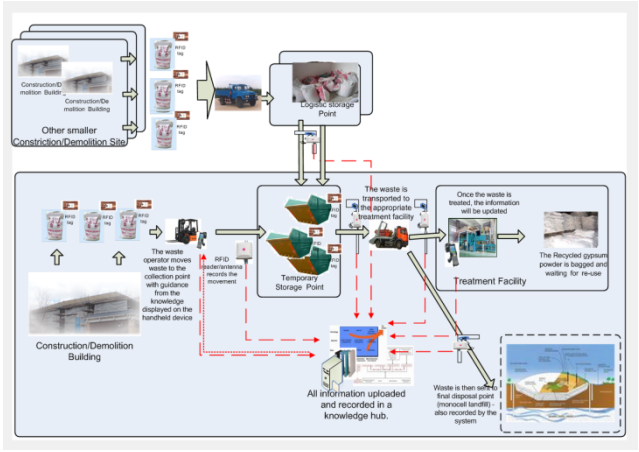


Figure 2: Proposed System Structure of the Waste Management System[11]

There are three databases to support the knowledge management function, two of them have been described in Section III: the 'Records' database is the central database in the framework, and stores the logistic records together with the reasoning results; The 'Rule' database is refer to as the knowledge base. Another database, called the 'Fact' Database, stores the variables and parameters that are acquired from the 'Records Database', and are used in the reasoning procedures.

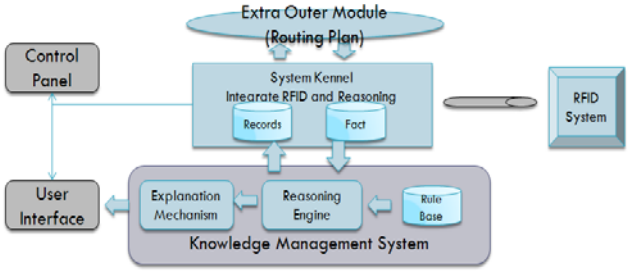


Figure 3: Technology View of System Structure.

The knowledge management system includes two other important components, the 'Reasoning Engine' and the 'Explanation Mechanism'. The 'Reasoning Engine' is the most important part that controls the progress of reasoning, and finally generates 'reasonable' results. The result is then passed to 'Explanation Mechanism' before it is displayed to the users. The 'Explanation Mechanism' is responsible for translating the data/code format result to human readable information, and it is located in the visualisation layer of the framework.

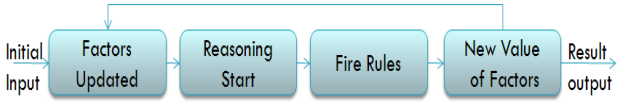


Figure 4: Reasoning Procedure[12].

The Reasoning Procedure

The Knowledge management reasoning procedure is started by data changes in the fact base. This is called 'Data Driven' Reasoning or 'Forward Chain' Reasoning. The fact base stores all the parameters and variables used in the conditions. The conditions are undoubtedly a part of a rule. Consequently, any change of data in the fact database will trigger a rule to start verification or fire. There is an exception in that some variables are used only to store the outcomes of the rules, and not linked to any condition.

Figure 4 shows the general reasoning procedure. The first trigger of reasoning normally starts from the main system kernel when it finds the need for reasoning. Once the system determines the need for reasoning, it will update parameters that relate to the current situation to the fact database. The update action must go through the Knowledge Management System and thus can be monitored and a reasoning cycle is triggered. A typical example is when an unexpected record is found or the time is due for planning the next period of a collection schedule.

Figure 5 illustrates further details concerning the reasoning procedure. After the initial updating is completed by the system, the following reasoning procedure is then controlled by the Knowledge Management system. The updated fact can be addressed by the reasoning engine and the conditions containing this fact will be determined. The next step is finding out the rules that are affected by these conditions, then the rules are verified and false rules are discarded and only valid rules are passed to the next step. The ideal situation is that only one rule has been left to be fired, but usually multiple rules are still present. Therefore, a conflict resolution method has to be applied to ensure that only one rule can be fired. In general, the fired rule will generate some facts and their value will be updated into the fact base, and the update action will trigger the next turn of reasoning until there is no rule to fire or no fact to be updated. The final status of the records in the fact base is consider as the result and will be passed to explanation mechanism with the reasoning procedure records to be translated to user readable information.

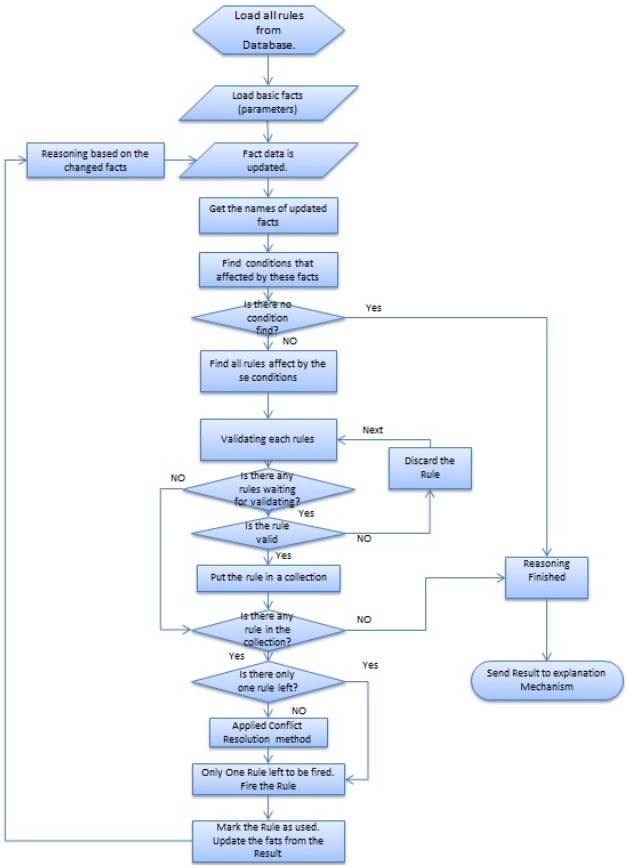


Figure 5 Flowchart of Reasoning Engine Procedure.

The Conflict Resolution Strategy

Three strategies for solving rule confliction are introduced to the prototype system: the longest match strategy, simple priority strategy, and certainly factor respectively. The three strategies are introduced to ensure that only one rule can be fired in each cycle and only one can be activated at the same time in a reasoning cycle.

The first method is the longest match strategy which aims to find the reasoning path that contains the most conditions. In this strategy, the reasoning engine will try every possible path of the reasoning chain, but only chose the path which has the most conditions as the final path, and passes the result as the final goal. Therefore, the reasoning engine needs to spend time trying those paths and also needs to create temporary records for these results, and consequently, the speed is slower than the simple priority strategy.

The simple priority strategy is the fastest as it does not need to try all the paths but only compares the priority in the current stage. In this strategy, each rule has been assigned a 'priority' value, which is an attribute of the rules and assigned when the rule base is built. Once conflict happens, the highest one will be fired, and others will just simply be ignored. If the priority is the same or there are two valid rules, then the first rule will be fired.

The third method is the 'Certainly Factor' (CF), which is a concept from uncertainly reasoning. The Certainly Factor is an extra attribute that is assigned to the facts represents the 'level of believes'. The reasoning engine will fire all match

rules in one cycle and each result will be stored in the fact database with the CF. After the reasoning cycle is completed the output is not a simple conclusion that has multiple possible situations/results with it's certainly factor built into it, but the system will intentionally choose the highest CF, or display all the result for users' query and judgment.

Demonstration for an Example Scenario

A prototype application was developed based on an SME waste recycling company. The company is a local waste recycling company that has about 30 vehicles including skips, trucks and vans etc. It recycles most types of waste from the local area, including wood, metals, cardboard, construction and also plasterboard waste. This company allocates vehicles to pick up waste from contracted sites and deliver to their recycling centre. Two members of staff are engaged in arranging the collection schedule on a manual daily basis, usually the day before, and the schedules are distributed to the operating staff and drivers at the start of their shift. The order of site scheduling is random and determined by the drivers based on their personal experience.

This procedure initially worked satisfactory in the past but with increased growth it is experiencing difficulties. The sites schedule is amended sometimes due to major regular customers requesting urgent pickups and/or additional empty waste container deliveries to site resulting in some customers being delayed by a day.

This case study scenario uses approximately 30 vehicles and 30 sites including the depot and recycling centre together with a plasterboard manufacturer which is also considered as a site because it's the final destination of the recycled product[11]. The system was developed based on this company's requirements focusing on automatic collection schedules for the drivers together with operational staff instructions. The system also includes additional features for audit and monitoring the tonnage of plasterboard waste movement, which could be used for verification of waste treatments and landfill for public scrutiny and government authentication.

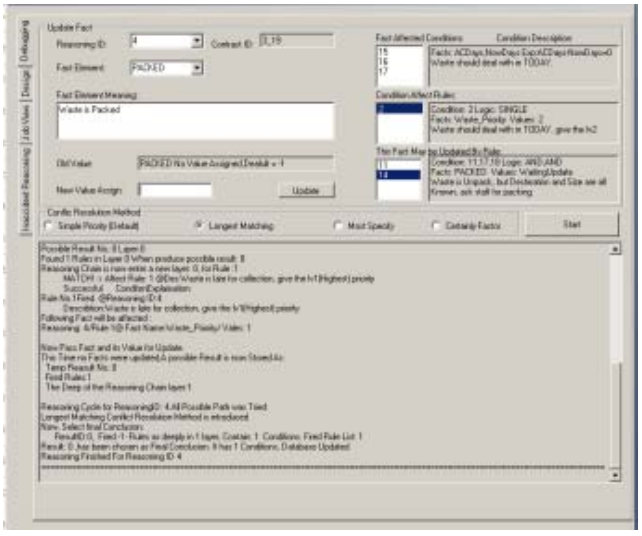
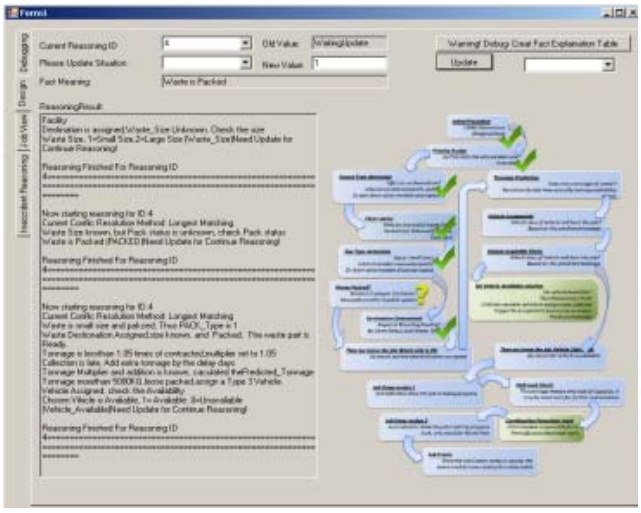


Figure 6(a): Debug Interface



(b) The User Interface.

```
Now starting reasoning for ID:4
Current Conflic Resolution Method: Longest Matching
Waste is small size and pakced, Thus PACK_Type is 1
Waste Destionation Assigned,size known, and  Packed,  The waste part is
Ready.
Tonnage is lessthan 1.05 times of contracted,multiplier set to 1.05
Collection is late, Add extra tonnage by the delay days
Tonnage Multiplier and addition is known, caculated thePredicted_Tonnage
Tonnage morethan 5000KG,loose packed,assign a Type 3 Vehicle.
Vehicle Assigned, check the Availablilty
Chosen Vihicle is Avaliable, 1= Avaliable. 0=Unavaliable
(Vehicle_Available)Need Update for Continue Reasoning!

Reasoning Finished For Reasoning ID
4============================================================
============================================================
========

Now starting reasoning for ID:4
Current Conflic Resolution Method: Longest Matching
Vehicle is OK
Waste and Vehicle both OK, Process enter STAGE 1
Process enter STAGE 1,caculate the total Transport Capacity
Tonnage more than 1/2 Capacity,Process Enter STAGE 2
Creating Stage is 2, and Route planed status is unknow, Check the status
Route planed(1) or Not (0) (RoutePlaned)Need Update for Continue
Reasoning!

Reasoning Finished For Reasoning ID 4
```

Figure 6 (c): Reasoning Procedure Explanation /Report.

Figure 6 illustrates the Knowledge Management System interface, which shows the reasoning procedure and the final report. Figure 6a is debug interface which displays the reasoning procedure in detail; and each step is listed with corresponding explanation. Figure 6b is a concise report to indicate the current stages and reminders of any missing information.

The reasoning explanation report explains the system decisions, and is outlined in Figure 6c: 'Tonnage more than 5000KG, loose packed, assign a Type 3 Vehicle'. It also shows the report of the reminders action to be made by the user as depicted (in Figure 6c): 'Vehicle Assigned, check the Availability; Chosen Vehicle is Available, 1= Available. 0=Unavailable (Vehicle_Available) Need Updated for Continue Reasoning!'

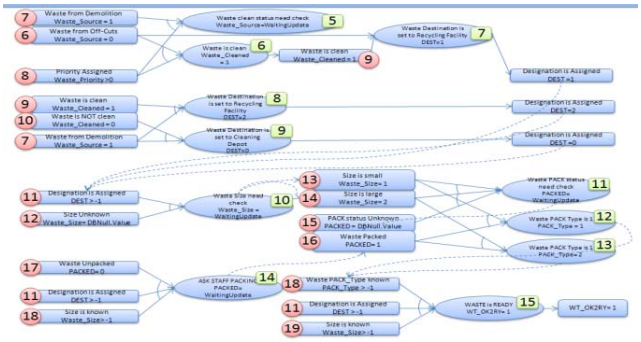


Figure 7. Part of the Reasoning Chain

The domain knowledge was gained from interviews and converted to 73 rules of which 40 rules were jobs creating and reminder for incident handling. Each rule represents a piece of knowledge, and all the rules in the rule base together constitute the domain knowledge, by rules' 'chain' structure. Figure 7 is an example of part of the reasoning chain that shows the relationship between the rules.

Conclusions

This paper discussed the 'Internet of Things' application to smart waste management system, and outlines the research motivation of this project. The UK has a plasterboard waste disposal issue because of the hazardous reaction of land filled plasterboard waste with organic waste producing H_2S gas emissions. The imbalanced geographic distribution of recycling facilities within the UK results in high transportation cost of recycling both in term of time and finances. Based on the understanding of the 'Internet of Things' concept, this paper proposes a 4-layer framework for a smart waste management system based on the 3 layer model of IoT, but extended to match the application domain.

From the framework and 'Internet of Things' concept, a smart waste management system solution is proposed based on a case study from a local SME waste recycling company. The system is a combination of RFID technology and Rule-Based Knowledge Management technology. The system can schedule the waste collection and provide guidance to operation staff either as normal or incident handling instruction, such as road hold ups or vehicle breakdowns etc. in addition, the logistic records of the waste can also be used for tracking waste movement which can provide verification and authentication for both public and government scrutiny.

Reference

[1] Weber Rolf H., "Internet of things - Need for a new legal environment?," *Computer Law & Security Review*, vol. 25, pp. 522-527, 2009.

[2] DEFRA, "Waste Strategy for England 2007," Department for Environment, Food and Rural Affairs, 2007.

[3] DEFRA, "Government Review of Waste Policy in England 2011," Department for Environment Food and Rural Affairs, 2011.

[4] Vermesan Ovidiu and Friess Peter, "European Research Cluster on the Internet of Things," in *Internet of Things - Globe Technology and Societal Trends*, O. Vermesan and p. Friess, Eds.: River Publishers, pp.67-99, 2011.

[5] Furness Anthony, "CASAGRAS DRAFT White Paper WP6:Applications in the emerging Internet of Things," 2009.

[6] Sensing_China, "Premier Jiabao Wen visit Wuxi Technology Centre," 2009.

[7] MIIT_of_China, " Internet of Things White Paper of China (2011)," China Academy of Telecommunication Resarch of MIIT, 2011.

[8] Zheng Lirong, Zhang Hui, Han Weili, Zhou Xiaolin, He Jing, Zhang Zhi, Gu Yun, and Wang Junyu, "Technologies, Applications, and Governance in the Internet of Things," in *Internet of Things-Global Technological and Societal Trends*, O. Vermesan and P. Friess, Eds., 2011.

[9] Zhang Lizong, Atkins Anthony S., and Yu Hongnian, "Application of RFID Technology and Knowledge Hub for Logistic Support in Scrapped Type Recycling," in *9th Informatics Workshop for Research Students* June, Bradford UK: pp.190-195, ISBN 978 1 85143 251 6, 2008.

[10] Atkins Anthony S., Zhang Lizong, and Yu Hongnian, "Application of Knowledge Hub and RFID Technology in Auditing and Tracking of Plasterboard for Environment Recycling and Waste Disposal," in *10th International Conference on Enterprise Information System- ICEIS* Vol. IV, pp.190-196 . ISBN:978-989-8111-48-7, 2007.

[11] Zhang Lizong , Atkins Anthony S and Yu Hongnain "RFID Technology in Intelligent Tracking Systems in Construction Waste Logistics using Optimisation Techniques," in *4th SKIMA (International Conference of Software, Knowledge and Information on Management Applications)* Paro, Bhutan 2010.

[12] Zhang Lizong, Atkins Anthony S., and Yu Hongnian, "RFID Technology in Smart Management System of Internet of Things in Construction Waste Logistics," in *2011 2nd International Conference on Management Science and Engineering (MSE 2011)* Chegndu, China: Southern Illinois University Carbondale, In press, 2011.

Application of RFID Technology in e-Health Management and Outsourcing in Bhutan

Atkins. A.S.[1], Lhamo D.[1, 2] and Yu. H.[1]

[1]Faculty of Computing, Engineering and Technology, Staffordshire University, Octagon, Beaconside, Stafford ST18 0AD, United Kingdom
[2]Department of Electronics and Communication, Royal University of Bhutan, College of Science and Technology, Phuentsholing, Bhutan

Abstract
The evolution of medicine and Information Technology has enabled emerging health informatics to be used in healthcare system in innovation ways. This has given rise to recent e-health applications that support public access to e-**health** facilities. This paper focuses on e-health management and its introduction in Bhutan through the use of RFID technology and also the use of digital health records together with communication between the health care-givers and patients over long distances (referred to as outsourcing). This paper proposes a health management system using RFID technology that tags patients, health records, health workers and equipment for an automated database update system for tracking and tracing.

Index Terms: auditing system, e-health, outsourcing, RFID technology, Bhutan.

Introduction
Recent evolution in Information Technology in health informatics has enabled us to look at existing healthcare systems in new ways, the ultimate goal being delivery of the best possible health care services for anyone, at anytime from anywhere. E-health is a recent term used in relation to healthcare provision '*combined use of electronic communication and information technology in the health sector; the use in the health sector of digital data - transmitted, stored and retrieved electronically - for clinical, educational and administrative purposes, both at the local site and at distance*' (Della Mea, 2001).

Public access to e-health applications is growing but e-health care services will not be used unless both patients' and clinicians' expectations and experiences are taken into account during their design and adoption (Gustafson and Wyatt, 2004). Recent improvements in IT technologies have significantly aided health care sectors in different parts of the world and it is now important to develop initiatives to improve service to patients. The Commission of the European Communities indicated that the e-Health market is currently some 2% of total healthcare expenditure in Europe and has the potential to more than double in size (Stroetmann et al, 2006; Sahar and Asi, 2009). E-health tools have tremendous potential to encourage people to adopt healthy behaviours to promote disease prevention, healthy life style, and early detection (Neuhauser and Kreps 2010).

People are increasingly using e-health applications, particularly the internet, to seek health information, to communicate with others who have a similar disease or illness, to receive prevention messages and health promotion advice, and to communicate with healthcare providers. Trends in 2000 indicated that Internet users seeking health information and healthcare services would more than double from 2000 to 2005, reaching 88.5 million people (Shumaker et al, 2009).

Areas of health informatics
According to Svensson and Per-Gunnar, 2002, there are three areas of health informatics based on the predominant type of user or use (Svensson and Per-Gunnar, 2002).

Consumer informatics: This category focuses on the patient and public communications regarding health topics.

Medical and clinical informatics: this category is related to health care structure, processes and outcomes. The main application is medical records like computer based personal records that will facilitate access to low cost therapies and computer based patient records that will facilitate clinical decision making.

Bio informatics: this category is related to creation and advancement of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from the management and analysis of biological data.

A survey was carried out in California to investigate the rate of adoption of service systems such Electronic Health Record (EHR), Computer Physicians Order Entry Systems (CPOE), Radiology Information Systems (IS), Pharmacy IS, Laboratory IS, Administration IS (AHA, 2008) and the results are depicted in Table 1. Table 1 show a survey of 33 Hospitals in California with the number of responses and corresponding percentages depicted respectively.

Table 1: Service Systems Adopted in California based Hospitals

	Fully Implemented	Implementing	Not Implementing	Total
EHR (Electronic Health Record)	13; 39.4%	12; 36.4%	8; 24.2%	33; 100%
CPOE (Computer Physicians Order Entry)	10; 30.3%	10; 30.3%	13; 39.4%	33; 100%
Laboratory IS	27; 81.8%	2; 6.1%	4; 12.1%	33; 100%
Radiology IS	28; 84.8%	2; 6.1%	3; 9.1%	33; 100%
Pharmacy IS	28; 84.8%	3; 9.1%	2; 6.1%	33; 100%
Administration IS	25; 75.8%	3; 9.1%	5; 15.2%	33; 100%
Radio Frequency Identification	2; 6.1%	5; 15.2%	26; 78.8%	33; 100%

Bhutan is a Himalayan kingdom with a population of about 700,000 and is a developing nation with basic health facilities. Although the country and its citizens have the benefit of free healthcare, because of its difficult terrain and mountainous nature, the health facilities are unavailable in some remote areas. A visit to the best facility that is situated in the capital would take a minimum of one day making it difficult and risky for patients. This paper proposes a possible solution to support better heath care management in Bhutan and discusses the solution using emerging technology such as Radio Frequency Identification (RFID), digital scanning and outsourcing etc.

National Referral Hospital, Bhutan

The main hospital in Bhutan is *The National Referral Hospital* known as the *Jigme Dorji Wangchuck National Referral Hospital* situated in the capital of the country, Thimphu. Since it was established in 1972, the hospital has been supplying free basic medical treatment as well as advanced surgeries and emergency services to citizens from all over the country. It provides the most sophisticated health evaluation and management services in the country.

In the late 1990s, a plan was launched to upgrade the then 175-bed Jigme Dorji Wangchuck Hospital to a National Referral Hospital with assistance from the government of India. Investigation showed the existing structure to have limited functionality and the planners proposed upgrading to a new 350-bed hospital. By 2002 a laboratory building, compound wall, gift shop, doctors' and nurses' quarters and an internal road system was completed.

The hospital caters to the population of the Thimphu district, non-referred patients from neighbouring districts and referred cases from the 20 district hospitals. The hospital not only functions as the National Referral Hospital for the entire country but also functions as:

- Human resources pool and technical backup for hospitals.
- Clinical training centre for the Royal Institute of Health Sciences (RIHS).
- Technical support to the Public Health Programs

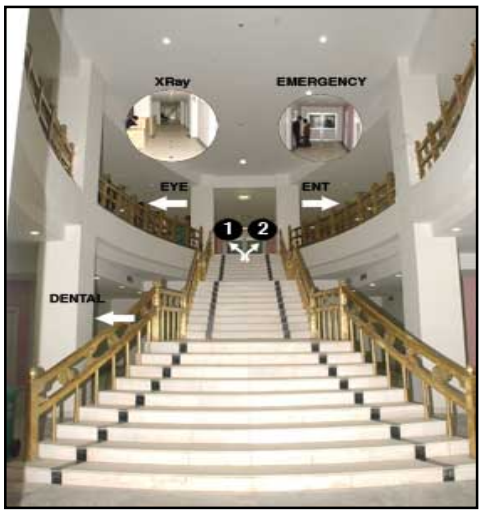


Figure1: The new 350-bed Jigme Dorji Wangchuck National Referral Hospital (JDWNRH) complex in Thimphu.
Source: http://www.kuenselonline.com/2010.

The number of patients attending JDWNRH has considerably increased over the years. This is probably due to greater health awareness among the general population, better accessibility due to improved transport and communication, actual increase in the population and increased services at JDWNRH. This has also led to an increase in referrals from district hospitals as well as referrals outside Bhutan to India etc. The hospital provides services such as accident and outpatient, inpatient, preventive and rehabilitative services together with some special clinics.

Medical records in Bhutan

JDWNRH is the only National Referral Hospital providing tertiary medical care to the people of 20 districts and technical backstopping to all the hospitals and health centers. The medical referral committee of this hospital also refers many patients requiring special investigation, treatment and/or surgery to India and third party countries. Consequently, it is of the utmost importance to maintain individual patient records for all patients, for reference purposes. In 1995 on the invitation of the Health Department Mr. Peter Parslow, who was the Administrative Officer of Mongar Leprosy Mission hospital introduced a computerized patient record keeping system for the In-patients at JDWNRH which was developed in FoxPro software and this was used until December 2004. However, the data currently in the system does not give adequate information and the system is in the process of being upgraded to accommodate new changes but this requires logistical as well as technical support.

There is a vision of a fully computerised medical record keeping system using a local area network (LAN). All patients attending JDWNRH will be registered using a computerised system at the Registration counter and disease and treatment information will be recorded in the Pharmacy unit. This type of computerised Individualized Patient Record (IPR) system will show a patient's lifetime record of hospital visits with all the information concerning patient illness and treatment.

Gradually all the wards, Doctor's chambers and Out Patient Department (OPD) chambers will be fully connected by LAN to enable information to be accessed from authorized health service personnel.

To maintain accurate individual records vital information such as full name, date of birth, National ID number of the patient or ID of parents in case of minors, father's name, permanent Dzongkhag where one's census is recorded, present Dzongkhag working or living at, contact address in case of emergency etc , would be recorded. This is some of the important information a patient or guardian would need to provide at the Registration Desk for the initial registration. However, for subsequent visits it will be relatively easy to retrieve the information just by entering the Patient Registration Number into the system. The primary function of Electronic Health Records (EHR) is to enable the delivery of safer, higher quality and more person-centred healthcare and especially to enable seamless care across the traditional health service boundaries. In providing a more comprehensive picture of health demands and resource utilisation, it also greatly supports the management information function (Department of Health and Children).

RFID System as a solution in Health Care

In Information and Communication Technologies (ICT) there is increasing interest in the use of Radio Frequency IDentification (RFID) in hospitals, which comes from the possibility of obtaining improvements in terms of efficiency, quality of health care treatment and error reduction (Correa et al. 2007, Thuemmler et al. 2007). RFID allows wireless detection of data that can then be automatically stored and retrieved from a database. In a hospital environment the adoption of RFID can assist in information processes and support staff in providing efficient medical services. Nevertheless, any organization that plans to adopt RFID has to face multiple challenges, including overcoming the technological, managerial and organisational problems (Wang et al, 2006). The main goal of the present study is to propose an RFID based service platform for hospitals, which is consistent with a service science driven design approach.

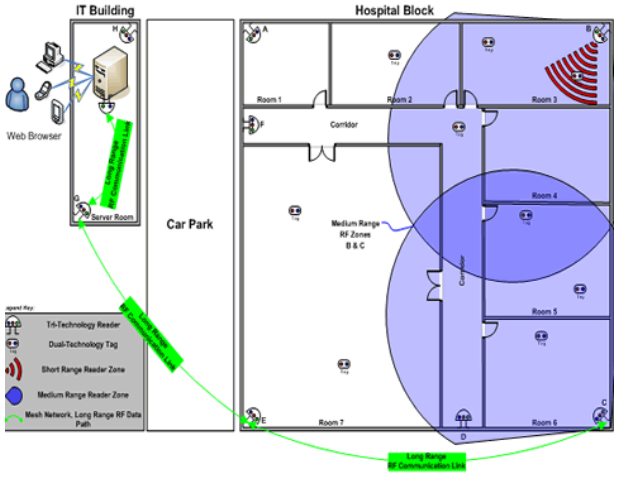


Figure 2: Conceptual Design of a Hospital using an RFID System

RFID systems have been suggested for a number of purposes such as tagging equipment for easy and fast retrieval, tagging patients in order to avoid human error, tagging folders or files of the patients in case a hard copy reference is required.

Figure 2 shows an example of a typical hospital layout. There are RFID reader antennas situated at various points in the hospital to pick up the tags that are either placed on patients or objects (Jervis, 2005). These readers would normally be operating continuously to monitored and record data from tagged objects such as patients and assets etc. This information is then continuously being logged into the database to maintain the hospital records.

The use of RFID provides an integrated solution to provide real time status information on hospital medical assets such as defibrillating equipment and pharmacy issues etc. Reducing the time taken to locate medical equipment for seriously ill patients could improve prognosis and avoid dangerous delays. RFID systems help locate medical equipment tagged with RFID tags within a shorter period of time and reduce manpower required in locating such equipment.

Another use of RFID in hospitals is the automatic update of the patients' information into their records during the required treatment and service provision at the hospital.

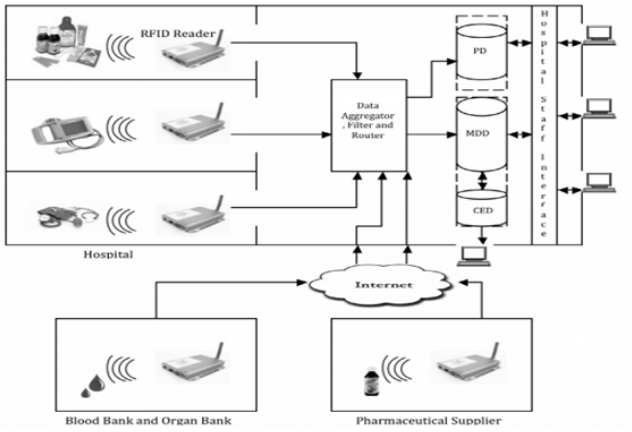


Figure 3: RFID System for Asset Tracking in a Hospital

Figure 3 outlines a proposed scalable platform for asset tracking in the hospital. The asset tracking platform has two major components, namely, Pharmaceutical tracking and Medical Devices tracking. The Medical Devices tracking solution also provides data to the Clinical Engineering Database (CED) of the hospital that enables faster repairs of critical equipment by tracking their usage log. In this platform, all the tracking data is first collected by the readers installed throughout the hospital premises. The readers could collect data from tagged medicines or medical devices or tagged blood containers and organs etc, which are in the vicinity. All the collected reader data is fed into the centralised data aggregator, filter and router. This unit will collect all the data flowing from each individual reader and filter it for false reads, multiple reads etc. Once the data has been filtered, it

will be channeled to the appropriate database. The information collected from the tagged medicines, blood containers and available organs would be stored in the Pharmaceutical Database (PD). The reason to group medicines data with blood containers and available organs data is because of the similarity in the Radio Frequency (RF) nature of these items. Blood or available organs would be retrieved for a patient only on the recommendation of a doctor in a similar way to medicines. Also these items have similar composition (organic and water based), which would result in similar tag responses when the RFID reader enquiries.

The second set of data collected from all the tagged medical devices is sent to the Medical Device Database (MDD). The Medical Device Database is connected to the Clinical Engineering Database (CED), which stores all the repairs and servicing information of the medical devices. The device usage history log can be used in addition to the CED data by the repair engineers for faster malfunction detection. The data stored in MDD, PD and CED is accessible to authorised hospital staff through the Hospital Staff Interface. The blood bank and pharmacies can be linked with the hospital system using similar RIFD readers by tagging their products. The data path between the blood/organ banks and the pharmacies could be through dedicated lines or through the Internet depending upon the infrastructure present in the hospital and capital investment requirements.

Due to the high read rate of RFID tags, the database might need to have improved data filtering and streaming capability. Compatibility with existing systems is essential to encourage wider use in adopting new technology and this can be easily arranged from the migration from bar code reading systems to RFID systems.

Bhutan being a mountainous country with dangerous terrain roads and unreliable weather makes the transfer of seriously ill patients time-consuming, dangerous and risky to the health of the patient. RFID technology along with wireless networks can be used to provide long distance assistance to patients. Thus the patient could still be in one area and receive health related advice from another area in Bhutan. Cases like regular tests (as in the case of diabetic patients) or medical advice and solutions for some minor problems can be dealt without having to travel.

Outsourcing

This paper has described the use of RFID technology for data collection and tracking of medical equipment and patients etc. Digitalising of patient records (Sangwan et al, 2005); treatment and medication etc would also be beneficial. The adoption of digitalised health care systems from the compilation of digital data would also enable transfer of data both internally or to any part of the world through satellite transmission. Figure 4 illustrates an application where a hospital in Bhutan could transfer information internationally for medical support. When a patient consults a local doctor and is advised for a CAT scan (Computerised Axial Tomography), the CAT scan is taken and its details are logged into their personal file in the database. This data can then be sent to another facility overseas for a specialized consultant (Ho and Atkins 2005) e.g. from UK to advise and support the local doctor in Bhutan regarding the patient's treatment and/or operations.

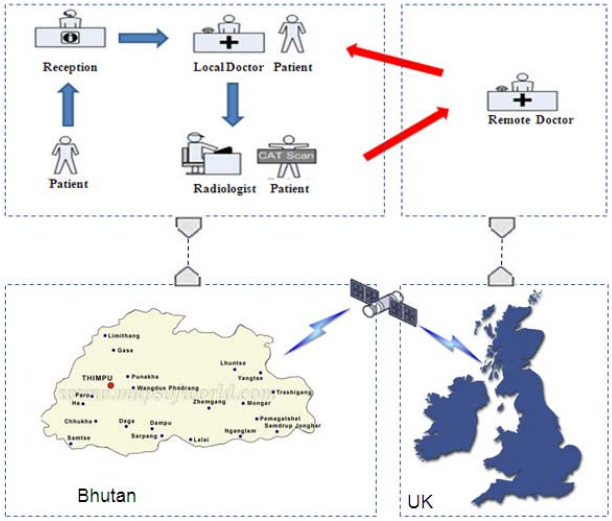


Figure 4: Outsourcing in Healthcare Application

Recommendations

Though the idea of e-health has been readily accepted in peoples' lives, there are number of challenges that need to be resolved:

Enhancing interaction: The majority of patient and clinician encounters take place for purposes of exchanging clinical information. It is estimated that only a small fraction of physicians offer e-mail interaction, a simple and convenient tool for efficient communication, to their patients (Institute of Medicine, 2001) The health messages and advice always originate from the experts and are directed down the chain to consumers. Too often, the authoritative delivery of health information without active consultation with consumers is perceived as intimidating and off-putting (Neuhauser and Kreps, 2003).

Increasing the interoperability: A wide range of consumers and care-givers must be able to work together in the e-health communication system.

Complex health problems often involve collaborative efforts between numbers of health care experts. Though the experts work in different offices, or health care systems, they must be able to have timely access to accurate health records for effective care coordination.

Creating e-health communication that is dynamic and engaging.

Designing communication to have the reach of mass media and the impact of interpersonal connections.

Conclusion

The paper uses the National Referral Hospital, Bhutan as an example to propose an RFID based system which is consistent with a service science driven design approach. The different

steps of the service design methodology have been drawn from the literature. Various operations that could be streamlined by the introduction of RFID technology to improve their efficiency and performance have been showcased. A sequential customisable adoption technique for the hospitals would be introduced depending on priories and finance. This allows the hospitals to optimise the RFID service platform in a way that benefits their operations. Future studies should focus on action research, in order to empirically identify organizational and technical issues concerning the implementation of the service system proposed.

There are numerous applications of information technology to health care practices. The e-health sector covers various topics such as telemedicine, electronic records that provide paperless systems, recruitment, procurement, healthcare score cards, audits, information systems etc. The recent rise in the use and study of e-health offers the advantages of using telecommunication to deliver improved health services. E-heath has the advantages of reducing costs, improving quality, and improving access to care in rural and underserved areas of the country. However the extent of these advantages is largely speculative at this time (Khanna, 2005).

References

[1] AHA. (2008). Continued Progress Hospital Use of Information Technology. American Hospital Association, Retrieved on October 2009, http://www.aha.org/aha/content/2007/pdf/070227-continuedprogress.pdf

[2] Correa, F.A, M.J.A. Gil, L.B. Redin. (2007). RFID and Health management: Is it a Good Tool against System Inefficiencies? International Journal of Health care Technology Management 8(3-4) 268-297.

[3] Della Mea, V. (2001). *What is e-Health (2): The death of telemedicine?* Retrieved December 2, 2010, from Journal of Medical and Internet Research: http://www.jmir.org/2001/2/e22/

[4] Department of Health and Children, authors. *Health Information: A National Strategy*. Dublin, Ireland: Retrieved April, 2011 http://www.dohc.ie/publications/pdf/nhis.pdf?direct=1.

[5] Gustafson D.H.; Wyatt J.C. (2004) Evaluation of e-Health Systems and Services. *BMJ*2004;328: 1150

[6] Ho, L., and Atkins, A. S,(2005). Outsourcing Decision-Making. In H. Kehal, *Outsourcing and Offshoring in the 21st Century: A Socio-economic Perspective* . Idea Group Inc. IRM Press.

[7] Institute of Medicine, authors. *Crossing the Quality Chasm: A New Health System for the 21st Century.* Retrieved April 20, 2011, from Washington, DC: National Academies Press; 2001. Jun 1, http://www.nap.edu/books/0309072808/html/

[8] Jervis, C. (2005). Chips with Everything: RFID and Healthcare. *British Journal of Healthcare Computing and Information Management* , Volume 22 Number 2.

[9] Khanna, S. (2005). *E-health and Telemedicine.* Retrieved December 02, 2010, from Pharmainfo.net; Pharmaceutical Information for You: http:www.pharmainfo.net/reviews/e-health-and-telemedicine

[10] Neuhauser and G.L. Kreps, (2010). E-health communication and behaviour change: promise and performance, *Social Semiotics* **20,** pp. 7–24.

[11] Sahar, F., and Asi, M. (2009). Citizen Access to eHealth Services in County . *Citizen Access to eHealth Services in County. MSC Thesis for Computer Science*.

[12] Sangwan, R.S., Qiu, R.G. and Jessen, D., (2005). Using RFID tags for tracking patients, charts and medical equipment within an integrated Health Delivery Network, *Networking, Sensing and Control, 2005. Proceedings. 2005 IEEE,*, pp1070-1074

[13] Shumaker, S, A., Ockene, J, K., Riekert, K A. (2009). The Handbook of Health Behavior Change, Springer Publishing Company, pp 170

[14] Stroetmann, K, A., Jones, T; Dobrev, A; Stroetmann, V N. (2006). The Economic Benefits of Implemented e-Health Solutions at Ten European Sites, Commission of the European Communities

[15] Svensson, and Per-Gunnar. (2002). ehealth Applications in Health Care Management. *eHealth International* .

[16] Thuemmler, C., W. Buchanan, V. Kumar. (2007). Setting Safety Standards by Designing a Low-Budget and Compatible Patient Identification System based on a Passive RFID Technology. International Journal of Health care Technology Management, 8(5) 571-583.

[17] Wang, Shang-Wei; Chen, Wun-Hwa; Chorng-Shyong Ong, Li Liu and Yun-Wen Chuang, (2006). RFID Application in Hospitals: A Case Study on a Demonstration RFID Project in a Taiwan Hospital, *System Sciences, HICSS. Proceedings of the 39th Annual Hawaii International Conference,* pp184

Comparative Analysis of Static and Dynamic CMOS Logic Design

Rajneesh Sharma[1] and Shekhar Verma[2]

[1]Asst. Prof., [2]Lecturer,
Electronics Engineering Department, Domain Robotics, Lovely Professional University, Jalandhar (PB) India

Abstract
The choice of the CMOS logic to be used for implementation of a given specification is usually dependent on the optimization and the performance constraints that the finished chip is required to meet. This paper presents a comparative study of CMOS static and dynamic logic. Effect of voltage variation on power and delay of static and dynamic CMOS logic styles studied. The performance of static logic is better than dynamic logic for designing basic logic gates like NAND and NOR. But the dynamic casecode voltage switch logic (CVSL) achieves better performance. 75% lesser power delay product is achieved than that of static CVSL. However, it observed that dynamic logic performance is better for higher fan in and complex logic circuits.

Index Terms: Static CMOS circuits, Dynamic CMOS circuits, Logic synthesis, Power delay product.

Introduction

It is well known that, for theoretical reasons, dynamic logic is less low-power consuming and have high speed than static logic. In particular, dynamic CMOS gates are supposed to be more advantageous than static ones mainly because of a total absence of output glitching and a reduced parasitic capacitance. However, the need of precharging operations introduces some extra dissipated power that does not affect static CMOS logic. In this project we observe experimentally how the choice of the CMOS technology influences the behavior, in terms of power consumption and delay, of digital circuit. An appropriate choice of logic can lead to design high performance, low power VLSI design.

A comparative study of CMOS static and dynamic logic [1-2] present power consumption which show that the power values for dynamic logic are lower than those for static logic. However the performance comparison on the basis of power delay product is not present so far. Power delay product (PDP) is a fundamental parameter which is often used for measuring the quality and the performance of CMOS logic. As a physical quantity, the power-delay-product can be interpreted as the average energy required for a gate to switch its output voltage from low to high and from high to low. The amount of energy required to switch the output has been calculated as the product of power and delay. It is mainly dissipated as heat when the NMOS and PMOS transistor conduct current during switching. It is desirable to minimize the power delay product (PDP) [3].

$$PDP = P_{avg} \times \tau_{avg}$$

$\tau_{avg} \rightarrow$Average delay

In this paper, static and dynamic 2 input NAND, NOR and dynamic cascode voltage switch logic (DCVSL) NAND are implemented with voltage ranging from 1V to 1.8V. ELDO simulation results for 180nm technology nodes are given.

This paper is organized as follows. Section II presents the operating principles Static and Dynamic logics. Section III compares the performance measures. Section IV discusses the results and section V concludes this paper.

Static and Dynamic Logic

Static logic

Static logic circuits allow versatile implementation [3] of logic functions based on static, or steady-state, behavior of simple CMOS structures. A typical static logic gate generates its output levels as long as the power supply is provided. This approach, however, may require a large number of transistors to implement a function, and may have cause considerable time delay. A basic function of static CMOS logic is explained with example of 2- input NAND gate [3]. There is conducting path between the output node and the ground only if input voltage VA and VB are equal to logic high value. If one of the inputs at low logic value then there is a path between voltage supply and output node is created i.e. except during switching, output connected to either VDD or GND via a low resistance path.

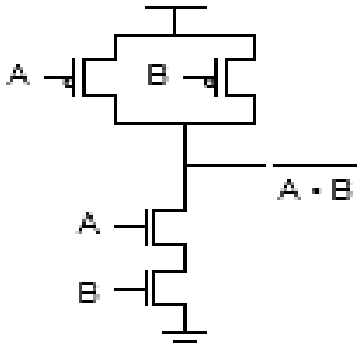


Figure 1. NAND logic using Static CMOS

Basic features of Static CMOS logic are [3]

- Very low static power dissipation
- High noise margins (full rail to rail swing)
- Low output impedance, high input impedance
- No steady state path between VDD and GND
- Delay is function of load capacitance and transistor resistance
- Comparable rise and fall times (under the appropriate transistor sizing conditions)

Dynamic logic

In high density, high performance digital implementations where reduction of circuit delay and silicon area is a major objective, dynamic logic circuits offer several significant advantages over static logic circuits. Fig. 2, shows a generalized CMOS dynamic logic circuit [3]. The operation of all dynamic logic gates depend upon on temporary storage of charge in parasitic [6]. This operational property necessitates periodic updating of internal node voltage levels, since stored charge in capacitor cannot retain indefinitely. Consequently, dynamic logic circuits require periodic clock signals in order to control charge refreshing. In the following, a dynamic CMOS circuit technique which allows us to significantly reduce the number of transistors used to implement any logic function is introduced. The circuit based on first precharging the output node capacitance and subsequently, evaluating the output level according to the applied inputs. The precharge phase is setting the circuit at a predefined initial state while the actual logic response is determined during the evaluation phase [7]. Static CMOS offers good performance but cannot keep up with dynamic logic styles in terms of propagation delay [8]. The shorter delays mostly have to be traded off for increased power dissipation.

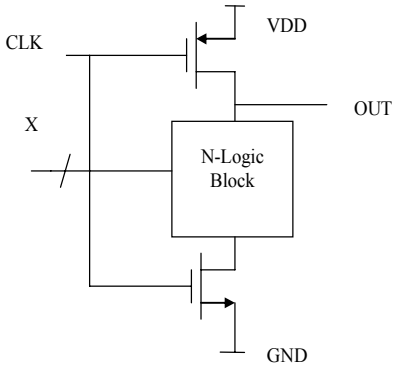


Fig. 2. Generalized Dynamic Logic circuit.

Results and Discussion

The performance analysis of static and dynamic CMOS circuits is carried out. ELDO simulation results for 180nm technology nodes are given. The effect of voltage variation on power dissipation and delay is studied .The result of static and dynamic 2 input NAND, NOR and dynamic cascode voltage switch logic (DCVSL) NAND are given with voltage ranging from 1V to 1.8V. It is observed from simulation result that with increased voltage, power dissipation of the circuits increasing and the delay decreased. Delay is inversely proportional to supply voltage and thereby it increases whereas power directly proportional therefore it decreases. The optimum power delay product for static NOR and NAND is 1.936×10^{-21} and 3.074×10^{-22} respectively as seen (as shown in results) is very less as compare to dynamic NOR and NAND. However Dynamic CVSL results in 6.129×10^{-21} power delay product whereas PDP value of static DCVSL result 20.14×10^{-21}, so dynamic CVSL perform 75% better than static. In the process technology utilized for this analysis, the attractive point for static logic circuit operation lies near 1.4V and for dynamic logic lies near 1V. Therefore it is concluded that dynamic logic can operate at much lower values of supply voltage.

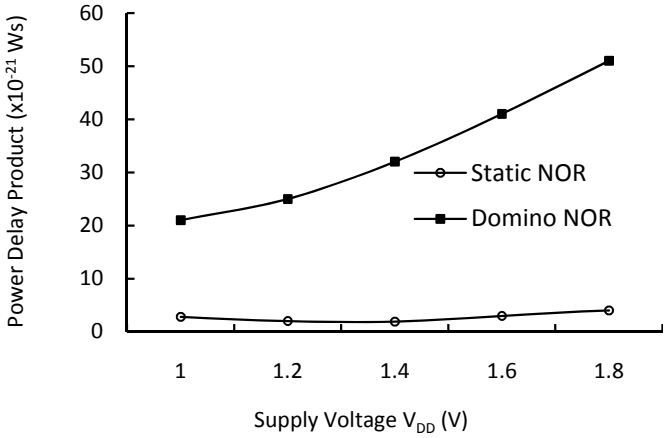


Fig. 3. Variation of PDP with supply voltage

Table 1. Static 2 Input NOR

V_{DD} (V)	Power (pW)	Average Delay (pS)	PDP ($x10^{-21}$ Ws)
1	14.237	193.035	2.748
1.2	20.260	99.362	2.013
1.4	27.773	69.721	1.936
1.6	37.024	78.766	2.916
1.8	48.299	84.957	4.103

Table 2. Dynamic 2 Input NOR

V_{DD} (V)	Power (pW)	Average Delay (pS)	PDP ($x10^{-21}$ Ws)
1	24.318	858.54	20.848
1.2	34.446	747.5	25.748
1.4	46.990	689.7	32.409
1.6	62.342	655.75	40.88
1.8	80.929	634.65	51.361

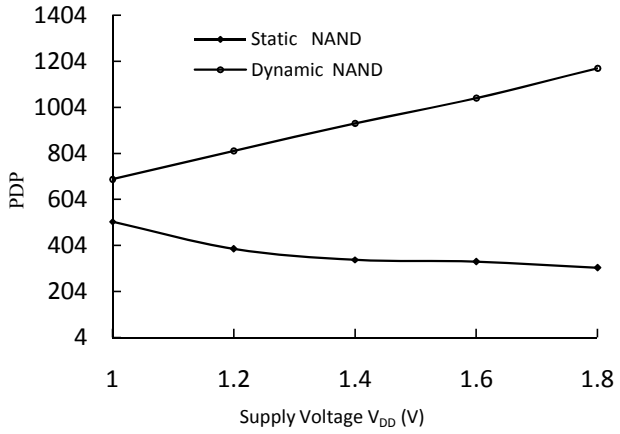


Fig. 4. Variation of PDP with supply voltage

Table 3. Power Delay Product (PDP) of Static NAND

V_{DD}(V)	Power(pW)	Average Delay(pS)	PDP(x10^{-24}Ws)
1	2.919	173.6	506.7
1.2	3.860	69.6	388.9
1.4	4.904	69.6	341.1
1.6	6.047	55.2	333.8
1.8	7.285	42.2	307.4

Table 4. Power Delay Product (PDP) of Dynamic NAND

V_{DD}(V)	Power(pW)	Average Delay(pS)	PDP(x10^{-24}Ws)
1	2.206	313.6	691.5
1.2	2.935	277.7	815.0
1.4	3.752	249.2	934.3
1.6	4.658	224.2	1300.1
1.8	5.648	42.2	1173.7

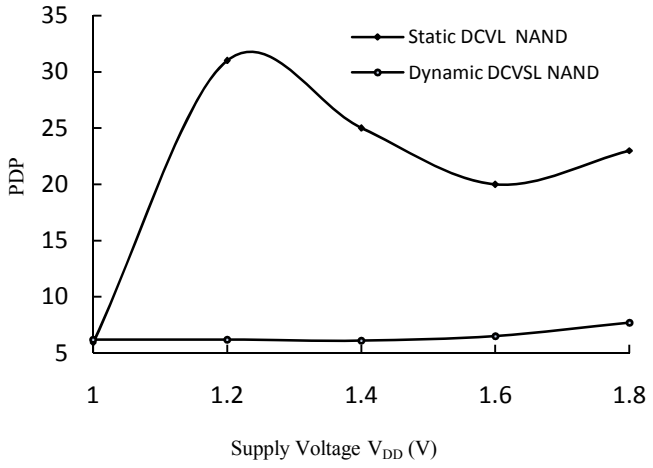


Fig. 5. Variation of PDP with supply voltage

Table 5. Static 2 input DCVSL NAND

V_{DD}(V)	Power (pW)	Average Delay (nS)	PDP(x10^{-21}Ws)
1	20.0715	0.295	5.924
1.2	27.9937	1.115	31.212
1.4	37.655	0.652	24.536
1.6	49.3151	0.408	20.135
1.8	62.9988	0.360	22.707

Table 5. Dynamic 2 Input DCVSL NAND

V_{DD}(V)	Power (pW)	Average Delay (pS)	PDP (x10^{-21}Ws)
1	21.439	290.32	6.224
1.2	30.336	203.05	6.159
1.4	41.368	148.179	6.129
1.6	54.859	119.04	6.302
1.8	71.19	108.08	7.694

Conclusions and Futurework

In this work the impact of voltage variation on delay and power in static and dynamic CMOS circuits has been carried out. It has been observed from the results that the choice of static and dynamic CMOS logic depends upon the requirements of application. For simpler logic implementation, e.g. NAND, NOR etc., we can use static logic because they provide comparable performance with respect to dynamic logic at low cost and less complexity, whereas dynamic logic preferable in complex logic circuit design, e.g. microprocessor, microcontroller etc. The present work is very useful for comparative study of analysis and simulation of static and dynamic CMOS circuits. An appropriate choice of logic along with voltage variation can lead to the design of high performance, low power VLSI chips.

This work shall be further carried out on bigger circuits like XOR, adder etc. so that we can analyze this comparative study more judiciously.

References

[1] E. M. M. Poncino, "Power Consumption of Static and Dynamic CMOS circuits," IEEE, 2nd International Conference on ASIC, pp. 425-427, October 1996.

[2] R. Chandel, Y. Nataraj, G. Khanna, "Performance Analysis of Voltage-Scaled Static and Dynamic CMOS circuits," Nanoelectronics and Optoelectronics, vol. 3, pp. 171-176, 2008.

[3] S.M. Kang and Y. Leblebici, CMOS Digital integrated Circuits- Analysis and Design, Tata McGraw Hill, New Delhi, India, 2003.

[4] M. Kontiala, M. Kuulusa and J. Nurmi, "Comparison of Static Logic Styles for Low-Voltage Digital Design," IEEE, The 8th International conference on Electronics, Circuits and System, vol. no. 3, pp. 1421 - 1424 September 2001.

[5] M. E. S. Elrabaa, "A New Static Differential CMOS Logic With Superior Low Power Performance," in Proc. Int. Sym. Circuits and Systems, vol. no. 2, pp.

810-813 December 2003.

[6] S. Perri and P.Corsonello, "Performance comparison between static and dynamic CMOS logic implementations of a pipelined square rooting circuits," IEEE Proc. Circuits devices system, vol. no. 147, pp. 347-355 December 2000.

[7] T. J. Thorp, G. S. Yee and C.M. Sechen, "Design and Synthesis of Dynamic Circuits" IEEE Transactions on Very Large Scale Integration (VLSI) systems, vol. no. 11, pp. 141-149 , February 2003.

[8] G. Yee and C. Sechen, "Clock-Delayed Domino for Dynamic Circuit Design", IEEE Transactions on VLSI Systems, vol. no. 8, pp. 425-430, Aug. 2000.

[9] N. H. E. Weste and D. Harris, CMOS VLSI Design: A Circuits and Systems Perspective, 3rd Edition, Addison-Wesley, 2005.

[10] F. Grassert, and D. Timmermann, "Dynamic Single Phase Logic with Self-timed Stages for Power Reduction in Pipeline Circuit Designs," IEEE International Symposium on Circuits and Systems (ISCAS), vol. no. 4, pp. 144-147 May 2001.

Reduction of Impulse Noise in images with Adaptive Window Length Recursive Weighted Median Filter

[1]Kiran P. Dange and [2]R.K. Kulkarni

[1,2]Department of ETC, V.E. Society's Institute of Technology, Chembur, Mumbai, India
E-mail:k.p.dange@gmail.com,rk1_2002@yahoo.com

Abstract
A Adaptive window length recursive weighted median filter (ARWMF) for effective reduction of impulse noise is presented in this paper. At high density, lower window size does not remove as well as larger window size may blur the images. So to avoid this, window size of the RWMF is adaptive, based on the presence of the noise density. The performance of proposed algorithm produces better edge and fine detailed preservations and reduces blurring at high density noise.

Index Terms: Impulse Noise, Adaptive Window Size, Recursive Weighted median Filter.

Introduction
Noise removal is one of the major concerns in the field of computer vision and image processing. Images are often contaminated by impulsive noise due to noisy sensors or channel transmission errors or faulty storage hardware. The goal of removing impulsive noise is primarily to suppress the noise as well as to preserve the integrity of edges and detailed information. Averaging filters are having low pass characteristics and therefore they tend to blur the edges and other fine image details. The nonlinear filtering techniques have been performed better and produce satisfactory results [1, 2]. The most nonlinear filter is Standard Median filter which is popular because of its simplicity in implementation and efficient noise removal characteristics [3]. In , Manglem Singh et al.[8] have proposed adaptive rank ordered median filter (AROM) that employs two stage switching schemes. Utilizing rank-conditioned median filter (RCM) and center weighted median (CWM) filter [9, 10]. Lin [4] and Huang[5] proposedsome adaptive algorithms for filtering impulse noise. But these algorithms are more complex and the results are not better compared to proposed algorithm. Two common types of impulse noise are the salt and pepper noise and random – valued noise. For images corrupted by salt-and–pepper noise, noisy pixel can take only the maximum and minimum values in the dynamic range. Let z_{ij} be the gray level of an original image Z at pixel location (i, j) and $[n_{min}, n_{max}]$ be the dynamic range of Z . Let X_{ij} be the gray level of the noisy image X at pixel (i, j) location. Impulsive Noise of density p may then be defined as:

$$x_{ij} = z_{ij} \text{ ; with probability 1-p} \quad \text{----(1)}$$

η_{ij} is the substitute for the original gray scale value at the pixel location (i, j). When $\eta_{ij} = [n_{min}, n_{max}]$, the image is said to be corrupted with Random Valued Impulsive Noise (RVIN) and when $\eta_{ij} \in [n_{min}, n_{max}]$, it is known as Fixed Valued Impulsive Noise or Salt &Pepper Noise (SPN).

In this paper an adaptive window size RWM filter algorithm using median controlled algorithm is proposed which provides high density of noise removal and also preserve image details.

Structure of the Filter
The general structure of the recursive weighted median filter is given as equation

$$Y(N) = \text{Median}\,(|A_1| \Diamond \text{sgn}(A_1) Y(n-1)^{N}_{L=1} + |B_k| \Diamond (X(n-k)^{M2}_{k=-M1})) \quad \text{----(2)}$$

The algorithm is explained in terms of the block diagram (Fig. 1).

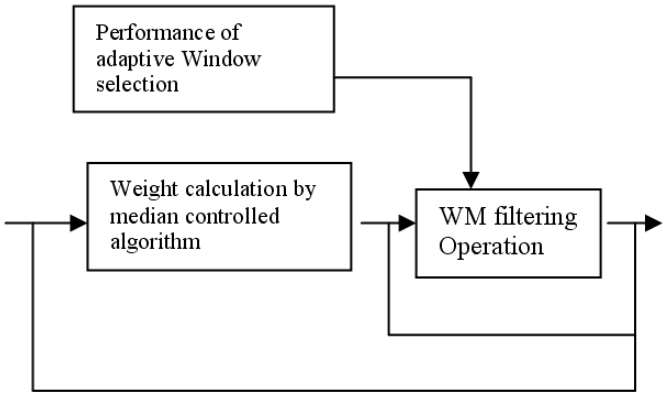


Figure 1: Block diagram of Median controlled adaptive RWM filter

Recursive WEIGHTED MEDIAN filter with adaptive median filter
The standard median filter is not providing edge preservation and fine details of the image. So this proposed filter is providing excellent impulse noise removal as well as preserving edge details. In case of adaptive window size selection, the window size selects based on amount of noise density. Because of this, the unwanted filtering of uncorrupted

pixels is reduced. So the blurring is reduced even at high density noise. In weighted median filter, weights are calculated for every iteration and output of the first iteration is to be reference signal, computing the new weights by comparing the new reference signal to the original signal and computing the output again using the new weights, it can continue the procedure till the number of iterations reached [5]. The following are the steps for this proposed algorithm.

Algorithm: This filtering operation works in two stages.

Stage1: Determination of the window size
Z_{min}=minimum intensity value in Sxy
Z_{max}=maximum intensity value in Sxy
Z_{RWM}=RWM intensity value in Sxy
Z_{xy} =intensity value at coordinates Sxy
The adaptive Recursive weighted median filtering algorithm works in two levels
Level A : If $Z_{min} < Z_{RWM} < Z_{max}$, go to level B
Else increase the window size
If window size · S_{max}, repeat level A
Else output Z_{RWM}
Level B : If $Z_{min} < Z_{xy} < Z_{max}$, output Z_{xy}
Else output Z_{RWM}

Stage 2: Filtering operation:
The Recursive weighted median filtering operation is explained with example.
Steps involved in the Median controlled algorithm are as follows
1. Get the median Filtered image using the sliding window W, store the result in REFERENCE image.
2. Calculate the weights as
Weight (i,j) = exp{ -α (|original(i,j) - Reference(i,j)|) }
Where α > 0.
This will be weight matrix of same size that of image
3. using the above weights perform the Recursive weighted median operation and store the output as reference image.

Extract 3x3 pixels of image and corresponding pixels of weight matrix as well, arrange them into row
Let us consider
Weight = [0.1 0.05 0.2 0.25 0.1 0.1 0.1 0.05 0.05], & image pixels =[1,5,8,11,2, 3, 6, 9,10] .
After sorting, we get the sorted input set with the corresponding weights

Sorted input = [11, 10, 9, 8, 6, 5, 3, 2, 1] and corresponding weights are (**Note:** weights are not sorted, they are arranged according to sorted image pixels)

W = [0.25, 0.05, 0.05, 0.2, 0.1, 0.05, 0.1, 0.1, 0.1]

Now starting from the left, add the weights until the sum reaches or exceeds 0.45. The first weight is smaller than 0.45, so add the next weight. The sum is now 0.5 exceeding 0.45. The weighted median is therefore 8.

i.e Addition of weights till sum reaches or exceeds 0.45.
0.25 +0.05+0.05+0.2=0.55

The sorted input pixel corresponding to last added weight is 8 hence weighted median is 8.

4. The process is done iteratively, so that output image is produced with least mean square error.

Results

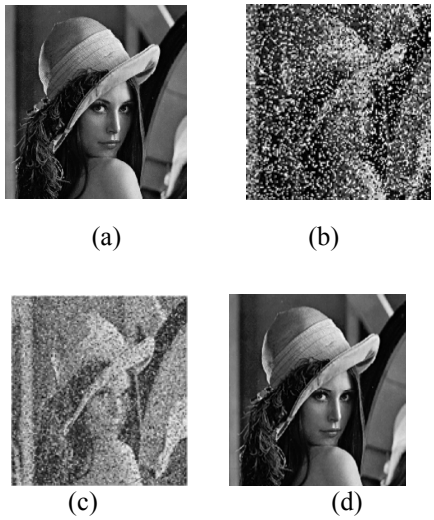
The performance measures of this proposed algorithm such bas Peak signal to noise ratio (PSNR), Mean square error(MSE) and Mean absolute error(MAE) are evaluated using the following formulas[10] :

$$PSNR = 10 \log_{10}(255^2/MSE) \quad \text{---- (3)}$$

$$MSE = 1/(MN) \sum_{ij} (y_{ij}-x_{ij})^2 \quad \text{---- (4)}$$

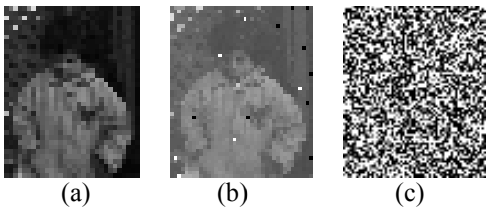
$$MAE = 1/(MN) \sum_{ij} | y_{ij}- x_{ij} | \quad \text{---- (5)}$$

In matlab 7.0 ,the proposed ARWM filter is tested using gray scale images(Lena,pout) show the results,Lena image corrupted by 40% of noise density Fig 2(a,b,c,d,)



(a) (b) (c) (d)

Figure 2: (a) Original 512X512 Lena image (b) Noisy Image (Noise Density =50%) Outputs of (c) WMF (d) Proposed Method



(a) (b) (c)

Figure 3: (a) Original 512×512 Pout image (b) Noisy image (Noise density 80%) (c) Output of proposed Method

Table 1: Comparison table of PSNR of different filter for Lena.tif (Gray scale Image)**.**

Noise Density	WMF	MC Lin's	Proposed Algorithm
10	34.22	34.48	44.04
20	27.08	31.70	43.33
30	21.66	27.53	42.60
40	17.57	22.30	41.93
50	14.22	17.40	41.25
60	11.64	13.72	40.52
70	9.49	10.70	40.07

Table 2. Comparison table of MSE of different filter for Lena.tif (Gray scale Image)

Noise Density	WMF	MC Lin's	Proposed Algorithm
10	20.34	27.24	10.32
20	56.25	41.99	12.15
30	179.56	76.03	14.40
40	444.36	184.96	16.79
50	895.80	466.56	19.64
60	1586.42	1041.99	23.20
70	2524.05	10.70	25.79

Table 3: Comparision table of MAE of different filters for Lena.tif (Gray scale image)

Noise Density	WMF	MC Lin's	Proposed Algorithm
10	2.12	1.08	2.98
20	3.17	1.93	3.42
30	5.70	3.15	3.94
40	10.75	5.79	4.55
50	19.87	12.06	5.33
60	33.45	23.63	6.21
70	52.44	42.88	7.16

Conclusions

The recursive weighted median filter is producing very effective outputs as shown in table 1, table 2 and that produce less blurring effect. Comparison with other types of algorithm shows that this proposed algorithm is having less MSE, MAE. But the drawback of this proposed algorithm is that it has three different computing stages, so the processing time requires more compare with other existing algorithms.

References

[1] I.Pitas and A.N.Venetsanopoulos, Nonlinear Digital Filters: Principles and Applications.Kluwer Academic Publishers,1990.

[2] J.Astola and P.Kuosmanen, Fundamentals of Nonlinear Filtering, CRC Press,1997.

[3] R.C.Gonzalez and R.E.Woods, Digital Image Processing, 2nd ed.Addison Wesley, 1992.

[4] Ho-Ming Lin and Alan, "Median filters with Adaptive Length,"IEEE transactions of the circuits and systems, vol.35, no.6, June 1998.

[5] H.Hwang and R.A.Haddad, "Adaptive median filters:new algorithms and results ",IEEE Transactions on Image Processing,4,pp.499-502,1995.

[6] Lin Yin and Yang, "Weighted Median filters: A tutorial, IEEE transactions of the circuits and systems, vol.43, no.3,March 1996.

[7] Kh.Manglem Singh and Prabin K.Bora, "Rank Threshold median filter for Removal of Impulse Noise From Images."

[8] R.C.Hardie and.E. Barner, "Rank conditioned rank selection filters for signal restoration,"IEEE Trans.Image Processing, vol.3,pp. 192-206,Mar.1994.

[9] S.J.Ko and Y.H.Lee, "Centre weighted median filters and their applications to image enhancement", IEEE Trans.Circuits System,vol.38,no.9,pp.984-993,Sep.1991.

[10] Zhou Wang, Alan Conard Bovik,Hamid Rahim sheik and Erno P.Simoncelli, "Image **Quality** Aseesment: From Error Visibility to Structral Similarity",IEEE Trans.Image Processing ,Vol.13,(2004)

Removal of Impulse Noise with Median based Detail Preserving Filter

Kiran P. Dange

Department of Electronics & Communication, S.N.D.T. Women's University, Mumbai, India
E-mail: k.p.dange@gmail.com

Abstract
For removing impulse noise, detail preserving median based filter is presented in this paper. This algorithm works on, detection of impulse pixel based on threshold value and later corrupted pixels are replaced by the median of the uncorrupted pixels in the filtering window. The proposed algorithm works well up to 90% noise density with less processing time. Implementation result shows that this proposed algorithm gives better results by comparing other different existing filters. The performance of proposed algorithm produces better edge and fine detailed preservations and reduces blurring at high density noise.

Index Terms: Impulse Noise, Median filter, Threshold.

Introduction

Noise removal is one of the major concerns in the field of computer vision and image processing. Images are often contaminated by impulsive noise due to noisy sensors or channel transmission errors or faulty storage hardware. The goal of removing impulsive noise is primarily to suppress the noise as well as to preserve the integrity of edges and detailed information. Averaging filters are having low pass characteristics and therefore they tend to blur the edges and other fine image details. The nonlinear filtering techniques have been performed better and produce satisfactory results [1, 2]. The most nonlinear filter is Standard Median filter which is popular because of its simplicity in implementation and efficient noise removal characteristics [3]. In , Mangle Singh et al.[8] have proposed adaptive rank ordered median filter (AROM) that employs two stage switching schemes. Utilizing rank-conditioned median filter (RCM) and center weighted median (CWM) filter [9, 10]. Lin [4] and Huang [5] proposed some adaptive algorithms for filtering impulse noise. But these algorithms are more complex and the results are not better compared to proposed algorithm.

Noise Model

Two common types of impulse noise are the salt and pepper noise and random-valued noise. For images corrupted by salt-and-pepper noise, noisy pixel can take only the maximum and minimum values in the dynamic range. Let z_{ij} be the gray level of an original image Z at pixel location (i, j) and $[n_{min}, n_{max}]$ be the dynamic range of Z. Let x_{ij} be the gray level of the noisy image X at pixel (i, j) location. Impulsive Noise of density p may then be defined as:

$$X_{ij} = Z_{ij} \text{ ; with probability 1-p} \quad \text{---- (1)}$$

η_{ij} is the substitute for the original gray scale value at the pixel location (i,j). When $\eta_{ij} = [n_{min}, n_{max}]$, the image is said to be corrupted with Random Valued Impulsive Noise (RVIN) and when $\eta_{ij} \in [n_{min}, n_{max}]$, it is known as Fixed Valued Impulsive Noise or Salt &Pepper Noise (SPN).

In this paper, a median based detail preserving filter is presented which provides high density of noise removal and also preserves image details.

Proposed Algorithm

This algorithm is having two stages In first stage, detection of impulse noise is provided and in second stage, the corrupted pixels replace with median value of uncorrupted pixel in the filtering window.

Let X be the noisy image of size M×N. Each pixel X_{ij}, a sliding window of size (2L+1) × (2l+1) centered at X_{ij} is defined. The steps of this algorithm are as follows:

1. Get the noisy image as X .
2. The current pixel X_{ij} to be processed,Y_{ij} is the sliding window of size (2L+1) × (2L+1) centered at X_{ij} .
3. The elements of this window are

$$Y_{ij} = \{ X_{i-u, j-v}, -L \le u, v \le L \text{ ----} \quad (2)$$

4. Apply 5× 5 window for noise detection to the entire image initially.
5. Find the absolute difference between the centre pixel and neighboring pixels in the corresponding window as

$$\delta_{i-u, j-v} = \{1 , 0< | X_{i-u,j-v} - X_{ij} | <= T_1 \text{ ----} \quad (3)$$
$$0 , \text{otherwise}$$

6. Count the no. of pixels whose absolute difference lies between zero to particular threshold ($0 < AD \le T$).For optimum performance T is selected as 25.

$$\zeta_{ij} = \sum_{-L \le u, v \le L} \delta_{i-u,j-v} \text{ ----} \quad (4)$$

ζ_{ij} denotes the number of pixels, which are similar to that of centre pixel

7. Let as assume Ψ_{ij} is same size of the filtering window and assigned to one when ζij greater than a threshold

T_2 .

$$\Psi_{ij} = \{1, \text{for } \zeta_{ij} \geq T_2 \text{ ---} \quad (5)$$
$$0 \text{ , } otherwise$$

Where T_2 is chosen as 4 for optimum performance.
($\Psi_{ij} = 1$ denotes as noise free pixel.)

8. Separation of uncorrupted pixel and corrupted pixels is be done by using following step

$$X^* \Psi = \{ 1, \Psi_{ij} = 1 \text{ ----} \quad (6)$$
$$0, otherwise$$

9. Apply a filtering window of initial size 5×5 to the noisy pixels which has zero value in the matrix (X* Ψ) and replace the noisy pixel with median value of the uncorrupted pixel in the window.

Table 1. Noise density vs. window size

Noise Density(ND)	Window size
0% < ND < 10%	5×5
10% < ND < 40%	7×7
40% < ND < 70%	9×9
70% < ND < 90%	13×13

Results

The proposed algorithm is tested for test images Lena and Elaine of size 512× 512,8 bits per pixel. It is observed that as high noise density as 90%, filtered image preserves edges and fine details. The performance is compared with existing filters such as standard median (SMF), centre weighted median filter (CWMF), Recursive weighted median filter, etc. The performance of this algorithm is calculated with different quality measures such as peak signal to noise ratio, mean absolute error and structural similarity index. They are defined as :

$$PSNR = 10 \log_{10} (255^2/MSE) \quad \text{---- (7)}$$

$$MAE = 1/(MN) \sum_{ij} | y_{ij} - x_{ij} | \quad \text{----(8)}$$

$$MSE = 1/(MN) \sum_{ij} (y_{ij} - x_{ij})^2 \quad \text{---- (9)}$$

$$SSIM(x,y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad \text{---(10)}$$

Where y_{ij} and x_{ij} denote the output and original image. Constant C1 and C2 represent small constant are added to avoid instability [7].

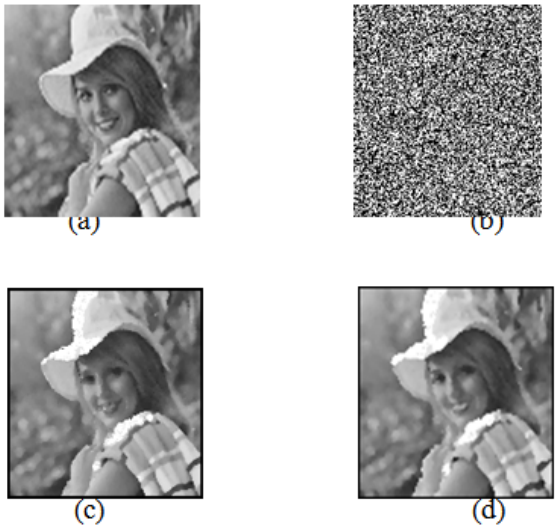


Figure 1. (a) Original Elaine.jpg image (b) corrupted by 80% noise (c) Filtered image of 80% noise (d) Filtered image of 90% noise.

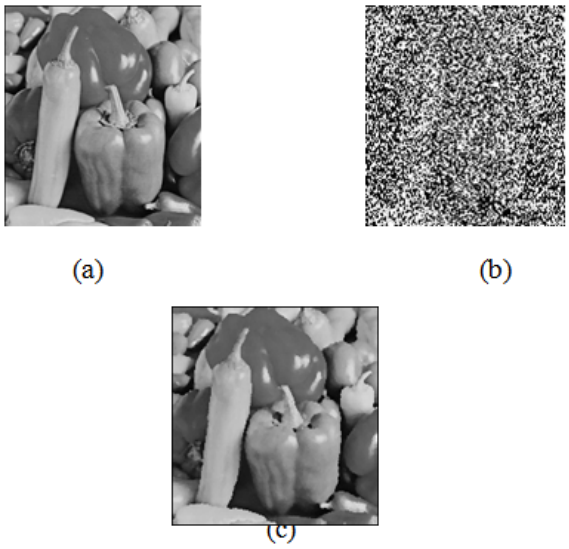


Figure 2. (a) Original pepper.png image (b) Corrupted by 70% image (c) Filtered image

Table 2. Comparison table of **PSNR** of different filters for Elaine.jpg (Gray scale image)

Noise Density	SMF	CWMF	WMF	RWMF	Proposed Algorithm
10	33.72	33.67	34.22	33.28	45.2265
20	29.62	25.81	27.08	32.22	44.3794
30	24.03	20.04	21.66	31.08	43.8565
40	19.03	16.19	17.57	29.14	43.3878
50	15.45	13.12	14.22	25.96	42.2871
60	12.44	10.59	11.64	21.88	41.7497
70	10.09	9.12	9.49	17.56	41.2915
80	8.19	7.64	7.90	14.14	39.7333
90	6.69	6.46	6.58	11.91	39.0835

Table 3. Comparison table of **MSE** of different filters for Elaine.jpg (Gray scale image)

Noise Den.	SMF	CWMF	WMF	RWMF	Proposed Algorithm
10	25.90	21.16	20.340	18.83	7.86
20	46.10	76.56	56.25	40.32	9.56
30	117.50	228.91	179.56	88.54	10.78
40	305.20	561.69	444.36	174.24	12.01
50	677.04	1101.57	895.80	237.16	15.48
60	1330.0	1882.69	1586.4	517.56	17.52
70	2241.0	3028.30	2524.0	1203.39	19.47
80	3464.5	3913.75	3672.3	2127.05	27.87
90	4883.2	5612.42	5031.0	4406.30	32.37

Table 4. Comparison table of **MAE** of different filters for Elaine.jpg (Gray scale image)

Noise Density	SMF	CWMF	WMF	RWMF	Proposed Algorithm
10	2.74	1.72	2.12	1.48	3.1546
20	3.40	3.08	3.17	1.68	4.2274
30	5.06	6.67	5.70	1.94	4.5715
40	9.10	13.21	10.75	2.40	4.9664
50	16.39	24.05	19.87	3.42	6.4831
60	28.92	38.18	33.45	5.73	7.0193
70	46.68	56.78	52.44	11.23	7.6515
80	70.01	78.18	73.9	20.76	10.8357
90	96.98	101.66	99.01	31.45	12.3161

Table 5. Comparision of different filters of **SSIM** (at 70% noise density)

Algorithms	SSIM
WMF (5× 5 Window size)	0.20708
CWMF (5× 5 Window size)	0.09442
RWMF(2 Iteration)	0.3398
PSMF	0.7113
Proposed Algorithm	0.6884

Conclusions

This is nonlinear filtering algorithm, also called as detail preserving median filter.The proposed algorithm is implemented in Mat lab 7.0.The filtered images are compared with different existing filters as shown in above tables. The PSNR is getting better than other filtering algorithms. The future scope is that by increasing the window size further, 99% noise can be filtered with this algorithm within very short processing time.

References

[1] I. Pitas and A. N. Venetsanopoulos , Nonlinear Digital Filters: Principles and Applications.Kluwer Academic Publishers, 1990.

[2] J. Astola and P. Kuosmanen, Fundamentals of Nonlinear Filtering, CRC Press,1997.

[3] R.C.Gonzalez and R.E.Woods, Digital Image Processing, 2nd ed.Addison Wesley, 1992.

[4] Ho-Ming Lin and Alan, "Median filters with Adaptive Length", IEEE transactions of the circuits and systems, vol.35, no.6, June 1998.

[5] H.Hwang and R.A.Haddad, "Adaptive median filters:new algorithms and results ",IEEE Transactions on Image Processing,4,pp.499-502,1995.

[6] Lin Yin and Yang, "Weighted Median filters: A tutorial, ",IEEE transactions of the circuits and systems, vol.43, no.3,March 1996.

[7] Z.Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity", IEEE Transactions on Image Processing, vol. 13, no. 4, Apr. 2004.

[8] R.C.Hardie and.E. Barner, "Rank conditioned rank selection filters for signal restoration,"IEEE Trans.Image Processing, vol.3,pp. 192-206,Mar.1994.

[9] S.J.Ko and Y.H.Lee, "Centre weighted median filters And their applications to image enhancement", IEEE Trans.CircuitsSystem,vol.38,no.9, pp.984- 993, Sep.1991.

[10] Zhou Wang, Alan Conrad Bovik,Hamid Rahim sheik and Erno P.Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity",IEEE Trans.Image Processing , Vol.13,(2004)

Tamil Search Engine for Unicode Standard

C.M.M. Mansoor[*] and H.M. Nasir[#]

[*]*Department of Computer Science, South Eastern University of Sri Lanka*
E-mail: mansoorr@seu.ac.lk
[#]*Department of Mathematics, University of Peradeniya, Sri Lanka*
E-mail: nasirh@pdn.ac.lk

Abstract
The web creates new challenges for information retrieval. Search engine technology has to scale dramatically to keep up with grow of the web. The Internet has been largely dominated by English till recently. The importance of reaching out to non-English speakers around the globe has been felt increasingly and this has lead to the spread of other languages on the Internet. A search engine capable of searching the web documents written in languages other than English is highly needed, especially when more and more sites are coming up with localized content in multiple languages. Tamil is the fastest growing language in the Internet among Indian languages. Thus it became necessary to develop a Tamil search engine which can look into Tamil web pages and retrieve relevant pages for the user. Tamil has more than 50 encoding. This created a proliferation of incompatible encodings on the net. Net users are forced to download fonts for each web site separately. It is very difficult to user to search and find their relative information using different search engine. In this paper, we discuss about developing the Search Engine for Tamil using Unicode encoding scheme.

Index Terms: Crawler, Encoding, Index, Rank, Unicode

Introduction

A wide number of languages are spoken by human beings in the world, and most of the people prefer to have information in their own language. But unfortunately, this information is not reaching to the widest of the masses. This is primarily because of the sweeping dominance of English and western-European languages on the web. However this situation is beginning to change, because most of the new Internet users will not have English as a mother-tongue, and almost everyone wants to use Internet in his or her native language. Hence we found the need for a search engine capable of searching Indian language documents.

Tamil is the fastest growing language in Internet among the India, Sri Lanka and other world. Searching for the required Information in the Tamil websites become increasingly difficult if not impossible

Search Engine is a software package that collects web pages on the internet through a robot program and stores the information on the pages with appropriate indexing to facilitate quick retrieval of desired information. The crawler continuously looks for updated information on the web and stores it in a database. When a user queries the search engine for a particular topic, the search engine looks up the database and lists the pages containing information on that topic. In displaying, some mechanism for ranking the relevant pages is used. [1]

The challenges faced for handling Tamil documents and the Tamil Language is different. There are innumerable Tamil sites available on the net. Moreover, the formats in which they are stored /represented are different. For instance, different fonts are used by different sites, each font having its own encoding.

Another challenge is with respect to the language itself. Tamil being a highly inflectional language, every root word takes on innumerable forms due to the addition of suffixes indicating person, gender, tense, number, cases etc. Thus the number of words to be handled is large. This can be an issue in the design of the database. [3]

Multilingualism on WWW

The arrival of the languages other than English, points out the need of the support in Internet based applications like HTML, HTTP server, Browser, Search Engines, etc. Now HTML has several tags which allow one to specify fonts and language attributes of a particular section of text. A character coding scheme, UNICODE [12], has already been designed to support the interchange, processing, and display of text in many languages of the modern world.

Web content being written in different languages of the world it has become important to have search engines that can search the documents written in different languages. So we need to have a standard for the character encoding that incorporates all the languages of the world. This standard is called Unicode. With this standard coming into vogue, search engines need to search the keywords in the documents written in Unicode.

Literature Review

During the past several years, different approaches have been introduced to search a non English language documents. Because of the different encoding in Tamil, it was no success to establish the search engine to search the all kind of documents. Therefore most the Tamil using people widely using the English based search engine to find their relative documents.

Later many of these enthusiasts and software developers joined together to create a singular standard encoding, namely TSCII. During the Tamilnet'99 conference in Chennai, then Tamilnadu Government also announced two standard encodings, namely TAB and TAM. TAB is a bilingual encoding and TAM is monolingual. TSCII and TAM seem to have considerable usage on the net. Unfortunately these standards do not seem to have many converts among the popular Tamil websites.

Each language has some encoding, English-ASCII. For Tamil, there are more than 50 encoding and the international convention has introduced 'Unicode' coding standard. The idea is that all languages would follow the same encoding. AU-KBC is the first in India to produce a search engine, which searches websites of all Indian language. But it was not fully capable of handling Tamil different encoding

Another search engine project was developed Tamil Nadu Virtual University (TVU) this is similar to the earlier and it has a problem with handling Tamil Unicode encoding system. Also Resource Centre for Indian Language Technology Solutions currently doing the project called Bavaani to handled different encoding for Tamil searching.

Tamil Character Encodings

Each language has some encoding, English-ASCII. For Tamil, there are more than 50 encoding standard. Tamil is a language, where in addition to the basic vowels (uyir) and consonants (mei), the compounded (uyirmei) characters, all have unique glyph forms. Popular Tamil font encoding schemes are TSCII, TAM, TAB, ISCII and Unicode.

Indian Standard Code for Information Interchange (ISCII)

ISCII is a bilingual character encoding (not glyphs-based) scheme. Roman characters and punctuation marks as defined in the standard lower-ASCII take up the first half the character set (first 128 slots). Characters for Indic languages are allocated to the upper slots (128-255). The Indian Standard ISCII-84 was subsequently revised in 1991 (ISCII-91). [10]

Tamil Standard Code for Information Interchange (TSCII)

TSCII is a single byte bilingual 8-bit glyph-based encoding where the lower 7-bit part (first 128 characters) is filled with standard plain ASCII characters. A select number of Tamil glyphs/characters are placed in the upper-ASCII part (slots 128-255). Hence conceptually TSCII is very similar to widely used 8-bit iso-8859-xxx schemes. [14]

Tamil Monolingual Encoding and (TAM) Tamil Bilingual Encoding (TAB)

TAB encodes the Roman script along with Tamil script. The first 128 code points of the TAB encoding scheme are exactly identical to the ASCII character set while the next 128 code points, which encodes the Tamil script, is a subset of the TAM monolingual Tamil encoding scheme. [11]

Unicode Standard

Unicode is a universal font encoding scheme, designed to cover all world languages. It is a 32-bit scheme with over 65500 slots to assign to various languages. Each language (except few like Chinese) is given a 128-slot block [9]

Unicode is intended to solve a large-scale problem in the world: interoperability of systems across not only national boundaries, but among different linguistic communities and with software from different vendors. [9]

Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. The Unicode Standard has been adopted by such industry leaders as Apple, HP, IBM, Just System, Microsoft, Oracle, SAP, Sun, Sybase, and Unisys.

Tamil Unicode Code Chart

The Unicode address space for Tamil language Address for specific character or symbol can be written as the combination of U+ upper most row heading followed by the left most column digit. For example, Address for #### is U+0B85.this can be directly found on table 02; while U+ represents that the address is in the Unicode standard.

Architecture of Search of Search Engine

A search engine is software that searches for documents dealing with a specific topic in the Internet. The basic architecture of a search engine is shown blow. It consists of two parts, a back-end database and a front-end graphical user interface (GUI), to facilitate the user to type the search term.

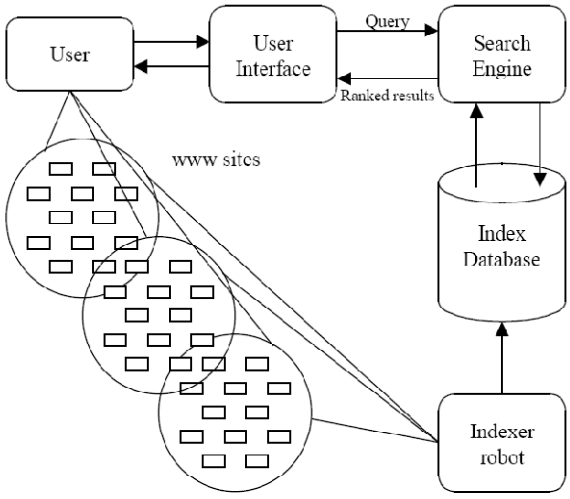


Figure 01: Architecture of Search Engine.

On the server side, the process involves creation of a database and its periodic updating done by software called spider. The spider also called as crawler, robot or indexer, crawls the web pages periodically and indexes these crawled web pages in the database. The hyperlinked nature of the Internet makes it possible for the spider to traverse the web. The front-end of the search engine is the client side having a graphical user interface, which prompts the user to type in the search query. The interface between the client and server side consists of matching the user query with the entries in the database and retrieving the matched web pages to the user's machine.

The database consists of a number of tables that are arranged so as to facilitate faster retrieval of the data. This database is housed in a database server, which is connected to the search engine.. The Tamil search engine uses a single database server, because of the small number of Tamil websites.

Methodology

Design Issues

In designing information retrieval engines it is very important to consider the human factors - how people search, how they make decisions. Normally, users specify some keywords of a subject for searching the documents. Our search engine is primarily designed to search Tamil language documents available on the Web. Its design is language independent. At the top level, our search engine can be divided into three parts: gatherer, indexer, and search processor. Gatherer collects the documents from the Web and passes them to the indexer.

Design of Tamil Search Engine

The Tamil Search Engine aims at looking up Tamil sites for information sought by a user. It searches for Tamil words in Tamil web sites available in Unicode font encoding schemes. The system gathers information from the Internet by the process of crawling. The information is gathered and stored in the database. An interface is provided to the user to enter the query. The query is analyzed with the information from the database, and the information that matches the query is returned to the user. Like any other search engine, it consists of a crawler, an indexing mechanism and a database to store the information.

Search Server

The operation of the searching is simple. In this case also, first the crawled data are stored in the database as mentioned earlier. . The search processor is implemented as a server to make the searching fast, since for every search we don't need to do all the initialization process. The search processor searches for the keywords and gives out the results according to the following algorithm:

For every keyword that is submitted to the Search
Server lookup the word in database
if found
end the search and print 'no results'
if not found
lookup the keyword
and take out the codeword for it.
lookup the codeword in inverted_index(I) and take
out the entry for the keyword.

Results and Discussion

Web crawlers are an essential component to search engines; running a web crawler is a challenging task. Crawling is the most fragile application since it involves interacting with hundreds of thousands of web servers and various name servers, which are all beyond the control of the system. Following is the process by which

The most crucial evaluation of focused crawling is to measure the harvest ratio, which is rate at which relevant pages are acquired and irrelevant pages are effectively filtered off from the crawl. This harvest ratio must be high, otherwise the focused crawler would spend a lot of time merely eliminating irrelevant pages, and it may be better to use an ordinary crawler instead

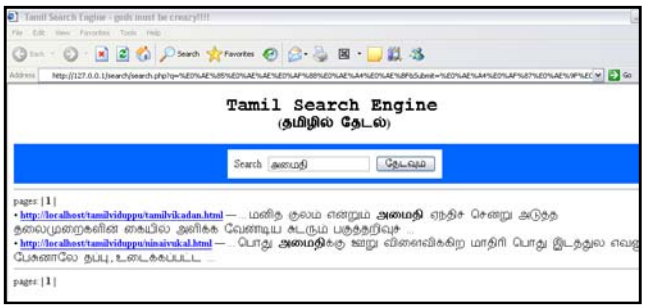


Figure 02. Result of web searched side

Conclusion

We focused in an attempt to enable users finding information they need in Tamil. The important points on search engines area Iare to crawl the Internet sites, index the resources, give every resource a rank, and check information retrieval relevancy.

This study attempted the creation of Tamil Search Engine Using Unicode System. Thus it became necessary to develop a Tamil search engine which can look into Tamil web pages and retrieve relevant pages for the user. So, this research the explained in detail the technology behind the Tamil search engine and the features of the search engine.

This research attained its main objective which is developing Tamil search engine using Unicode standard. with this application, The front-end of the search engine is the client side having a graphical user interface, which prompts the user to type in the search query. The interface between the client and server side consists of matching the user query with the entries in the database and retrieving the matched WebPages to the user's machine.

References

[1] Dell Zang, Yisheng Dong, An Efficient Algorithm to Rank Web Resources. [Online]. Available: , http://www9.org/w9cdrom/ 251/251.html

[2] c Franklin Curt, "How Internet Search Engines Work", Accessed on [July 2004], . [Online]. Available: http://computer.howstuffworks.com/search-engine1.htm

[3] C N, "New Tamil search engine", Accessed on [March 2005], Published on Nov 18th, 2004. . [Online]. Available: http://www.chennaionline.com/science/Technology/11tamilsearch.asp

[4] ZHANG Wen-hui et al, "A Multilingual (Chinese, English) Indexing, Retrieval, Searching Search

Engine", Accessed on [March 2005], [Online]. Available: http://www.isoc.org/inet99/proceedings/posters/210/,

[5] Belew K. Richard, 2000, Finding Out About – "A Cognitive Perspective on Search Engine Technology and the WWW". Cambridge University Press, Cambridge, UK

[6] V.S. Rajam, A reference grammer of Classical Tamil Poetry, American Philosophical Society, Philadelphia, , p.1, (1992); Encyclopaedia Britannica, vol. 21, p.647-648, 1972 ed.

[7] R.M. Suresh, S. Arumugam and K.P. Aravanan, "Recognition of handwritten Tamil characters using fuzzy classificatory approach", Proc. The Tamil Internet 2000 Conference, Singapore, July 2000

[8] Acharya, 2005, "Multilingual Computing for Literacy and Education", SDL, IIT Madrass, India. . [Online]. Available: http://acharya.iitm.ac.in/acharya.html

[9] Unicode Consortium - Universal Code Standard. 1991. [Online]. Available: http://www.unicode.org.

[10] Web Reference: Center for Development of Advanced Computing, CDAC, Pune, India. .[Online].Available: http://www.cdac.org.in/html/gist/articles.html

[11] P. Chellappan, "Tab to Unicode Conversion", [Online]. Available: http://unicode.org/notes/tn17/tab_to_unicode.pdf

[12] Samaranayake, V. K., Nandasara, S. T., Dissanayake, J. B., Weerasinghe, A.R.,Wijayawardhana, 2003,H., "An Introduction to UNICODE for Sinhala Characters", University of Colombo School of Computing

[13] S. Brin and L. Page, "The Anatomy of a Large-Scale Hyper-textual Web Search Engine", Computer Science Department, Stanford University, 1998

[14] V.S. Rajam, A reference grammar of Classical Tamil Poetry, American Philosophical Society, Philadelphia, , p.1, (1992); Encyclopedia Britannica, vol. 21, p.647-648, 1972 ed.

[15] Baeza-Yates Ricardo, Ribeiro-Neto Berthier, 1999, "Modern Information Retrieval". Addison-Wesley, USA.

Inter Color Local Ternary Patterns for Image Indexing and Retrieval

[1]P.V. N. Reddy and [2]K. Satya Prasad

[1]Research Scholar, Dept. of Electronics and Communication Engineering, JNTU, Kakinada, India- 533003
[2]Rector &Professor, Dept. of Electronics and Communication Engineering, JNTU, Kakinada, India-533003
E-mail: pvnreddy_alfa@rediffmail.com , Prasad_kodati@yahoo.co.in

Abstract
Content Based Image Retrieval (CBIR) system using Inter Color Local Ternary Patterns (ICLTP) based features with high retrieval rate and less computational complexity is proposed in this paper. The property of LTP is, it extracts the information based on distribution of edges in an image. This property made it a powerful tool for feature extraction of images in the data base. First the image is separated into red(R), green(G), and blue(B) color spaces, and these are used for inter color local ternary patterns (ICLTP), which are evaluated by taking into consideration of local difference between the center pixel and its neighbors by changing center pixels of one color plane with other color planes. Improved results in terms of computational complexity and retrieval efficiency are observed over recent work based on Local Binary Pattern (LBP) based CBIR system. The $d_1$distance is used as similarity measure in the proposed CBIR system.

Keywords: CBIR, Feature Extraction, Local Binary Patterns, Inter color Local Ternary Patterns.

Introduction

With the rapid growth of digital image and video, Content Based Image Retrieval (CBIR) has become important research area to help people to search and retrieve useful information. High retrieval efficiency and less computational complexity are the desired characteristics of CBIR system. CBIR finds applications in advertising, medicine, crime detection, entertainment and digital libraries. Computational Complexity and retrieval efficiency are the key objectives in the design of CBIR system [1]. However, designing of CBIR system with these objectives becomes difficult as the size of image data base increases. CBIR based on color, texture, shape and edge information are available in the literature [2, 3, 4, 5, 6].This paper describes an image retrieval technique based on ICLTP. Texture is an important feature of natural images. Features of an image should have a strong relationship with semantic meaning of the image. CBIR system retrieves the relevant images from the image data base for the given query image, by comparing the feature of the query image and images in the database. Relevant images are retrieved according to minimum distance or maximum similarity [7] measure calculated between features of query image and every image in image database.CBIR systems can be based on many features viz., texture, color, shape and edge information. Texture contains important information about the structural arrangement of surfaces and their relationship to the surroundings. Varieties of techniques are developed for texture analysis [8, 9]. Most of the texture features are obtained from the application of a local operator, statistical analysis or measurement in transform domain.

Swain et al. proposed the concept of color histogram in 1991 and also introduced the histogram intersection distance metric to measure the distance between the histograms of images [10]. Stricker et al. (1995) used the first three central moments called *mean, standard deviation* and *skewness* of each color for image retrieval [11]. Pass et al. (1997) split the each histogram bin into two parts called a color coherence vector (CCV) [12]. CCV partitions the each bin into two types, i.e., coherent, if it belongs to a large uniformly colored region or in coherent, if it does not. Huang et al. (1997) used a new color feature called color correlogram [13]. Color correlogram characterizes not only the color distributions of pixels, but also spatial correlation of pair of colors. Lu et al. (2005) proposed color feature based on vector quantized (VQ) index histograms in the DCT domain. They computed 12 histograms, four for each color component from 12 DCT-VQ index sequences [14].

Texture is another salient and indispensable feature for CBIR. Smith et al. used the mean and variance of the wavelet coefficients as texture features for CBIR [15]. Moghaddam et al. proposed the Gabor wavelet correlogram (GWC) for CBIR [16, 17]. Ahmadian et al. used the wavelet transform for texture classification [18]. Moghaddam et al. introduced new algorithm called wavelet correlogram (WC) [19]. Saadatmand et al. [20, 21] improved the performance of WC algorithm by optimizing the quantization thresholds using genetic algorithm (GA). Birgale et al. [22] and Subrahmanyam et al. [23] combined the color (color histogram) and texture (wavelet transform) features for CBIR. Subrahmanyam et al. proposed correlogram algorithm for image retrieval using wavelets and rotated wavelets (WC+RWC) [24].

Ojala et al. proposed the local binary pattern (LBP) features for texture description [25] and these LBPs are converted to rotational invariant for texture classification [26]. Pietikainen et al. proposed the rotational invariant texture classification using feature distributions [27]. Ahonen et al. [28] and Zhao et al [29] used the LBP operator facial expression analysis and recognition. Heikkila et al. proposed the background modeling and detection by using LBP [30]. Huang et al. proposed the extended LBP for shape localization

[31]. Heikkila et al. used the LBP for interest region description [32]. Li et al. used the combination of Gabor filter and LBP for texture segmentation [33]. Face recognition under different lighting conditions by the use of local ternary patterns is discussed in [34] where emphasis lays on the issue of robustness of the local patterns.

To improve the retrieval performance in terms of retrieval accuracy, in this paper, we constructed the inter space local ternary patterns (LTP) histograms between the red (R), green (G), and blue (B) spaces. These histograms are used as the feature vectors for image retrieval. The experimentation has been carried out on Corel and MIT VisTex databases for proving the worth of our algorithm. The results after being investigated shows a significant improvement in terms of their evaluation measures as compared to LBP histogram on R, G, B spaces separately.

The organization of the paper as follows: In section I, a brief review of image retrieval and related work is given. Section II, presents a concise review of local patterns, and presents the proposed system framework is given in section III. Experimental results and discussions are given in section IV. Based on above work conclusions are derived in section V.

Local Patterns

Local Binary Patterns (LBP)

Ojala et al. introduced the LBP [25] for texture description as shown in Fig. 1. For a given center pixel in the image, a LBP value is computed by comparing it with those of its neighborhoods:

$$LBP_{P,R} = \sum_{i=0}^{P-1} 2^i \times f(g_i - g_c) \tag{1}$$

$$f(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{2}$$

where g_c is the gray value of the center pixel, g_i is the gray value of its neighbors, P is the number of neighbors and R is the radius of the neighborhood. Fig. 2 shows the examples of circular neighbor sets for different configurations of (P,R).

The uniform LBP pattern refers to the uniform appearance pattern which has limited discontinuities in the circular binary presentation. In this paper, the pattern which has less than or equal to two discontinuities in the circular binary presentation is considered as the uniform pattern and remaining patterns considered as non-uniform patterns.

Fig. 3 shows all uniform patters for *P=8*. The distinct values for given query image is $P(P-1)+3$ by using uniform patterns.

After identifying the LBP pattern of each pixel *(j, k)*, the whole image is represented by building a histogram:

$$H_S(l) = \sum_{j=1}^{N_1} \sum_{k=1}^{N_2} f(BLP_{P,R}^{u2}(j,k), l);\ l \in [0, P(P-1)+3] \tag{3}$$

$$f(x,y) = \begin{cases} 1 & x = y \\ 0 & othrwise \end{cases} \tag{4}$$

where the size of input image is $N_1 \times N_2$.

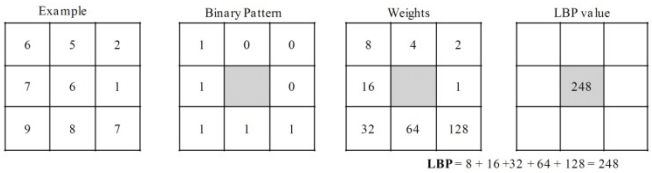


Fig. 1: LBP calculation for 3×3 pattern

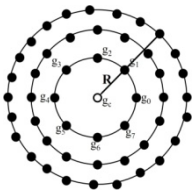


Fig. 2: Circular neighborhood sets for different *(P,R)*

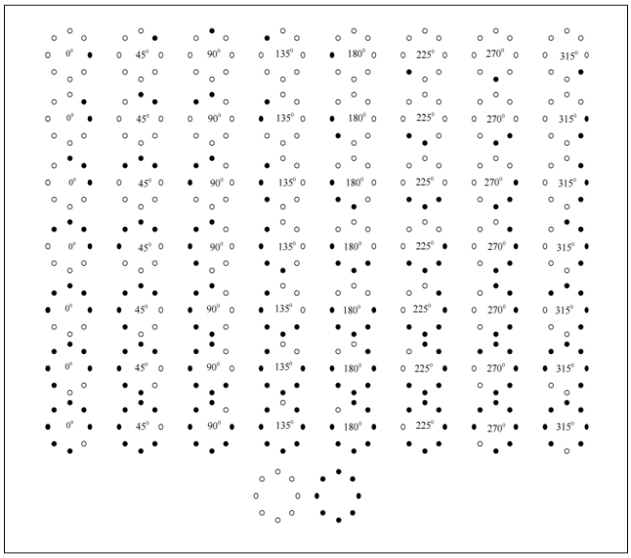


Fig. 3: Uniform patters when P=8. The black and white dots represent the bit values of 1 and 0 in the LBP operator.

Local Ternary Patterns (LTP)

Xiaoyang Tan and Bill Triggs proposed the 3-valued codes, LTP, in which gray-levels in a zone of width ±*t* around g_c are quantized to zero, ones above this are quantized to 1 and ones below it to -1, i.e., the indicator *f(x)* is replaced with a 3-valued function

$$f(x, g_c, t) = \begin{cases} 1 & x \geq g_c + t \\ 0 & |x - g_c| < t \\ -1 & x \leq g_c - t \end{cases} \tag{5}$$

and the binary LBP code is replaced by a ternary LTP code. Here *t* is a user-specified threshold—so LTP codes are more resistant to noise, but no longer strictly invariant to gray-level transformations. The LTP encoding procedure is illustrated in Fig. 4. Here the threshold was set to 5, so the tolerance interval is [40, 50].

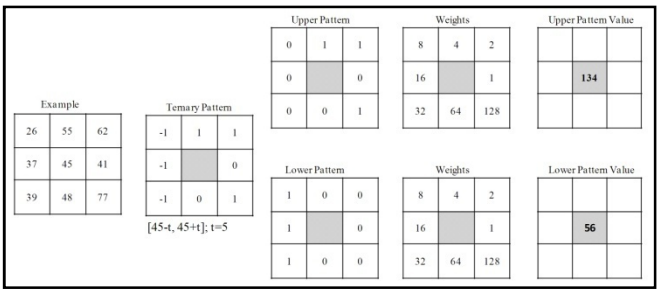


Fig. 4: LTP calculation for 3×3 pattern.

Proposed System Framework (ICLTP)

In this paper, we proposed the new technique by constructing inter LTP histograms between R, G, B spaces. The proposed algorithm combines the color feature (RGB histograms) and texture feature (LBPs) for image retrieval. Fig. 5 shows the flowchart of the proposed image retrieval system and algorithm for the same is given below:

Algorithm:

Input: Image;Output: Retrieval results.

1. Load the input image.
2. Separate the RGB color spaces.
3. Calculate three LTPs on R space by replacing the R pattern center pixel with G and B patter center pixels.
4. Construct the LTPs on G and B spaces also.
5. Construct the inter space LTP histogram between RGB color spaces.
6. Form the feature vector by using LTP histograms.
7. Calculate the best matches using Eq. (6).
8. Retrieve the number of top matches.

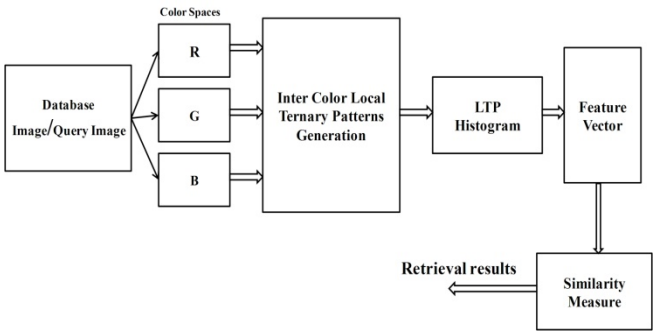


Fig. 5: Proposed image retrieval system framework

Similarity Measurement

In the presented work d_1 similarity distance metric is used as shown below:

$$D(Q,I_1)=\sum_{i=1}^{Lg}\left|\frac{f_{I,i}-f_{Q,i}}{1+f_{I,i}+f_{Q,i}}\right| \quad (6)$$

where Q is query image, Lg is feature vector length, I_1 is image in database; $f_{I,i}$ is i^{th} feature of image I in the database, $f_{Q,i}$ is i^{th} feature of query image Q.

A query image may be any one of the data base images. This query is then processed to compute feature vector as given in the algorithm. The d_1 similarity distance is computed as given in equation 6. The distances are then sorted in increasing order and the closest sets of images are then retrieved. The top 'N' retrieved images are used for computing the performance of the proposed method. The retrieval efficiency is measured by counting the number of matches.

Experimental results

The retrieval performance in terms of average retrieval rate and retrieval time of the proposed CBIR system is tested by conducting an experiment on MIT VisTex data base.

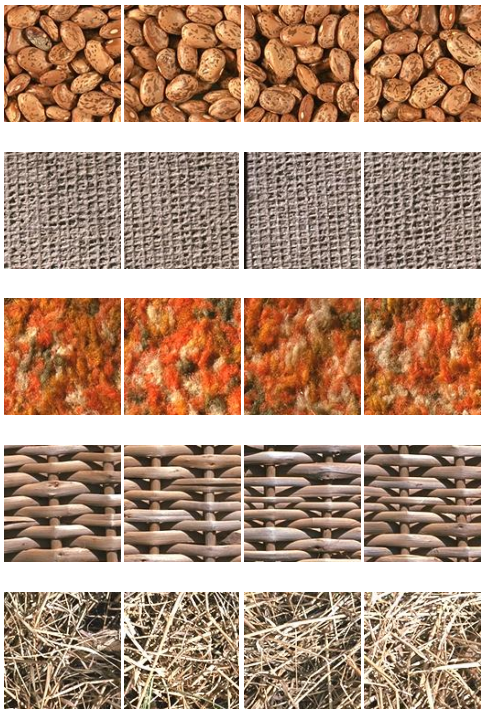


Fig. 6: sample images from MIT VisTex database of five different subjects

MIT VisTex database consists of 400 colorful different textures [35]. The size of each texture is 512×512. Each 512×512 image is divided in to sixteen 128×128 non over lapping sub images, thus creating a data base of 640(40×16) images. In this work, all images in the data base are converted into R, G and B planes. Some sample images in this data base are shown in Fig. 6. All the images in the data base are scaled to a size of 128x128. For creating the feature data base on each image local ternary patterns (LTP) are computed. The ICLTP histograms are computed between RGB color spaces. This ICLTP histogram is used as a feature vector.

Some of the retrieval results when all the 16 images (N=16) in one subject of the image data base are retrieved are shown in Fig. 7.

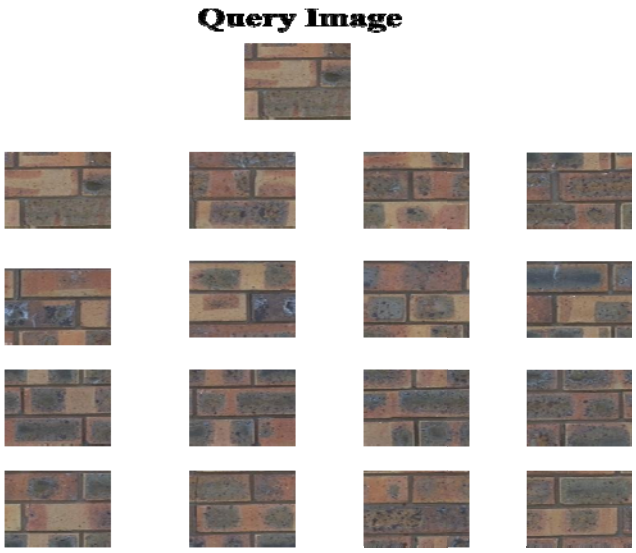


Fig. 7 Retrieved top 16 similar images from MIT VisTex database

Average retrieval rate

The average retrieval rate for the query image is measured by counting the number of images from the same category which are found in the top 'N' matches. Comparative retrieval performance of the proposed CBIR system on the MIT VisTex data base using ICLTP histogram features is shown in Table1.

Table 1. Percentage average Retrieval efficiency on MIT

Method	Number of top matches considered						
	1	3	5	7	9	11	13
LBP	100	99.11	97.81	95.78	93.05	90.34	87.70
ICLTP	100	99.68	99.06	98.08	96.85	95.41	93.67

From Table1, it is observed that the proposed CBIR system with d_1 similarity distance is providing improved retrieval performance over LBP based CBIR system. The superiority of the proposed method is also observed in all the cases. i.e., when 'N' is considered as 1, 3, 5, 7, 9, 11, 15(N is the no. of top retrieved images). When all the images in a class are retrieved (N=16) proposed method with d_1 distance is able to produced an improved average retrieval rate as shown in Table 1.

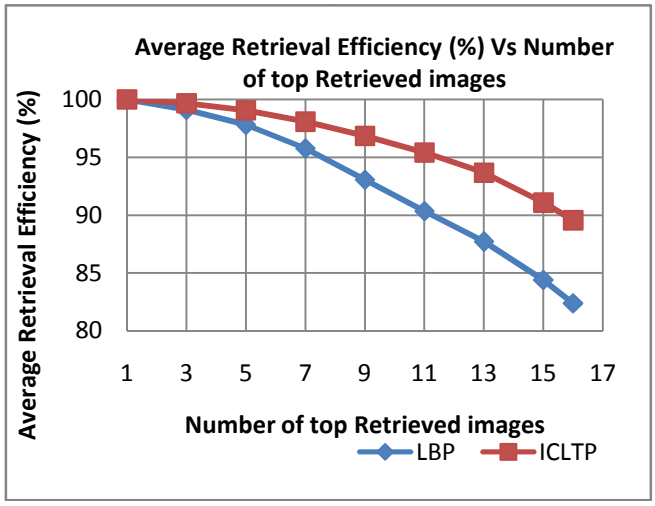


Fig. 8 Percentage average Retrieval efficiency.

Comparative performance in terms average retrieval rate is shown in Fig. 8 indicates the superiority of the proposed ICLTP based CBIR system with d_1 distance when compared to LBP method.

The Average Recall Rate on MIT VisTex database is given in Table 2 and a comparative Recall graph is also presented in Fig. 9.

Table2. Average Recall Rate on MIT VisTex.

Method	Number of top matches considered						
	16	32	48	64	80	96	112
LBP	82.37	90.24	93.06	94.50	95.42	96.16	96.69
ICLTP	89.56	95.01	96.67	97.57	98.22	98.59	98.85

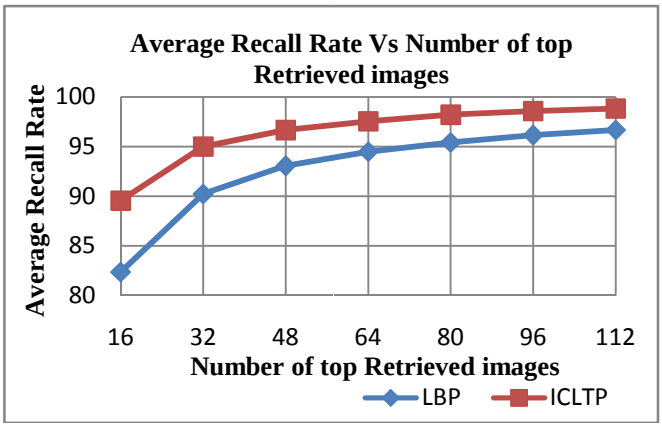


Fig. 9 Average Recall Rate Vs Number of top Retrieved images

Retrieval Time

Proposed CBIR system with ICLTP features is tested on MIT VisTex data base. The average retrieval time for the proposed method is 0.573 seconds, but the average retrieval time in LBP based CBIR system is 0.962 seconds. Hence the proposed method is superior in terms of retrieval time over LBP based

system. Experiments are conducted using MATLAB version 7.8.0 with Pentium IV, 3.00 GHZ.

Conclusions

The performance of the CBIR system dependent on the feature vector that represents the image in the data base. The property of the LTP i.e., it extracts the information based on distribution of edges in an image, is explored in this work. The ICLTP histogram features are used as feature vector representing the image. Superiority of this work is observed in terms of retrieval rate and computational time over LBP based CBIR system.

References

[1] Arnold. W. M. Smeulders, M. Worring, S. Satini,A. Gupta, R. Jain. Content – Based Image Intelligence, Vol. 22, No. 12, pp 1349-1380 , 2000.

[2] Belongie S., Carson C., et al., Color - and Texture Based Image Segmentation using EM and its Application to Content-Based Image Retrieval. Proceedings of 8th International Conference on Computer Vision, 1998.

[3] Manesh Kokare, B .N.Chatterji and P.K.Biswas. A Survey on current content based Image Retrieval Methods, IETE Journal of Research, Vol. 48, No. 3 & 4, pp 261-271, 2002.

[4] Gupta. A.Visual Information Retrieval Technology: A Virage Perspective, Virage Image Engine. APISpecification, 1997.

[5] Smith. J and Chang S.F. Visual SEEK: a fully automated content-based image query system. Proceedings of ACM Multimedia 96, pp 87-98, 1996.

[6] N. Monserrat, E. de Ves, P. Zuccarello. Proceedings of GVIP 05 Conference, December 2005.

[7] S. Satini, R. Jain. Similarity Measures. IEEE Transactions on pattern analysis and machine Intelligence, Vol.21, No.9, pp 871-883, September 1999.

[8] Hiremath.P.S, Shivashankar . S. Wavelet based features for texture classification, GVIP Journal, Vol.6, Issue 3, pp 55-58, December 2006.

[9] B. S. Manjunath and W.Y. Ma. Texture Feature for Browsing and Retrieval of Image Data. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, No. 8, 1996.

[10] M. J. Swain and D. H. Ballar, Indexing via color histograms, Proc. 3rd Int. Conf. Computer Vision, Rochester Univ., NY, (1991) 11–32.

[11] M. Stricker and M. Oreng, Similarity of color images, Proc. SPIE, Storage and Retrieval for Image and Video Databases, (1995) 381–392.

[12] G. Pass, R. Zabih, and J. Miller, Comparing images using color coherence vectors, Proc. 4th ACM Multimedia Conf., Boston, Massachusetts, US, (1997) 65–73.

[13] J. Huang, S. R. Kumar, and M. Mitra, Combining supervised learning with color correlograms for content-based image retrieval, Proc. 5th ACM Multimedia Conf., (1997) 325–334.

[14] Z. M. Lu and H. Burkhardt, Colour image retrieval based on DCT domain vector quantization index histograms, J. Electron. Lett., 41 (17) (2005) 29–30.

[15] J. R. Smith and S. F. Chang, Automated binary texture feature sets for image retrieval, Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Columbia Univ., New York, (1996) 2239–2242.

[16] H. A. Moghaddam, T. T. Khajoie, A. H Rouhi and M. Saadatmand T., Wavelet Correlogram: A new approach for image indexing and retrieval, Elsevier J. Pattern Recognition, 38 (2005) 2506-2518.

[17] H. A. Moghaddam and M. Saadatmand T., Gabor wavelet Correlogram Algorithm for Image Indexing and Retrieval, 18th Int. Conf. Pattern Recognition, K.N. Toosi Univ. of Technol., Tehran, Iran, (2006) 925-928.

[18] A. Ahmadian, A. Mostafa, An Efficient Texture Classification Algorithm using Gabor wavelet, 25th Annual international conf. of the IEEE EMBS, Cancun, Mexico, (2003) 930-933.

[19] H. A. Moghaddam, T. T. Khajoie and A. H. Rouhi, A New Algorithm for Image Indexing and Retrieval Using Wavelet Correlogram, Int. Conf. Image Processing, K.N. Toosi Univ. of Technol., Tehran, Iran, 2 (2003) 497-500.

[20] M. Saadatmand T. and H. A. Moghaddam, Enhanced Wavelet Correlogram Methods for Image Indexing and Retrieval, IEEE Int. Conf. Image Processing, K.N. Toosi Univ. of Technol., Tehran, Iran, (2005) 541-544.

[21] M. Saadatmand T. and H. A. Moghaddam, A Novel Evolutionary Approach for Optimizing Content Based Image Retrieval, IEEE Trans. Systems, Man, and Cybernetics, 37 (1) (2007) 139-153.

[22] L. Birgale, M. Kokare, D. Doye, Color and Texture Features for Content Based Image Retrieval, International Conf. Computer Grafics, Image and Visualisation, Washington, DC, USA, (2006) 146 – 149.

[23] M. Subrahmanyam, A. B. Gonde and R. P. Maheshwari, Color and Texture Features for Image Indexing and Retrieval, IEEE Int. Advance Computing Conf., Patial, India, (2009) 1411-1416.

[24] Subrahmanyam Murala, R. P. Maheshwari, R. Balasubramanian, A Correlogram Algorithm for Image Indexing and Retrieval Using Wavelet and Rotated Wavelet Filters, Int. J. Signal and Imaging Systems Engineering.

[25] T. Ojala, M. Pietikainen, D. Harwood, A comparative sudy of texture measures with classification based on feature distributions, Elsevier J. Pattern Recognition, 29 (1): 51-59, 1996.

[26] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell., 24 (7): 971-987, 2002.

[27] M. Pietikainen, T. Ojala, T. Scruggs, K. W. Bowyer, C. Jin, K. Hoffman, J. Marques, M. Jacsik, W. Worek, Overview of the face recognition using feature

distributions, Elsevier J. Pattern Recognition, 33 (1): 43-52, 2000.

[28] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: Applications to face recognition, IEEE Trans. Pattern Anal. Mach. Intell., 28 (12): 2037-2041, 2006.

[29] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell., 29 (6): 915-928, 2007.

[30] M. Heikkil;a, M. Pietikainen, A texture based method for modeling the background and detecting moving objects, IEEE Trans. Pattern Anal. Mach. Intell., 28 (4): 657-662, 2006.

[31] X. Huang, S. Z. Li, Y. Wang, Shape localization based on statistical method using extended local binary patterns, Proc. Inter. Conf. Image and Graphics, 184-187, 2004.

[32] M. Heikkila, M. Pietikainen, C. Schmid, Description of interest regions with local binary patterns, Elsevie J. Pattern recognition, 42: 425-436, 2009.

[33] M. Li, R. C. Staunton, Optimum Gabor filter design and local binary patterns for texture segmentation, Elsevie J. Pattern recognition, 29: 664-672, 2008.

[34] X. Tan and B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, IEEE Tans. Image Proc., 19(6): 1635-1650, 2010.

[35] MIT Vision and Modeling Group, Vision Texture. [Online]. Available: http://vismod.www.media.mit.edu.

Clinical Analysis of MR Brain Images using 2-D Rigid Registration Method

[1]Jainy Sachdeva, [2]Vinod Kumar, [3]Indra Gupta, [4]Niranjan Khandelwal and [5]Chirag Kamal Ahuja

[1]Research Scholar, Department of Electrical Engineering, Indian Institute of Technology Roorkee, Roorkee-247667, India
E-mail: jainysachdeva@gmail.com
[2&3]Department of Electrical Engineering, Indian Institute of Technology Roorkee, Roorkee-247667, India
E-mail: vinodfee@iitr.ernet.in, indrafee@iitr.ernet.in
[4&5]Department of Radiodiagnosis, Post Graduate Institute of Medical Research & Education (PGIMER), Chandigarh, India
E-mail: khandelwaln@hotmail.com, chiragkahuja@rediffmail.com

Abstract
In this paper, clinical analysis of subjects diagnosed with malignant tumor/cyst is performed using a 2-D rigid multimodal registration technique- a principal axes method. Highly descriptive rigid control points are detected in reference image using area-based technique. The corresponding control points are initialized in target image. Inverse affine transformation is applied to get the registered image. After removing high geometric deformations such as scaling, translation and change in patient posture (rotation of head) by the area based technique, clinical information obtained after registration (CIAR) is ascertained. In CIAR, radiological and pathological information from the registered datasets is discussed. The method is tested on real multi-modal MR - T1, T2, post contrast T1 unregistered image datasets of 10 patients obtained from PGIMER (Post Graduate Institute of Medical Education & Research, Chandigarh, India). The robustness achieved by the proposed method (proved by the cited examples) makes it a useful tool in routine clinical practice.

Index Terms: Clinical information obtained after registration (CIAR), Clinical analysis, multimodal image registration, control points, PGIMER

Introduction

In medical imaging, MR - T1, T2 and post contrast T1 sequences provide anatomical information of various tissues, tumors and cysts present within the brain. The three signal intensities patterns by which tumors or cysts are differentiated from normal tissues are: Isointense signal - same signal or brightness as that of a standard comparative medium; Hypointense signal -reduced signal or less bright as compared to the comparative medium; Hyperintense signal- increased signal or brighter than the comparative medium. MR imaging predicts the solid (T1 hypo to isointense and T2 hypo to isointense) and necrotic (T1 hypointense and T2 hyperintense) components of a tumor. The degree of enhancement on post contrast T1 (gadolinium) sequence (which is different for different tumors) and the extent of the brain oedema it induces (seen as hyperintense signal on T2 weighted sequences) can be determined. These images are taken at different times or at different viewpoints. Clinical MR images-T1, T2, and post contrast T1-weighted sequences consist of affine deformations. Therefore, rigid registration of MR-T1, T2, and post contrast T1 images has to be done before deducing information of clinical relevance. Rigid transformations removes geometric distortions like rotation (arising due to change in patient posture), scaling and translation differences between the reference and the target images. The rigid transformation transforms all pixels in the image with a single geometric transformation.

Rigid registration serves the following purposes:

1. For precise diagnosis of the pathology, preoperative evaluation, neurosurgical planning and evaluation of follow up for the treatment response.
2. For assessment of an intracranial tumor by a neuroradiologist who has to localise a tumor, define its extent and relationship with the adjacent structures including the change it is causing to the intracranial environment.
3. For generating complementary and functional information from multiple modalities.

Image registration techniques [1-3] can be broadly classified into two classes: (i) Area (pixel intensity) based and (ii) Points (feature) based. The area based techniques make use of the intensity of the images as the resemblance criteria for registration. The feature based techniques use the control point correspondences using spatial relations and localized invariant feature descriptors as their matching criteria between the reference and the target images. These techniques maximize the likelihood criteria by initializing a geometric transformation field to get an optimum transformation.

Many researchers have used Mutual Information (MI) [4-6] as the base for multimodal image registration. In MI, similarity between the intensity pairs from corresponding spatial locations in two images is estimated. However, overall registration accuracy is affected as the joint histogram or interpolation methods may introduce artefact patterns. Johnson et al. conducted similar experiments using corresponding landmarks and intensity as matching criteria. Corresponding landmarks are matched between two images and image intensities are matched for the regions away from the

landmark locations. However, this method is applicable to same modality images and the pre-processing is required in some cases so as to equalize the intensities of the images [7]. Wang et al. registration, a non-iterative wavelet-based hierarchical registration is proposed. However, in this approach the control points are taken on the whole image rather than on the regions to be registered. This increases the deformation error while calculating affine transformation. Error in estimation of point transformation gets amplified and transferred to higher resolution levels [8].

Moreover, all these methods are employed on normal brain MR and CT images. Clinical interpretation of images consisting of tumor and cyst is not provided. There may be intensity mismatching when tumor and cyst regions are considered.

In this paper we have proposed a use of 2-D multimodal image registration for clinical analysis of MR brain images. For this application, area based registration technique is used that overcome the limitations of above mentioned methods.

Proposed method

Information of clinical relevance is deduced by an area based technique- a principal axes approach [9]. This approach is applied for automatic multimodal 2-D registration of brain images i.e. the reference image- I_R and the target image- I_T. In the area based module, control points are made invariant to high geometric distortions like scaling, rotation and translation providing approximate correspondence between the point pairs. Area based module is followed by the inverse affine transformation for final alignment of I_R and I_T to get a registered image. From the registered and the input images, analysis is performed to figure out the relevant clinical details which otherwise would have been vague. For clinical analysis, high degree geometrical misalignment between the subject images like deformations in angles, scaling and translation are easily and efficiently correctable by the proposed algorithm.

Area Based Module

In the proposed method active control points are initialized in I_R using standard gradient inertia matrix. Let $\{X_R(s), Y_R(s)\}$ are the active control points in I_R. In the reference image, these points are detected by locating points that have local maxima in edge measures. The points are located by detecting local maxima in Eigen space of gradient inertia matrices in the neighbourhood of 5×5 for each pixel in an image. The gradient inertia matrix for a point *(x, y)* in an image is given by:

$$C(x,y)=\begin{bmatrix}\overline{I_x(x,y)}\times\overline{I_x(x,y)} & \overline{I_x(x,y)}\times\overline{I_y(x,y)}\\ \overline{I_y(x,y)}\times\overline{I_x(x,y)} & \overline{I_y(x,y)}\times\overline{I_y(x,y)}\end{bmatrix} \quad (1)$$

where, $\overline{I_x(x,y)}$ and $\overline{I_y(x,y)}$ are average gradient of image I at point *(x, y)* along *x* and *y*-axis. The smaller Eigen value for gradient inertia matrix on each point is stored and local maximum points in neighbourhood of 5×5 are taken as active control points [10]. After initialization, binary brain masks of I_R and I_T are generated. These masks are produced for background removal in I_R and I_T, which acts as ROI (region of interest). These masks are obtained by smoothing the image and removing Gaussian random noise by use of an averaging filter of size 11×11. Intensity normalization on the filtered image is applied on I_R and I_T using Eq. (2) so that the highest intensity level in the image is 1. Intensity normalized image I_N is the normalized image derived from any image I and is given by:

$$I_N(x,y)=\frac{I(x,y)-\min(I(x,y))}{max(I(x,y))-\min(I(x,y))} \quad (2)$$

Rough binary brain mask is generated by an automatic adaptive threshold on the normalized image. The mask is refined by using morphological operations. The binary discontinuities are removed by morphological closing operation and image filling operations to get a final binary brain mask. A disk shaped structure element of radius 10 pixels is used in the above morphological operations because of its radially symmetric shape. Let *B* be the binary brain mask represented as:

$$B(x,y)=\begin{cases}1, & \text{if } (x,y) \text{ is in binary brain mask}\\ 0, & \text{if } (x,y) \text{ is not in binary brain mask}\end{cases} \quad (3)$$

where, B_R and B_T represents the binary brain masks of I_R and I_T respectively. Translation, rotation and scaling parameters are found by using B_R and B_T. After generation of binary brain masks, translation parameters (t_x, t_y), scaling parameters $(s_{yy}, s_{yx}, s_{xx}, s_{xy})$ and rotation parameters (θ) are estimated [9]. These parameters are discussed below:

Translational parameter estimation

The translational parameters are determined from shift of the centroid of the B_T from centroid location of B_R. The centroid points x_c, y_c are estimated as follows:

$$x_c=\frac{1}{N}\frac{\sum_x x\,B(x,y)}{\sum_x B(x,y)} \qquad y_c=\frac{1}{N}\frac{\sum_y y\,B(x,y)}{\sum_y B(x,y)} \quad (4)$$

where, *N* is number of brain segment pixels in binary brain mask *B*. If $\{x_r, y_r\}$ and $\{x_t, y_t\}$ are the centroid coordinates of B_R and B_T, then translation parameters (t_x, t_y) are defined as:

$$t_x = x_r - x_t \ , \ \ t_y = y_r - y_t \quad . \quad (5)$$

Rotation parameter estimation

The rotation parameter (θ) is calculated as angle between the Eigen vectors of inertia matrices of I_R and I_T. The inertia matrix (I) from brain mask *B* is defined as:

$$I=\begin{bmatrix}I_{xx} & I_{xy}\\ I_{yx} & I_{yy}\end{bmatrix}$$

where

$$I_{xx}=\sum_y (y-y_c)^2\,B(x,y)$$

$$I_{yy} = \sum_{x} (x - x_c)^2 B(x, y) \quad (6)$$

$$I_{xy} = I_{yx} = \sum_{x,y} (x - x_c)(y - y_c) B(x, y)$$

The major Eigen vectors e_r, e_t of inertia matrices for B_R and B_T are directed along the principle axis of B_R and B_T. Orientations ϑ_R, ϑ_T are defined as:

$$\vartheta_R = \tan^{-1}\left(\frac{u_r}{v_r}\right), \quad \vartheta_T = \tan^{-1}\left(\frac{u_t}{v_t}\right) \quad (7)$$

$$\begin{bmatrix} u_r \\ v_r \end{bmatrix} = e_r, \quad \begin{bmatrix} u_t \\ v_t \end{bmatrix} = e_t$$

The rotation parameter θ is the rotational misalignment of B_T with respect to B_R given as:

$$\theta = \vartheta_R - \vartheta_T \quad (8)$$

Scaling parameter estimation

A two-step procedure is adopted for estimation of scaling parameters. In the first step, global scaling i.e. area normalization factor s is calculated as:

$$s = \sqrt{\frac{area(B_R)}{area(B_T)}} \quad (9)$$

In the second step, the relative scaling parameters $(s_{yy}, s_{yx}, s_{xx}, s_{xy})$ are calculated from ratio of the lengths of principle axis of B_R and B_T given as:

$$s_{yy} = s_{yx} = s \times \frac{Lr_{major}}{Lt_{major}}, \quad s_{xx} = s_{xy} = s \times \frac{Lr_{minor}}{Lt_{minor}} \quad (10)$$

where, L_r and L_t are the lengths of the major and the minor principle axis of B_R and B_T respectively.

The estimated parameters are used to find affine transformation T. Let $\{X_T(s), Y_T(s)\}$ are the control points in I_T. These control points are initialized in I_T by applying an affine transformation to $\{X_R(s), Y_R(s)\}$ based on structural misalignment of B_T with respect to B_R. General affine transformation T is defined as:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s_{xx}.\cos\theta & -s_{xy}.\sin\theta & t_x \\ s_{yx}.\sin\theta & s_{yy}.\cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (11)$$

where, (x, y) are input pixel coordinates (x', y') are transformed pixel coordinates.

Inverse Affine transformation generation

Affine transformation generated to get the registered image as given below:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} A_{xx} & A_{xy} \\ A_{yx} & A_{yy} \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \quad (12)$$

In matrix form the above equation can be represented as:

$$X' = AX + T. \quad (13)$$

Here transformation parameters A and T maps points from I_T to I_R. This affine transformation is applied on target image to get registered image.

Dataset and Software implementation

Dataset

The dataset is obtained from Post Graduate Institute of Medical education and Research (PGIMER), Chandigarh, India. It consists of unregistered axial slices of MR (T1, T2 and post contrast T1) with 5mm slice thickness of 10 subjects with malignant tumors/cyst. All the experiments are performed on images of size 256×256 pixels. Dataset is collected over the time period of January 2010 to March 2011. The consent of the patients for using these images for research was taken prior to image recording. The study is approved by the medical ethics committee of the PGIMER, Chandigarh. All the images are obtained using the same MRI equipment (Siemens's Verio, Erlangen Germany, 3Tesla MR scanner). Histological characteristics of brain tumor and cyst MR images in present study are confirmed by the expert radiologists.

Software Implementation

Proposed method is implemented in MATLAB and is tested on various brain tumor MR images of size 256×256. The algorithm takes 53s when run on PC having Intel ™ Core 2 Duo ® 2.0 GHz processor with 3 GB RAM.

Experimental Setup

Evaluation Criteria

The assessment of clinically relevant information of registered datasets is performed via superimposed registration and visual evaluation by the senior radiologists. Superimposed registration means the process of overlaying two images, through registration to a common co-ordinate system, such that the resultant image contains the data from both the images. The superimposed registration is shown by checkerboard representation. Visual evaluation is defined as the ability of an expert radiologist to extract useful anatomical information from registered datasets. Though visual evaluation varies from expert to expert and is subjected to observer's variability. Therefore, registered datasets were assessed by two expert radiologists associated with an Institute of national repute; one of them is Professor & Head and the other is Senior Resident in the Department of Radiodiagnosis, Postgraduate Institute of Medical Education & Research, Chandigarh, India.

Implementation of Proposed Method on Sample Images

In this example, registration of MR (T2) and MR-post contrast T1 images is performed. The reference image of brain MR (T2) image of a subject with astrocytoma tumor is shown in Fig. 1(a). The target image is shown in Fig. 1(b). Active control points on the reference image generated by the use of gradient inertia matrix are shown in Fig. 1(c). The mapped points on the target image generated by binary brain mask alignment are clearly visible on Fig.1 (d). The number of control points on the reference and the target images remain same as same window size is taken. Registered image obtained is shown in Fig. 1(e) and superimposed registration is shown in Fig. 1(f). Other examples of registration of T1, T2, post contrastT1 images are illustrated from Fig.2 and Fig.3 where in each Fig., sequence (a) represents the reference image; sequence (b) represents the target image, sequence (c) represents the registered image and sequence (d) represents the superimposed registration.

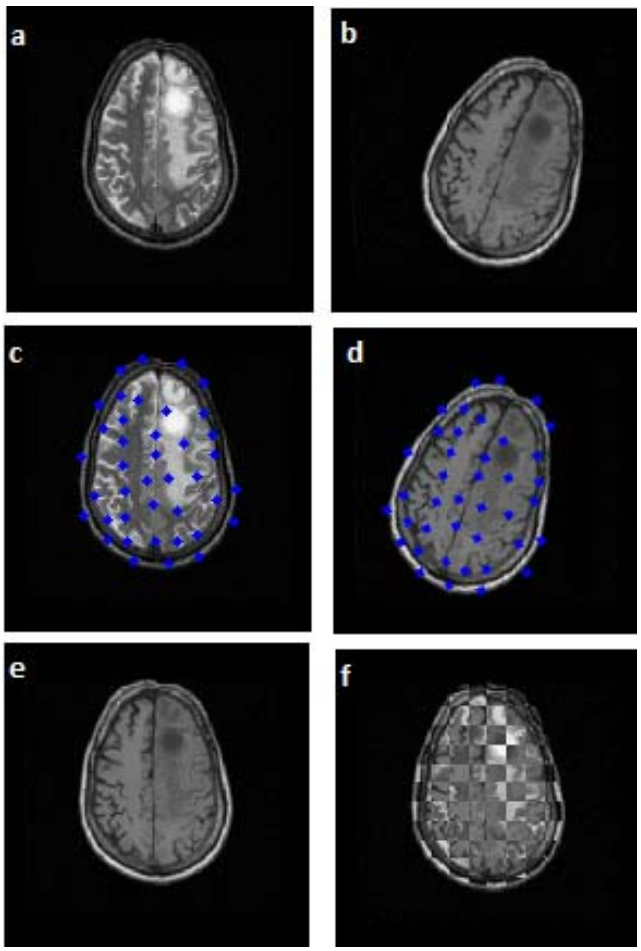


Fig. 1: Registration of axial slice MR (T2) image and MR (post contrast T1) image of a subject with astrocytoma tumor (a) Reference image (MR- T2) (b) Target image (MR-post contrast T1) (c) Active control points on reference image (d) Mapped control points on target image (e) Registered image (f) Superimposed registration

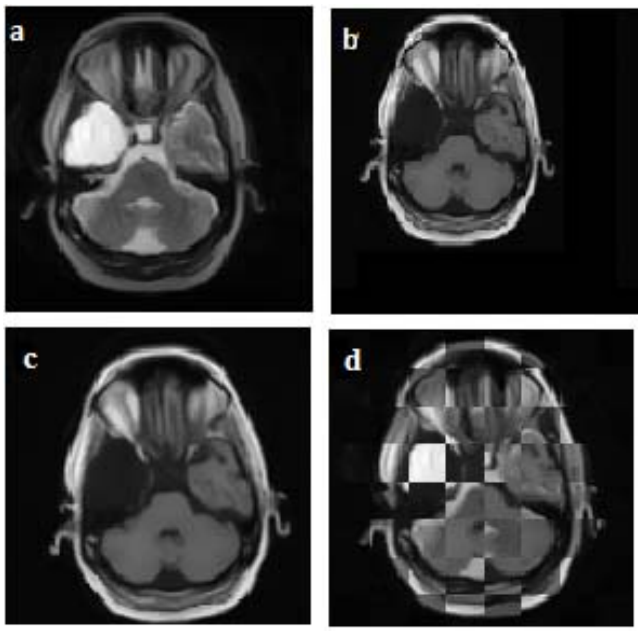


Fig. 2: Registration of axial slice MR (T2) image and MR (T1) image of a subject with arachnoid cyst (a) Reference image (MR- T2) (b) Target image (MR-T1) (c) Registered image (d) Superimposed registration

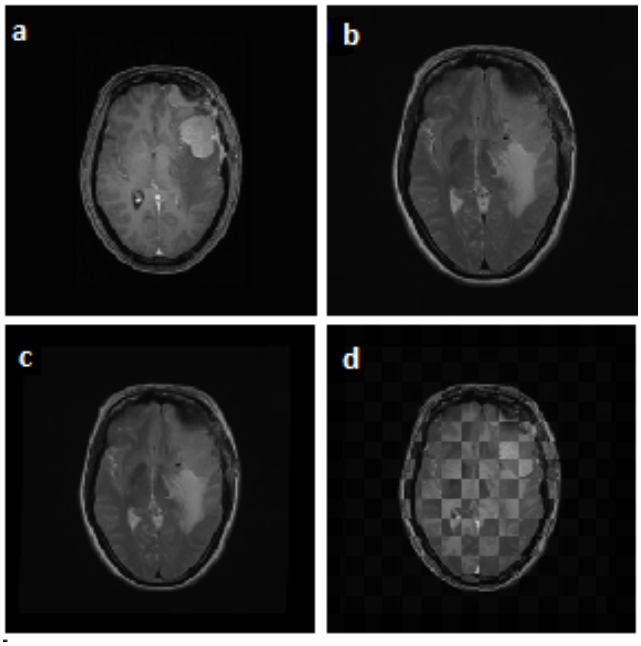


Fig. 3: Registration of axial slice MR (post contrast T1) image and MR (T2) image of a subject with astrocytoma tumor (a) Reference image (MR- post contrast T1) (b) Target image (MR-T2) (c) Registered image (d) Superimposed registration

Results and Discussions

The details of each registered pair are given in Table 1. Column 1 represents the Fig. numbers of the registered datasets. In column 2 the modality details used for registration are given. Tumor/cyst properties depicted by the modality are given in column 3. Column 4 represents the tumor visibility

details on the reference image and the registered image In addition, clinical information deduced after registration is provided in this section. Execution time taken by the registration algorithm to register each dataset is given in Column 5.

It is observed in the superimposed registration that the skull regions finely overlap each other. The continuity of contours in adjacent patches appreciates the quality of registration.

Table I: Results and Discussions for PGIMER dataset

Registered Dataset	Modality	Tumor/Cyst Type	Clinical information after registration (CIAR)	Execution Time
Fig 1(a) and Fig 1(e)	Axial MR- T2 weighted & post contrast T1weighted images	Astrocytoma tumor	Reference Image- Hyperintense T2 lesion (in white) with moderate perilesional oedema (in yellow). Registered Image- Post contrast T1 weighted axial section of the brain with a peripheral rim of solid component (in white). CIAR-Tumor is having a necrotic core. It is a relatively well defined focal lesion in left frontal lobe having isointense signal on T1 (in white).	53 sec
Fig 2(a) and Fig 2(c)	Axial MR- T2 and T1 weighted images	Arachnoid cyst	Reference Image - Cystic lesion showing hyperintense signal on T2 with no perilesional oedema (in white). Registered Image - Cystic lesion showing hypointense signal on T1 (in white). CIAR- A well circumscribed homogenous cyst to be removed by surgery, no perilesional oedema near or around the cyst.	52sec
Fig 3(a) and Fig 3(c)	Axial MR- Post contrast T1 and T2 weighted images	Meningioma Tumor	Reference Image – Post contrast axial T1 MR image shows moderately enhancing extra axial lesion (in white). Registered Image - In the left temporal region, this lesion shows iso to hyperintense signal on T2 (in white). CIAR: The lesion causes compression of the adjacent brain parenchyma leading to white matter oedema seen on T2 (in white, just below the tumor on the reference image).	52 sec

Conclusion

In this paper, clinical information obtained after 2-D rigid multimodal image registration method is discussed. Area based brain mask alignment technique is used which provides the finer mapping of control points between the reference image and the target image. The proposed method provides efficient and fast registration even at high degree of subject movements during the scans and variations in scaling and translation parameters. Experiments performed on real datasets have demonstrated the clinical relevance of the method. The execution time for registering multimodal images is 53 seconds. It can be used as an online system for precise localization, diagnosis, and interpretation by the medical expert and as an initial step in image fusion.

References

[1] B. Zitova, J. Flusser, Image registration methods: a survey, *Image and Vision Computing*, vol.21, pp. 977–1000, 2003.

[2] Leslie G. Brown, A Survey of Image Registration Techniques, *ACM Computing Surveys*, vol. 24, no. 4, pp.325–376, 1992.

[3] J.B.A. Maintz, M.A. Viergever, A survey of medical image registration, *Medical Image Analysis*, vol.2, no.1, pp. 1–36, 1998.

[4] Xuesong Lu, Su Zhang, He Su, Yazhu Chen, Mutual information-based multimodal image registration using a novel joint histogram estimation, *Computerized Medical Imaging and Graphics*,vol.32, pp.202–209, 2008.

[5] Zhiyong Gao, Bin Gu, Jiarui Lin, Monomodal image registration using mutual information based methods, *Image and Vision Computing* ,vol.26, pp.164–173, 2008.

[6] X L.u, S. Zhang, H. Su, Y.Chen, Mutual information-based multimodal image registration using novel joint histogram estimation, *Computer Medical Imaging and Graphic*s, vol.32, no.3, pp.202–209, 2008.

[7] H.J. Johnson, G.E. Christensen, Consistent landmark and intensity-based image registration, *IEEE Transactions of Medical Imaging*, vol.21, no.5, pp. 450-461, 2002.

[8] X. Wang, DD. Feng, Non-iterative hierarchical registration for medical images, *Journal of Signal Processing and Systems*, vol.54, pp.65–77, 2009.

[9] L. K. Arata, A. P. Dhawan, J. P. Broderick, M. F. Gaskil-Shipley, A. V.Levy, and N.D.Volkow, Three-dimensional anatomical model-based segmentation of MR brain images through principal axes registration, *IEEE Transactions of Biomedical Engineering*, vol. 42, no. 11, pp. 1069–1078, 1995.

[10] A. Ardeshir Goshtasby, 2-D and 3-D Image Registration for Medical, Remote Sensing, and Industrial Applications, *Wieley publications*, pp.43-45, 2005.

Protecting Copyright Multimedia Files by Means of Digital Watermarking: A Review

G.S. Kalra[1], *Member IEEE*, Dr. R. Talwar[2], Dr. H.Sadawarti[2], *Member IEEE*

[1]Lovely Professional University, Phagwara, Punjab, India
[2] RIMTIET, Mandi Gobingarh, Punjab, India

Abstract
Digital information is easy to transfer and store but this property of digital information becomes harmful to itself as it can be easily copied and distributed on the internet. Thus, number of efforts are going on to protect the copyright of the owner like Steganography, digital signatures etc. But digital watermarking comes out to be most effective tool among these. It can be applied on text, image, audio and video files in number of ways which are effective for any specific application.

Keywords: Watermarking; image; Audio; video; text watermarking.

Introduction

The growth of high speed computer networks has created new definitions for entertainment, scientific, business and social opportunities. As a result, it causes the growth of digital data. Digital media have several advantages over analog media such as high fidelity copying, easy editing, high quality etc. The digital information can be copied very easily and can be distributed easily which led to the need for effective copyright protection tools. Recent studies [1][2] show that 35% of the software programs installed in 2006 are pirated. It can be prevented if the copyright or the mark of ownership will be added in the original file in such a manner so that in case of any dispute, the actual owner can be identified. It is done by hiding data (information) within digital audio, images and video files. The ways of such data hiding is digital signature, copyright label or digital watermark that completely characterizes the person who applies it and therefore, marks it as being his intellectual property. Digital watermarking is the process that embeds data, called a watermark, into a multimedia object in such a manner that the watermark can be detected or extracted later to make a decision about the copyright of the object. The process of embedding the watermark with the secret key and detection of the watermark is shown in fig.1 (a) and fig.1 (b).The object may be an image, audio, video or text only. A simple example of a digital watermark would be a visible "seal" placed over an image to identify the copyright. However the watermark might contain additional information including the identity of the purchaser of a particular copy of the material. In addition to copyright protection, watermarking is also used in data integrity and data confidentiality. Unlike data integrity and confidentiality applications, watermarks for copyright protection applications need to be robust and invisible.

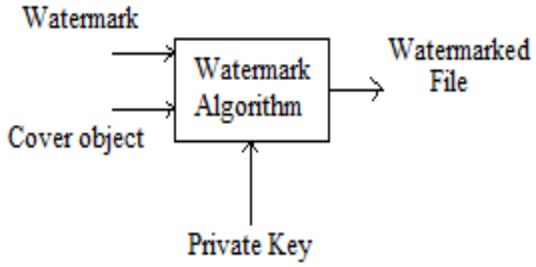


Fig.1(a): Watermark embedding process

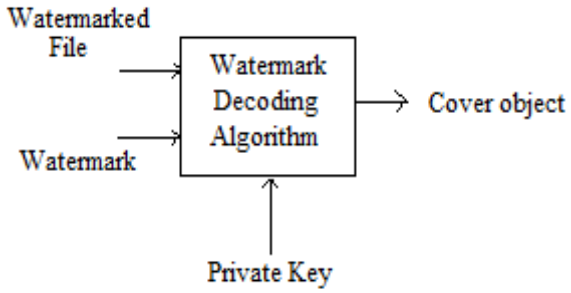


Fig.1:(b) Watermark extraction process

In general, any watermarking scheme (algorithm) consists of three parts.

- The watermark.
- The encoder (insertion algorithm).
- The decoder and comparator (verification or extraction or detection algorithm).

For a digital watermark to be effective, it should exhibit the following characteristics [3]:

1. *Adjustability.* The algorithm should be tunable to various degrees of robustness, quality, or embedding capacities to be suitable for diverse applications.
2. *Robustness.* The embedded watermarks should not be removed or eliminated by unauthorized distributors using common processing techniques, including compression, filtering, cropping, quantization and others.

3. *Security.* The watermarking procedure should rely on secret keys to ensure security, so that pirates cannot detect or remove watermarks by statistical analysis from a set of images or multimedia files. An unauthorized user, who may even know the exact watermarking algorithm, cannot detect the presence of hidden data, unless he/she has access to the secret keys that control this data embedding procedure.
4. *Imperceptibility.* The watermark should be invisible in a watermarked image/video or inaudible in watermarked digital music. Embedding this extra data must not degrade human perception about the object. Evaluation of imperceptibility is usually based on an objective measure of quality, called peak signal-to-noise ratio (PSNR) or a subjective test with specified procedures.
5. *Real-time processing.* Watermarks should be rapidly embedded into the host signals without much delay.

Watermarking Issues

There are certain issues regarding to the digital watermarking which are tried to answer ere but these are not limited to these only.

What is it?

The answer to this issue is already described above. It is a copyright mark or logo inserted in the multimedia or text file in a specific manner called algorithm so that it can be extracted only if the copyright owner wants to do it and that too if proper decoding algorithm is known exactly.

How can a digital watermark be inserted or detected?

It can be inserted and extracted with proper algorithm as shown in fig. 1(a) and fig. 1(b).

How robust does it need to be?

There are two types of watermarking as far as robustness is concerned robust and fragile. Higher robust watermark is needed if the copyright owner does not want the watermark to be extracted by himself or anyone. But fragile watermarking is also needed in those cases when authorized distribution is going to be done. The receiver receives the multimedia file along with the decoding algorithm and the key to decode the same.

Why and when are digital watermarks necessary?

When the distribution of multimedia file is done by the means of internet or by digital storage system then it can be copied and edited very easily, then watermarking becomes necessary to protect the copyright owner.

What can watermarks achieve or fail to achieve?

Watermark achieved its purpose to protect the ownership logo or copyright mark and the disputes regarding the ownership can be easily solved. But watermarking methods are not perfect against digital reformations of the file. There are different methods for different types of files (text, image, audio and video) by means of which the copyright mark can be destroyed completely or at least the mark can be destroyed in such an extent that it cannot be considered in case of any dispute.

How should digital watermarks be used?

Digital watermark must be used in a particular manner, called watermarking algorithm, which must not be a common or known method for the public. The algorithm must be kept secret. It can be extracted only with the specific decoding algorithm.

How might they be abused?

If the watermark is not embedded in a specific algorithm then the watermark can be extracted to get the original file without any copyright mark. This file can be misused and anyone can insert their watermark and can prove that this belongs to the imitator.

What are the business opportunities?

As far as business opportunities are concerned, there are two business ends which can have benefits of watermarking. One is of course the owner of the particular file and second one is the person or company which develops such method of watermarking which can be extracted without the concern of owner.

What roles can digital watermarking play in the content protection infrastructure?

The most watermarking methods developed are such that if the watermark is tried to be extracted without proper decoding algorithm, whole file gets corrupted or at least the few contents be deleted. In other words, the quality of the extracted file is much reduced.

How can we evaluate the technology?

The watermarking methods can be evaluated by certain parameters like peak signal to noise ratio (PSNR), signal to noise ratio (SNR), bit error rate (BER) or non correlation (NC).

Multimedia files and text

The watermarking technique can be applied to any text or multimedia file like image, audio or video file. The different methods of watermarking are shown in fig.2.

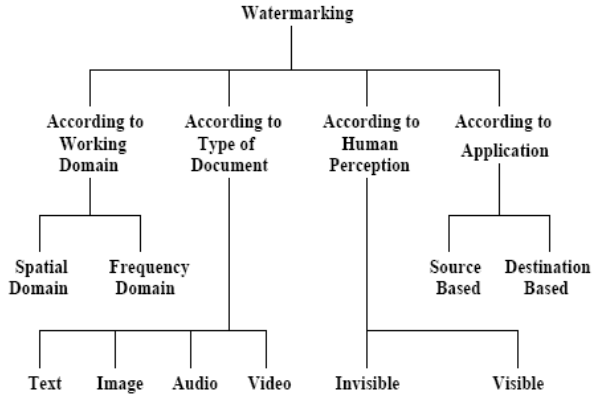


Fig.2: Watermarking methods

Watermarking techniques can be divided into four categories according to the type of document to be watermarked as follows:

- Image Watermarking [10] [11] [12]
- Video Watermarking[13] [14] [15]
- Audio Watermarking [16] [17] [18]
- Text Watermarking [19] [20] [21]

According to the human perception, the digital watermarks can be divided into two different types as follows:

- Visible watermark [22] [23]
- Invisible watermark [24] [25]

A visible watermark is a secondary image, ownership mark or a logo overlaid into the primary image. The watermark appears visible to a casual viewer on a careful inspection. The invisible-robust watermark is embedded in such a way that alternations made to the pixel value are perceptually not noticed and it can be recovered only with appropriate decoding mechanism. The invisible-fragile watermark is embedded in such a way that any manipulation or modification of the image would alter or destroy the watermark. In some cases dual watermarking is also done which means both visible and invisible watermarking is done on the same file. In this type of watermark, an invisible watermark is used as a backup for the visible watermark as clear from the fig. 3.

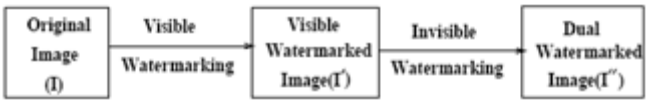


Fig. 3: Dual watermarking

According to the working domain the watermarking can be done in the spatial domain or frequency domain and according to application it can be source oriented or destination oriented.

Desired characteristics of digital watermark

The desired characteristics of digital watermark can be different depending upon the type file and type of watermarking required, that is, robust or fragile.

Desired characteristics of visible watermarks

- A visible watermark should be obvious in both color and monochrome images.
- The watermark should spread in a large or important area of the image in order to prevent its deletion by clipping.
- The watermark should be visible yet must not significantly obscure the image details beneath it.
- The watermark must be difficult to remove. Rather, removing a watermark should be more costly and labor intensive than purchasing the image from the owner.
- The watermark should be applied automatically with little human intervention and labor.

Desired Characteristics of Invisible Fragile Watermarks

- The invisible watermark should neither be noticeable to the viewer nor should degrade the quality of the content.
- An invisible fragile watermark should be readily modified when the image pixel values have been altered.
- The watermark should be secure. This means that it is impossible to recover the changes, or regenerate the watermark after image alternations, even when the watermarking procedure, and/or the watermark itself are known.
- For high quality images, the amount of individual pixel modification should be as small as possible.

Desired Characteristics of Invisible Robust Watermarks

- The invisible watermark should neither be noticeable to the viewer nor should degrade the quality of the content.
- An invisible robust watermark must be robust to common signal distortions and must be resistant to various intentional tampering solely intended to remove the watermark.
- Retrieval of watermark should unambiguously identify the owner.
- It is desirable to design a watermark whose decoder is scalable with each generation of computer.
- While watermarking high quality images and art works, the amount of pixel modification should be minimum.
- Insertion of watermark should require little human intervention or labor.

Desired Characteristics of Video and/or audio Watermarks

- The presence of watermark should not cause any visible or audible effects on the playback of the video.
- The watermark should not affect the compressibility of the digital content.
- The watermark should be detected with high degree of reliability. The probability of false detection should be extremely small.
- The watermark should be robust to various intentional and unintentional attacks.
- The detection algorithm should be implemented in circuitry with small extra cost.

Application of Digital Watermarks

Visible Watermark

Visible watermarking can be used for copyright protection for image or video files. In such cases, the content owner is in need that the images will be used commercially without payment of royalties. The content owner desires an ownership mark, that will be visually apparent, but which does not prevent image being used for other purposes. In this case,

images are made available through the internet and the content owner desires to indicate the ownership of the underlying materials.

Invisible Robust Watermark

Invisible watermarking is used to detect misappropriated images. In this case, fee-generating images may be purchased by an individual who will make them available for free. Invisible watermarking can be used as evidence of ownership. In this case, the seller of the digital images suspects that one of his images has been edited and published without payment of royalties. Here, the detection of the seller's watermark in the image is intended to serve as evidence that the published image is property of seller.

Invisible Fragile Watermarks

Invisible watermarking can be used for a trustworthy camera. In this case, images are captured with a digital camera for later use in the news articles. Here, it is the desire of a news agency to verify that an image is true to the original capture and has not been edited. In this case, an invisible watermark is embedded at capture time, its presence at the time of publication is intended to indicate that the image has not been amended since it was captured. Invisible watermarking can be used to detect alternation of images stored in a digital library. The content owner desires the ability to detect any alternation of the images, without the need to compare the images to the scanned materials.

Performance evaluation of watermarking methods

The performance of any watermarking technique can be measured in any one of the parameters like BER [26], PSNR [27], SNR [28] or Non correlation [29]. They can be calculated as:

$$BER = \frac{100}{B}\sum_{n=0}^{B-1}\begin{cases}1, & \tilde{w}(n) \neq w(n)\\ 0, & \tilde{w}(n) = w(n)\end{cases}$$

Where B is the number of blocks, which is total number of bits divided by number of samples (bits) in each block, w(n) is watermark and $\tilde{\omega}(n)$ is the extracted watermark.

PSNR = 10 ——

Where,

———————

X_1 and X_2 are the original audio sample and watermarked audio sample. 'R' is 255 as data type used is 8-bit unsigned number representation. If the data type is double precision floating type then 'R' will be 1.

Signal to noise ratio can be calculated as

$$SNR = 10 \cdot \log_{10}\left\{\frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1}[\tilde{x}(n) - x(n)]^2}\right\}$$

Where *x(n)* is the original audio signal and *x (n)* as the watermarked audio signals and the normalized correlation (NC) is used to evaluate the similarity measurement of extracted binary watermark which can be calculated as

$$NC(W, W^*) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} W(i,j)W^*(i,j)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{M} W(i,j)^2}\ \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{M} W^*(i,j)^2}}$$

Where W and W^* are original and extracted watermarks respectively, i and j are indexes of the binary watermark image.

Attacks on Watermarks

A watermarked image is likely to be subjected to certain manipulations, some intentional such as compression and transmission noise and some intentional such as cropping, filtering, etc. They are summarized in Fig.4.

Many compression schemes like JPEG and MPEG can potentially degrade the data's quality through irretrievable loss of data. Geometric distortions are specific to images videos and include such operations as rotation, translation, scaling and cropping. Common Signal Processing Operations include the following.

- D/A conversion
- A/D conversion
- Re-sampling
- Re-quantization
- Dithering distortion
- Recompression
- Linear filtering such as high pass and low pass filtering
- Non-linear filtering such as median filtering
- Colour reduction
- Addition of a constant offset to the pixel values
- Addition of Gaussian and Non Gaussian noise
- Exchange of pixels

Some other possible attacks can be as follows:

- Printing and Rescanning
- Watermarking of watermarked image (re-watermarking)
- Collusion: A number of authorized recipients of the image should not be able to come together (collude) and like the differently watermarked copies to generate an un-watermarked copy of the image.
- Forgery: A number of authorized recipients of the image should not be able to collude to form a copy of watermarked image with the valid embedded watermark of a person not in the group with an intention of framing a 3rd party.
- IBM attack [7] [8]: It should not be possible to produce a fake original that also performs as well as the original and also results in the extraction of the watermark as claimed by the holder of the fake original.
- The Unzign and Stir mark have shown remarkable success in removing data embedded by commercially available programs.

Conclusions

In this paper a clear overview on watermarking concept is provided. The types of watermarking technique, whichever is required according to the application, can be applied. Further, the algorithm developed to embed the watermark can be evaluated b means of anyone of the parameters BER, PSNR, SNR and NC and at last the embedding algorithm can also be checked that weather the watermark can be survived after the attack or not.

Lossy Compression	Geometrical Distortions	Common Signal Processing Operations	Other Intentional Tampering
JPEG	Rotation	D/A or A/D conversion	Printing
MPEG	Translation	Re-sampling	Rescanning
	Scaling	Re-quantization	Rewatermarking
	Cropping	Dithering	Collusion
		compression	Forgery
		Linear filtering	IBM attack
		Non linear filtering	Unzign attack
		Colour reduction	Stirmark attack
		Addition of offset value	
		Addition of noise	
		Exchange of pixels	

Fig. 4: Possible attacks on watermarked file

References

[1] Business Software Alliances, *Piracy study, Fourth Annual BSA And IDC Global Software, 2007,* http://www.bsa.org/globalstudy/ .

[2] Industrial Design and Construction (IDC) and Business Software Alliance (BSA), *"Piracy study," July 2004.* http://www.bsaa.com.au/downloads/ PiracyStudy070704.pdf

[3] Wen-Nung Lieand Li-Chun Chang, "Robust and High-Quality Time-Domain Audio Watermarking Based on Low-Frequency Amplitude Modification", *IEEE Transactions On Multimedia, Vol. 8, No. 1, February 2006,* pp.46-59.

[4] S.P.Mohanty, "A Dual Watermarking Technique For Images", *Proc. 7th ACM International MultimediaConference, ACM-MM'99,* Part 2, pp. 49-51, Orlando, USA, Oct. 1999.

[5] S.P.Mohanty, "A Dual Watermarking Technique For Images", *Proc. 7th ACM International MultimediaConference, ACM-MM'99,* Part 2, pp. 49-51, Orlando, USA, Oct. 1999.

[6] Saraju P. Mohanty, *Dept of Comp Sc and Eng, Unversity of South Florida*, Tampa, FL 33620.

[7] S.Craver, and alliance, "Can Invisible Watermarks Resolve Rightful Ownership?", *IBM Research Report, RC205209,* July25 1996.

[8] S. Craver and alliance, "Resolving Rightful Ownerships with Invisible Watermarking Techniques: Limitations, Attacks and Implications", *IEEE Journal. on Selected Areasin Communications,* Vol.16, No.4, May 1998, pp.573-586.

[9] Paraskevi Bassia, Ioannis Pita, and Nikos Nikolaidis, "Robust Audio Watermarking in the Time Domain", *IEEE Transactions On Multimedia,* Vol. 3, No. 2, June 2001, pp.-232-241.

[10] Myung-Ho Lee and Oh-Jin Kwon, "Color Image Watermarking Based on DS-CDMA Using Hadamard Kernel", *ICACT 2008*, Feb. 2008, pp.1592-1597.

[11] Lihong Ma and Hanqing Lu, "Adaptive Spread-transform Dither Modulation for Color Image Watermarking", *IEEE "GLOBECOM" 2008 proceedings.*

[12] A. AI-Gindya, H. AI-Ahmad, R. Qahwaji and A. Tawfik, "A Novel Blind Image Watermarking Technique for Colour RGB Images in the DCT Domain Using Green Channel", *MIC-CCA 2008*, pp.26-31.

[13] Noorkami, M. Mersereau and R.M., "Digital Video Watermarking in P-Frames With Controlled Video Bit-Rate Increase", *IEEE Transactions on information Forensics and Security, Sept. 2008*, pp 441-455.

[14] Honq-Mei Liu, Ji-Wu Huang and Zi Mei Xiao, "AN ADAPTIVE VIDEO WATERMARKING ALGORITHM", *2001 IEEE International Conference on Multimedia and Expo (ICME'01)*, August 2001.

[15] Brunton, A. Jiying Zhao , "Real-time video watermarking on programmable graphics hardware", *Canadian Conference on Electrical and Computer Engineering 2005*, May 2005, pp. 1312-1315.

[16] Xiangyang Wang, Wei Qi and Panpan Niu, "A New Adaptive Digital Audio Watermarking Based on Support Vector Regression", *IEEE Transactions On Audio, Speech, And Language Processing*, Vol. 15, No. 8, November 2007, pp. 2270-2277.

[17] Paraskevi Bassia, Ioannis Pita, and Nikos Nikolaidis, "Robust Audio Watermarking in the Time Domain", *IEEE Transactions On Multimedia, Vol. 3, No. 2, June 2001,* pp.-232-241.

[18] Lin Kezheng, Fan Bo and Yang Wei, "Robust Audio Watermarking Scheme Based on Wavelet Transforming Stable Feature", *2008 International Conference on Computational Intelligence and Security*, pp.-325-329.

[19] Xinmin Zhou, Weidong Zhao, Zhicheng Wang and Li Pan, "Security Theory and Attack Analysis for Text Watermarking", *International Conference on E-Business and Information System Security, EBISS '09,* May 2009, pp. 1-6.

[20] Hongtao Ge, Fulin Su and Yong Zhu, "**Color image text watermarking using wavelet transform and error-correcting code", *6th International Conference on Signal Processing*,** August 2002, pp.1584-1587.

[21] Qadir and Ahmad, "**Digital text watermarking:**

secure content delivery and data hiding in digital documents", *International Carnahan Conference on Security Technology, CCST '05*, Oct.2005, pp.101-104.

[22] Shang-Chih Chuang, Chun-Hsiang Huang and Ja-Ling Wu, "UNSEEN VISIBLE WATERMARKING", *ICIP 2007*, pp.III_261-III_264.

[23] Shu-Kei Yip, Oscar C. Au, Chi-Wang Ho and Hoi-Ming Wong, "Lossless Visible Watermarking", *ICME 2006*, pp.853-856.

[24] Na Li, Xiaoshi Zheng, Yanling Zhao, Huimin Wu and Shifeng Li, "Robust Algorithm of Digital Image Watermarking Based on Discrete Wavelet Transform", *International Symposium on Electronic Commerce and Security*, 2008, pp.942-945.

[25] Dumitru Dan Burdescu, Liana Stanescu, Anca Ion and Razvan Tanasie, "A NEW 3D WATERMARKING ALGORITHM*", 3DTV-CON'08*, May 2008, pp.381-384.

[26] Guo ZhiChuan and Wang JinLin Ni Hong, "A low complexity reversible data hiding method based on modulus function", *ICSP2008 Proceedings*, pp.2221-2224.

[27] Li Jing and Fenli Liu, "Applying General Regression Neural Network in Digital Image Watermarking", *Fourth International Conference on Natural Computation 2008*, pp.452-456.

[28] J. D. Gordy and L. T. Bruton, "Performance Evaluation of Digital Audio Watermarking Algorithms", *Proceedings of43rd IEEE Midwest Symposium on Circuits and Systems, Lansing MI*, Aug. 2000, pp.456-459.

[29] Chu-Hsing Lin, Jung-Chun Liu and Pei-Chen Han, "On the Security of the Full-Band Image Watermark for Copyright Protection", *2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, pp.74-80.

Morphological based Particle Analysis: A Review

Prabhdeep Singh[1] and Dr. A.K. Garg[2]

[1]Electronics & Communication Department, M M University, India
E-mail: prabhsingh13@gmail.com
[2]Electronics& Communication Department, M M University, India
E-mail:garg_amit03@yahoo.co.in

Abstract
Illumination is one of the most important factors affecting the appearance of an image. It often leads to diminished structures and inhomogeneous intensities of the image due to different texture of the object surface and shadows cast from different light source directions. This effect generates from non uniform backround illumination and its effects are adverse in case of biological images. Techniques such as segmentation, edge detection and general image processing algorithms based on 'region of interest' could not differentiate between some of the particles and their backround or neighbouring pixels. This paper is aimed to remove these problems in microscopic image processing by firstly removing the problem of non-uniform backround illumination from the image using Morphological Opening, Adaptive Histogram Equalization and Image Enhancement to transform the input image to its indexed form with maximum accuracy involving thresholding and contrast adjustment techniques and finally employing the connected-connected component strategy to find out number of particles present in the image, creating pseudo colored index image to compute the characteristics of every particle clearly and computing the characteristics of every particle and finally plotting area based statistics and histogram of the final image.

Keywords: Morphology, Histogram Equalization, Thresholding, Dialation.

Introduction
Image processing is used to modify pictures to improve them (enhancement, restoration), extract information (analysis, recognition), and change their structure (composition, image editing). Images can be processed by optical, photographic, and electronic means, but image processing using digital computers is the most common method because digital methods are fast, flexible, and precise.Image processing technology is used by planetary scientists to enhance images of Mars, Venus, or other planets. Doctors use this technology to manipulate CT scans and MRI images. Image Enhancement improves the quality (clarity) of images for human viewing. Removing blurring and noise, increasing contrast, and revealing details are examples of enhancement operations.Image Processing basically includes analysis, manipulations, storage and display of graphical images from sources such as photographs, drawings and so on. Image processing spans a sequence of 3 phases, which are the image acquire, processing and display phase. The image acquires phase converts the differences in colouring and shading in the picture into binary values that a computer can process. The enhancement phase can include image enhancement and data compression. The last phase consists of display or printing of the processed image. The term **morphology** means form and structure of an object. Sometimes it refers to the arrangements and inter-relationships between the parts of an object. Morphology is related to the shapes and digital morphology is a way to describe and analyse the shape of a digital object. In biology, morphology relates more directly to shape of an organism such as bacteria. Morphological opening is a name specific technology that creates an output image such that value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbours. By choosing the size and shape of the neighbourhood, one can construct a morphological operation that is sensitive to specific shapes in the input image. Morphological functions could be used to perform common image processing tasks, such as contrast enhancement, noise removal, thinning, skeletonization, filling and segmentation.

Related Work, Issues and Possible Solutions
There are many research papers that propose Histogram modelling techniques to modify an image so that its histogram has a desired shape. This is useful in stretching the low-contrast levels of an image with a narrow histogram, thereby achieving contrast enhancement. This is driven by the lack of removal of non-uniform illumination, as a result of which presence of extra light pixels at some positions in the image and extra dark pixels around other particles in the image is variable, so this contrast enhancement at the starting of the image processing does not create the accurate boundaries of the objects to be detected.[8] Komal Vij,et al. has proposed image enhancement method to increase image visibility and details. They covers all the factors like enhancement efficiency, computational requirements, noise amplification, user intervention, and application suitability. [10] Yan Wan,et al.has proposed a dual threshold calculating method to obtain accurate and continuous fiber edge, as well as to control the image noise.[11] M. Kowalczyk,et al.has proposed conception of effectively working groups of morphology functions in particular image cases. [12]David Menotti has proposed two methodologies for fast image contrast enhancement based on

histogram equalization (HE), one for gray-level images, and other for color images. For gray-level images, technique called Multi-HE has been proposed.[13] Ley,et al.has proposed a simple background illumination correction based approach for improving matting problems with uneven or poor lit blue-/green screens.[14] Joanna Sekulska,et al.has proposed general methods of biological images processing. These techniques are oriented to better image interpretation.[7] has proposed an automatic method for estimating the illumination field using only image intensity gradients.[15] has proposed a novel model-based correction method is proposed, based on the assumption that an image corrupted by intensity inhomogeneity contains more information than the corresponding uncorrupted image.[9]has proposed a new approach to the correction of intensity inhomogeneities in magnetic resonance imaging (MRI) that significantly improves intensity-based tissue segmentation.[3] Yadong Wu,et al. has proposed an image illumination correction algorithm based on tone mapping. The proposed algorithm combined color space decomposition and tone mapping based image brightness adjustment, which can improve the image contrast while maintaining the better color of the original image, and cannot increase noise. [4] M Rama Bai has Introduced a novel algorithm based on multi-scale morphological method for the purpose of border detection. Standard morphological border detection methods use single and symmetrical structure elements. [2] **YanFeng Sun**,et al. has proposed a Multi-scale Fusion TV-based Illumination Normalized (MFTVIN) model. In this Illumination effects in the large-scale part are removed by region-based histogram equalization and homomorphic filtering.[1] Emerson Carlos Pedrino,et al.has proposed an original reconfigurable architecture using logical, arithmetic, and morphological instructions generated automatically by a genetic programming approach. They also presented Binary, gray, and color image practical applications using the developed architecture.

Technology Overview

Mathematical morphology (MM) is a theory and technique for the analysis and processing of geometrical structures, based on set theory, lattice theory, topology, and random functions. MM is most commonly applied to digital images, but it can be employed as well on graphs, surface meshes, solids and many other spatial structures, Topological and Geometrical space concepts such as size, shape, convexity,and geodesic distance, can be characterized by MM on both continuous and discrete spaces. MM is also foundation of morphological image processing, which consists of a set of operators that transform images according to the above characterizations.

Denoting an image by *f(x)* and the structuring function by *b(x)*, the grayscale dilation of *f* by *b* is given by:

$$(f \oplus b)(x) = \sup_{y \in E}[f(y) + b(x - y)] \quad -1$$

where "sup" denotes the supremum.Similarly, the erosion of *f* by *b* is given by

$$(f \ominus b)(x) = \inf_{y \in E}[f(y) - b(y - x)] \quad -2$$

MM was originally developed for binary images, and was later extended to grayscale functions and images. The subsequent generalization to complete lattices is widely accepted today as MM's theoretical foundation. In binary morphology, an image is viewed as a subset of a Euclidean space Rd or the integer grid Zd, for some dimension *d*.

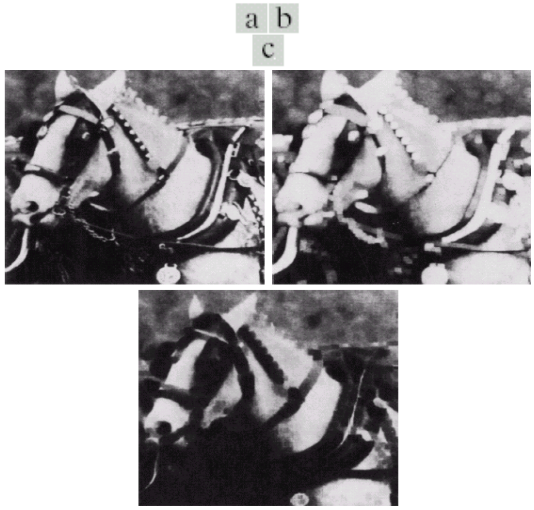


Fig.1(a) Original Image (b) Result of Dialation (c) Result of Erosion.

Some of the basic elements of binary morphology includes Structuring Element, Basic Operators : Euclidean Distance and Shift Invariant, Erosion and Dilation and Opening and Closing of binary Images. The basic idea in binary morphology is to probe an image with a simple, pre-defined shape, drawing conclusions on how this shape fits or misses the shapes in the image. This simple "probe" is called structuring element, and is itself a binary image. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. In greyscale morphology, images are functions mapping a Euclidean space or grid *E* into R U {∞, -∞}, where R is set of reals, ∞ is an element larger than any real number, -∞ is an element smaller than any real number. Grayscale structuring elements are also functions of the same format, called "structuring functions".

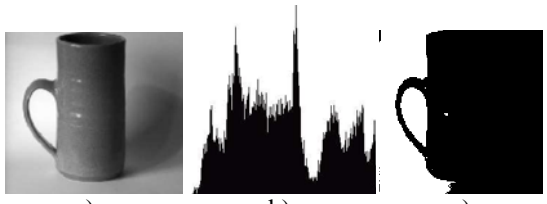


Fig.2 (a) The mug image; (b) its histogram; (c) thresholded using global segmentation at T=16.[16]

Various techniques and common approaches to solve the problem of particle identification are Image Filtering, Boundary detection, Edge Detection, Linear Filtering,

Segmentation, Morphological operations: Dilation and Erosion etc. But most of these techniques alone fail to accurately determine the objects real boundaries due to the problem of non-uniform illumination in the backround of the image due to which most of the particles appear to be either dark or light in an image and using techniques such as segmentation, edge detection and general image processing algorithms based on 'region of interest' could not differentiate between some of the particles and their backround or neighboring pixels. Even when the particles are extracted, there are changes to their shape and size which leads to faulty readings in the computations of area of such particles. So, advanced image processing and image enhancement tools have to be used for maximum accuracy of the results and to identify the particles accurately from the image without even missing a single object.

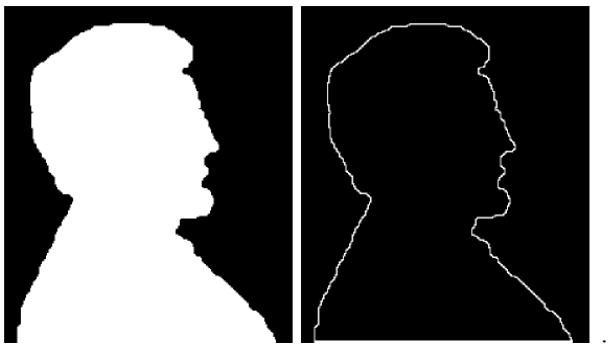


Fig.3 Boundary extraction usng 3*3 square element.[4]

The Following is the flow-diagram of algorithms followed in **previous approaches** to solve this desired problem:

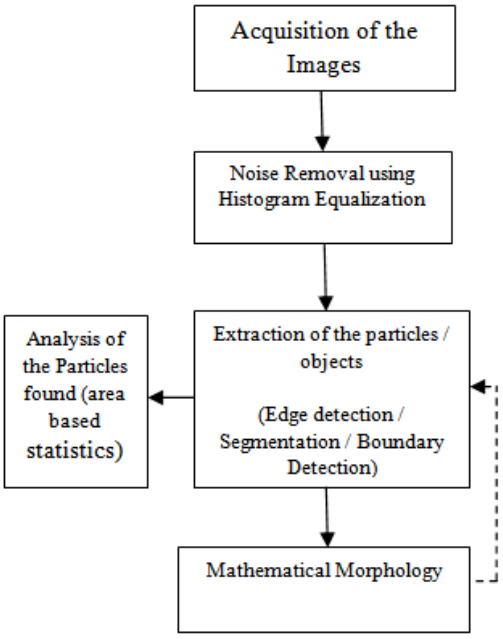


Fig.4 Image processing algorithm previously used.

Considering the above algorithm, the microscopic images are first undergone through histogram equalization, a technique that represents the relative frequency of occurrence of grey levels within an image. Histogram modelling techniques modify an image so that its histogram has a desired shape. This is useful in stretching the low-contrast levels of an image with a narrow histogram, thereby achieving contrast enhancement. In histogram equalization (HE), the goal is to obtain a uniform histogram for the output image, so that "optimal" overall contrast is perceived.

Conclusions and Future Work

It has been concluded that due to non-uniform background illumination, most of the particles appear to be either dark or light in an image and using techniques such as segmentation, edge detection and general image processing algorithms based on 'region of interest' could not differentiate between some of the particles and their background or neighbouring pixels. In future, it is plan to perform the simulations and analysis of an image and its enhancement to correct for non uniform illumination, then use the enhanced image to identify discrete objects/particles present in the image. The work to be done is based upon images taken for biological studies such as images consisting of cluster of cells, bacteria, or other particles where it is important to find out the concentration of the particles. So, a particular defined area of a photographic plate is taken and exposed by the particles the characteristics of which are to be computed. So, the technique used would be to make an algorithm to finally examine every particle of the image, to see clearly every object in the image, and remove any of the problems such as non-uniform illumination, less brightness etc. that make it difficult to differentiate between the particles. Finally, when all the visualization and analyzing problems are removed, the characteristics of each particle, its area has to be computed and results would be shown in area based statistics and histogram equalization.

References

[1] Emerson Carlos Pedrino,et al."A Genetic Programming Approach to Reconfigure a Morphological Image Processing Architecture", pp1-10,2011.

[2] YanFeng Sun,et al.**"A Multi-scale TVQ-based Illumination Normalization Model" ,vol.I pp1-6, 2011.**

[3] Yadong Wu,et al."An Image Illumination Correction Algorithm based on Tone Mapping", IEEE pp245-248, 2010.

[4] M Rama Bai,"A New Approach For Border Extraction Using Morphological Methods", pp3832-3837, 2010.

[5] Kevin Loquin,et al. "Convolution Filtering And Mathematical Morphology On An Image: A Unified View", pp1-4, 2010.

[6] Przemysław Kupidur,"Semi-automatic method for a built-up area intensity survey using morphological

granulometry",pp271-277, 2010.
[7] Abhishek Acharya,et al. "FPGA Based Non Uniform Illumination Correction in Image Processing Applications Vol 2", pp349-358, 2009.
[8] Komal Vij,et al. "Enhancement of Images Using Histogram Processing Techniques Vol 2" , pp309-313, 2009.
[9] Jean-Michel Morel,et al. "Fast Implementation of Color Constancy Algorithms Vol.19", pp2825-2837, 2009.
[10] Yan Wan,et al."A Dual Threshold Calculating Method for Fiber's Edge Extraction", IEEE pp 247-254,2009.
[11] M. Kowalczyk,et al."Application of mathematical morphology operations for simplification and improvement of correlation of images in close-range photogrammetry",pp153-158, 2008.
[12] David Menotti,"Contrast Enhancement in Digital Imaging using Histogram Equalization" ,pp1-85, 2008.
[13] Ley,et al."GPU-Based Background Illumination Correction for Blue Screen Matting" , pp1-102, 2007.
[14] Joanna Sekulska,et al." Digital image processing methods in biological structure recognition a short review",pp24-27, 2006.
[15] M. Ranzato,et al."Automatic Recognition of Biological Particles in Microscopic Images",pp1-19, 2006.
[16] Torsten Seemann,"Digital Image Processing using Local Segmentation",Thesis B. Sc (Hons),pp-61,2002.

Effect of Pulse Shaping on BER Performance of QAM Modulated OFDM Signal

D.K. Sharma[1], A. Mishra[2] and Rajiv Saxena[3]

[1]Ujjain Engineering College, Ujjain, MP, India
[2]Madhav Institute of Technology & Science, Gwalior, MP, India
[3]Jaypee Institute of Engineering & Technology, Guna, MP, India
E-mail: dilip_sharma1172@yahoo.com; drabhaymishra@yahoo.com; rajiv.saxena@jiet.ac.in

Abstract
Pulse shaping of a signal at base-band forms a more important tool in controlling and analyzing the parameters like ISI, the modulated signal envelope uniformity, the phase continuity as these parameters lead to the efficient spectral confinement of a binary signal which is a future and nowadays a major important aspect in the design of a wireless cellular communication system.

In this paper, different time-limited waveform are discussed which are basically available as windows functions and comparing the performance of Rectangular, Raised cosine and BTRC and prove that the BTRC pulse is supposed to be a batter pulse shaping for OFDM in terms of obtaining better BER performance and proposed two pulse shapes which are giving better $E[\rho_i^2]$ in respective cases which in turn provides better BER as compared to BTRC which was considered as one of the best pulse shape.

Keywords: Inter Symbol Interference, Inter Channel Interference, Signal to Noise Ratio, Bit Error Rate.

Introduction

A linear filter channel distorts the transmitted signal and the parameter like ISI is an undesired overlap of the neighboring symbols both in time and in phase. The channel distortions results in ISI at the output of the demodulator which leads to an increase in ISI at the output of the demodulator and leads to an increase in the probability of error at the detector and hence BER performance is degraded. Greater ISI allows the spectrum to be more compact, making demodulator more complex, hence, spectral compactness is the primary trade-off in going from one type of pulse shaping to other type i.e. the pulse shaping can take care of the BER performance of the OFDM system and can be a major tool to reduce the ISI and ICI effects efficiently in an OFDM system [1, 2]. The pre-modulation low pass filter (LPF) must have a narrow Bandwidth with sharp cutoff frequency and very little overshoot in its impulse response.

In this paper, different time-limited waveform are discussed which are basically available as windows functions. The synthesis of window function is an important technique for pulse shaping [1], [3] and the pulse shaping gives a considerable performance improvement of OFDM systems in multipath radio channels as compared to the conventional OFDM. Therefore, windowing finds a wide and large area of applications extending from spectrum estimation, digital filtering, speech processing, surface acoustic wave (SAW) filter design [4], [5] and many other fields of communication and signal processing.

An exhaustive review of many pulses shaping functions or time limited waveform is done on the basis of their application and characteristics in [3] with a conclusion that the Comparing the performance of Rectangular, Raised cosine and BTRC, the BTRC gives smaller equivalent noise power and hence lower BER values. Therefore, the BTRC pulse is supposed to be a batter pulse shaping for OFDM in terms of obtaining better BER performance [6].

There after a good number of windows have been proposed by various authors with some performing better in the frequency domain and some in the time domain. A concise comparative study has also been performed in different way [7]. The main criterion for the selection of a window is application oriented, means that whether the application requires better spectral performances or the better time domain performances [8] and it can be concluded that the two proposed pulses are giving better $E[\rho_i^2]$ in respective cases which in turn provides better BER as compared to BTRC which was considered as one of the best pulse shape [6].

ICI IN OFDM

The pulse shaping plays a major role in improvising the Bit Error Rate (BER) performance of modulated OFDM system [9], [7], since the BER is related to the signal to noise ratio (SNR) of a signal where signal power is supposed to be the second moment of the amplitude of the pulse shaping signal, where it can be observed that the role off factor ‘α’ is present in the second moment of amplitude which indicates that the role off factor is responsible for the performance of BER of the signal. This indicates that a play in the role off factor for a particular pulse shaping can improve the BER performance.

Since, ICI power is the nothing but the mean square value of an ICI term present in an OFDM transmitted signal and is dependent on the amplitude or magnitude of the weighting function which is otherwise also known as pulse shaping function. Thus, it is evident that the play in the role off factor of a pulse shape will affect the ICI power also of the transmitted OFDM signal.

Thus, in order to analyze various pulse shapes in the light of BER performance firstly the amplitude moment of each of them has to be calculated, the ICI of an OFDM signal is calculated as per equation (1).

where, ρ_i= imperfect carrier offset (ICI)

$$\rho_i = \sum_{k=0}^{N-1} a_i[k] \cdot g(-kT - \tau_i) = \sum_{\substack{k \neq m \\ l=0}}^{N-1} a_{l,i} \cdot C_{l-m} \cong \sum_{\substack{k \neq m \\ k=0}}^{N-1} a_k \cdot C_{k-m} \quad (1)$$

and the equivalent noise $E[\rho_i^2]$ in geometrical progression (G. P.) is derived as-

$$E\left[\rho_i^2\right] = \sum_{k=0}^{N-1} E\left(\left|a_k\right|^2\right) \cdot E\left(\left|C_{k-m}\right|^2\right) \quad (2)$$

$$E\left[\rho_i^2\right] = \sum_{-\infty}^{\infty} E\left(\left|C_{k-m}\right|^2\right)$$

$$E\left[\rho_i^2\right] = \frac{1}{T}\sum_{-\infty}^{\infty} g^2(-kT-\tau)d\tau = \frac{1}{T}\sum_{-\infty}^{\infty} \int_{-kT}^{-(k+1)T} g^2(t)dt$$

replacing the factor (- Kt - τ) by t as-

$$E\left[\rho_i^2\right] = \frac{1}{T}\int_{-\infty}^{\infty} g^2(t)dt = \frac{1}{T}\int_{-\infty}^{\infty} |G(f)|^2 df \quad (3)$$

which is nothing but the Parseval's Theorem.

Types of Pulse Shapes in OFDM

As lot of pulse shaping functions have been designed, optimized and applied for various applications, with a result some performing better in time domain whereas some in frequency domain. In this section some most commonly used pulse shaping functions which define a class of modulation techniques are included [3, 10-11, 12 & 13].

Case-I: The Rectangular Pulse

$$G(f) = \frac{T}{\pi} rect\left(\frac{fT}{\pi}\right) \quad (4)$$

Case-II: Raised Cosine (RC) Pulse

$$G(f) = T \text{ , for } 0 \le |f| \le \frac{(1-\alpha)}{2T}$$

$$G(f) = \frac{T}{2}\left[1 + \cos\left\{\frac{\pi T}{\alpha}\left(|f| - \frac{1-\alpha}{2T}\right)\right\}\right],$$

for

$$\frac{(1-\alpha)}{2T} \le |f| \le \frac{(1+\alpha)}{2T}$$

$$G(f) = 0 \text{ , otherwise} \quad (5)$$

where, the alpha (α) is known as the pulse shaping factor lies $0 \le \alpha \le 1$.

Case-III: Better than Raised Cosine (BTRC) Pulse

$$G(f) = T \text{ , for } 0 \le |f| \le \frac{(1-\alpha)}{2T}$$

$$G(f) = \frac{T}{2} e^{-\frac{2T\log 2}{\alpha}\left[|f| - \frac{1-\alpha}{2T}\right]} \text{, for } \frac{(1-\alpha)}{2T} \le |f| \le \frac{1}{2T}$$

$$G(f) = \frac{T}{2} e^{-\frac{2T\log 2}{\alpha}\left[\frac{1+\alpha}{2T} - |f|\right]} \text{, for } \frac{1}{2T} \le |f| \le \frac{(1+\alpha)}{2T}$$

$$G(f) = 0 \text{ , otherwise} \quad (6)$$

Thus, the BTRC pulse shape has a small BER in presence of ISI and symbol timing error in an OFDM or base band system in comparison to the rectangular and raised cosine pulse modulated system and a BTRC modulated signal represents a better eye diagram with respect to that of the Rectangular and Raised cosine modulation [6].

Case-IV: Proposed-I Pulse Shapes

$$G(f) = T \text{ , for } 0 \le |f| \le \frac{(1-\alpha)}{2T}$$

$$G(f) = \frac{T}{2}\left[(\beta/2) + (1-\beta)\cos\left\{\frac{\pi T}{\alpha}\left(|f| - \frac{1-\alpha}{2T}\right)\right\} + (\beta/2)\cos\left\{\frac{2\pi T}{\alpha}\left(|f| - \frac{1-\alpha}{2T}\right)\right\}\right]$$

$$\text{, for } \frac{(1-\alpha)}{2T} \le |f| \le \frac{(1+\alpha)}{2T}$$

$$G(f) = 0 \text{ , otherwise} \quad (7)$$

where, "α" is role off factor lies in between zero and one and "β" is the weighting coefficient of proposed-I such that $0 \le \beta \le 1$.

Case-V: Proposed-II Pulse Shape

$$G(f) = T \text{ , for } 0 \le |f| \le \frac{(1-\alpha)}{2T}$$

$$G(f) = \frac{T}{2}\left[\beta - (4\beta - 2)\frac{|f|}{\alpha} + (1-\beta)\cos\left\{\frac{\pi T}{\alpha}\left(|f| - \frac{1-\alpha}{2T}\right)\right\}\right]$$

$$\text{, for } \frac{(1-\alpha)}{2T} \le |f| \le \frac{(1+\alpha)}{2T}$$

$$G(f) = 0 \text{ , otherwise} \quad (8)$$

where, β ($0 \le \beta \le 1$) is the weighting coefficient of proposed-II pulse shape and "α" is roll of factor of proposed-II pulse shapes.

These above said pulse shapes are expected to provide better BER performance than the other pulse shapes defined earlier in this section.

Effect of Pulse Shaping on BER Performance in OFDM

Case-I: Effect of Rectangular Pulse Shaping on BER

Considering the spectrum of a rectangular pulse from equation (4) and using equation (3) the second moment or power contained in this type of pulse can be evaluated as-

$$E\left[\rho_i^2\right]=\frac{1}{T}\int_{-\infty}^{\infty}|G(f)|^2\,df$$

$$E\left[\rho_i^2\right]=\frac{1}{T}\int_{-\infty}^{\infty}\left|\frac{T}{\pi}\right|^2 df=\frac{1}{T}\cdot\frac{T^2}{\pi^2}\int_{-\frac{\pi}{2T}}^{\frac{\pi}{2T}}1\,df$$

$$E\left[\rho_i^2\right]=\frac{1}{T}\cdot\frac{T^2}{\pi^2}1\Big|_{-\frac{\pi}{2T}}^{\frac{\pi}{2T}}=\frac{T}{\pi^2}\cdot\frac{\pi}{T}$$

$$E\left[\rho_i^2\right]=\frac{1}{\pi} \tag{9}$$

Case-II: Effect of RC Pulse Shaping on BER

Consider the spectrum of RC pulse from equation (5) and using the equation (3) for evaluating the second moment as-

$$E\left[\rho_i^2\right]=\frac{1}{T}\int_{-\infty}^{\infty}|G(f)|^2\,df$$

$$E\left[\rho_i^2\right]=\frac{1}{T}\left\{2T^2\frac{(1-\alpha)}{2T}+\frac{T^2}{2}\int_{\frac{1-\alpha}{2T}}^{\frac{1+\alpha}{2T}}\left[1+\cos\left(\frac{\pi T}{\alpha}\left(f-\frac{1-\alpha}{2T}\right)\right)\right]^2 df\right\}$$

$$E\left[\rho_i^2\right]=(1-\alpha)+\frac{\alpha}{2}\int_0^1\left(1+\cos(\pi\cdot u)\right)^2 du$$

$$E\left[\rho_i^2\right]=(1-\alpha)+\frac{\alpha}{2}\int_0^1\left(1+\cos^2(\pi\cdot u)+2\cos(\pi\cdot u)\right)du$$

$$E\left[\rho_i^2\right]=(1-\alpha)+\frac{\alpha}{2}\left[1+\int_0^1\frac{1+\cos(2\pi\cdot u)}{2}du\right]$$

$$E\left[\rho_i^2\right]=(1-\alpha)+\frac{\alpha}{2}\left[1+\frac{1}{2}\right]=(1-\alpha)+\frac{3\alpha}{4}$$

$$E\left[\rho_i^2\right]=1-\frac{\alpha}{4} \tag{10}$$

Case-III: Effect of BTRC Pulse Shaping on BER

Consider the spectrum of BTRC pulse from equation (6) using the equation (3) for evaluation of second moment for this type of pulse as-

$$E\left[\rho_i^2\right]=\frac{1}{T}\int_{-\infty}^{\infty}|G(f)|^2\,df$$

$$E\left[\rho_i^2\right]=\frac{1}{T}\left\{\begin{array}{l}2T^2\left(\frac{1-\alpha}{2T}\right)+2T^2\int_{\frac{1-\alpha}{2T}}^{\frac{1}{2T}}\exp\left[\frac{4T\log 2}{\alpha}\left(\frac{1-\alpha}{2T}-f\right)\right]df\\ +2T^2\int_{\frac{1}{2T}}^{\frac{1+\alpha}{2T}}\left[1-\exp\left\{\frac{2T\log 2}{\alpha}\left(f-\frac{1+\alpha}{2T}\right)\right\}\right]^2 df\end{array}\right\}$$

$$E\left[\rho_i^2\right]=(1+\alpha)+\frac{3}{8}\left(\frac{\alpha}{\log 2}\right)+\left(\alpha-\frac{\alpha}{\log 2}+\frac{3}{8}\left(\frac{\alpha}{\log 2}\right)\right)$$

$$E\left[\rho_i^2\right]=1-\frac{\alpha}{4\log 2} \tag{11}$$

Case-IV: Effect of Proposed-I Pulse Shaping on BER

Consider the spectrum of Proposed-I pulse from equation (7) using the equation (3) for evaluation of second moment for this type of pulse as-

$$E\left[\rho_i^2\right]=\frac{1}{T}\int_{-\infty}^{\infty}|G(f)|^2\,df$$

$$E\left[\rho_i^2\right]=\frac{1}{T}\left\{2T^2\left(\frac{1-\alpha}{2T}\right)+2T^2\int_{\frac{1-\alpha}{2T}}^{\frac{1+\alpha}{2T}}\left|\begin{array}{l}(\beta/2)+(1-\beta)\cos\left(\frac{\pi T}{\alpha}\left|f\right|-\frac{1-\alpha}{2T}\right)\\ +(\beta/2)\cos\left(\frac{2\pi T}{\alpha}\left|f\right|-\frac{1-\alpha}{2T}\right)\end{array}\right|^2 df\right\}$$

$$E\left[\rho_i^2\right]=(1-\alpha)+\frac{\alpha}{2}\left\{(\beta/2)^2+\frac{1}{2}(1-\beta)^2+(\beta^2/8)\right\}$$

at β = 1/2, above equation becomes as-

$$E\left[\rho_i^2\right]=1-\frac{57\alpha}{64} \tag{12}$$

Case-V: Effect of Proposed-II Pulse Shaping on BER

Consider the spectrum of Proposed-II pulse from equation (8) using the equation (3) for evaluation of second moment for this type of pulse as-

$$E\left[\rho_i^2\right]=\frac{1}{T}\int_{-\infty}^{\infty}|G(f)|^2\,df$$

$$E\left[\rho_i^2\right]=\frac{1}{T}\left\{2T^2\left(\frac{1-\alpha}{2T}\right)+2T^2\int_{\frac{1-\alpha}{2T}}^{\frac{1+\alpha}{2T}}\left|\begin{array}{l}\beta-\left((4\beta-2)\frac{|f|}{\alpha}\right)\\ +(1-\beta)\cos\left(\frac{\pi T}{\alpha}|f|-\frac{1-\alpha}{2T}\right)\end{array}\right|^2 df\right\}$$

$$E\left[\rho_i^2\right]=(1-\alpha)+\frac{\alpha}{2}\left\{\left(\frac{3\beta^2-2\beta+1}{2}\right)+\left(\frac{4\beta-2}{3\alpha^2}\right)(4\beta-2\alpha\beta-2)\right\}$$

at, β = 1/2, above equation becomes as-

$$\mathrm{E}\left[\rho_i^2\right] = 1 - \frac{29\alpha}{32} \tag{13}$$

Result and Discussion:

Thus, the overall pulse shaping performance in OFDM QAM based system as discussed in case of Rectangular, Raised Cosine, BTRC, Proposed-I and Proposed-II pulse shapes can be summarized and comparisons between the mean square value of ICI terms in all five cases is presented in Table-1.

Table-1: The value of $E[\rho_i^2]$ for different types of pulse shapes in OFDM with QAM.

Types of Pulse Shapes	Second Moment
Rectangular Pulse	$\mathrm{E}_R\left[\rho_i^2\right] = \frac{1}{\pi}$
RC Pulse	$\mathrm{E}_{RC}\left[\rho_i^2\right] = 1 - \frac{\alpha_{RC}}{4}$
BTRC Pulse	$\mathrm{E}_{BTRC}\left[\rho_i^2\right] = 1 - \frac{\alpha_{BTRC}}{4\log 2}$
Proposed-I Pulse	$\mathrm{E}_{Proposed\text{-}I}\left[\rho_i^2\right] = 1 - \frac{57\alpha_{Proposed\text{-}I}}{64}$
Proposed-II Pulse	$\mathrm{E}_{Proposed\text{-}II}\left[\rho_i^2\right] = 1 - \frac{29\alpha_{Proposed\text{-}II}}{32}$

In order to maintain the same BER performance, it is mandatory that the equivalent noise power must be equivalent to each other, i.e.

$$\mathrm{E}_{RC}\left[\rho_i^2\right] = \mathrm{E}_{BTRC}\left[\rho_i^2\right] \text{ or } 1 - \frac{\alpha_{RC}}{4} = 1 - \frac{\alpha_{BTRC}}{4\log 2}$$

Hence, $$\alpha_{BTRC} = 0.693\alpha_{RC} \tag{14}$$

But, it is clear from equation (10) and (11) that BTRC pulse has lower equivalent noise power; therefore it will have better BER performance as compared to Rectangular and RC pulses.

The second moment of Interference "I" completely described the Gaussian distribution and its variance defined as-

$$Var[\mathrm{I}] = \sigma_{ICI}^2 = \mathrm{E}\left[\mathrm{I}^2\right] = \frac{\Omega_1 \mathrm{T}\sum_{i=1}^{L} \mathrm{P}_i \mathrm{E}\left[\rho_i^2\right]}{4} \tag{15}$$

The value of $E[\rho_i^2]$ is derived in equations (9-13) for different types of pulse shapes in OFDM with QAM. The total equivalent noise power is defined as-

$$\sigma_t^2 = 1 + Var[\mathrm{I}] \tag{16}$$

Hence, average BER, condition on R_s, is defined by-

$$\mathrm{P}_b\Big|_{R_s = r_s} = Q\left\{\sqrt{\frac{\mathrm{P}_s \mathrm{T}}{2\sigma_t^2}} \cdot r_s\right\} \tag{17}$$

$$\alpha \uparrow - E[\rho_i^2] \downarrow - Var[\mathrm{I}] \downarrow - \sigma_t^2 \downarrow - P_b \text{ (BER)} \downarrow \tag{18}$$

eq. (15 to 19) - eq. (21)-eq. (22)-eq. (23) (19)

Conclusion

It is observed that the BER performance obtained from the Gaussian Approximation depends only on total interference signal power as per equation (15). It is also observed that larger value of α gives smaller value for mean square value $\{E[\rho_i^2]\}$ and hence smaller values for Var[I] and σ_t^2, as from equation (17), it is clear that the BER will also be correspondingly reduced. Therefore, from the Gaussian Approximation equation (17), it is clear that the value of excess bandwidth α provides a tradeoff between spectral efficiency and detection performance in an OFDM system.

Comparing the performance of Rectangular, Raised cosine and BTRC form equation (9) to (11), the BTRC gives smaller equivalent noise power and hence lower BER values. Therefore, the BTRC pulse is supposed to be a batter pulse shaping for OFDM in terms of obtaining better BER performance [6].

It can be concluded that the two proposed pulses are giving better $E[\rho_i^2]$ in respective cases which in turn provides better BER as compared to BTRC which was considered as one of the best pulse shape [6].

References

[1] A. Mishra, R. Saxena and Y. M. Gupta, "Signaling Techniques in CPM," Indian Journal of Telecommunications, vol. 54, no. 4, pp. 19-29, July-August 2004.

[2] H. G. Ryu, C. X. Wang and S. B. Ryu, "Nonlinear Distortion Reduction for the Improvement of the BER Performance in OFDM Communication Systems," IEEE Proceedings of International Symposium on Communications and Information Technologies, pp. 66-71, 2007.

[3] J. K. Gautam, A. Kumar and R. Saxena, "WINDOWS: A Tool in Signal Processing," IETE Technical Review, vol. 12, no. 3, pp. 217-226, May-June 1995.

[4] A. Vallavaraj, B. G. Stewart and D. K. Harrison, "An evaluation of modified μ-Law companding to reduce the PAPR of OFDM systems," International Journal of Electronics and Communications (AEU), vol. 64, pp. 844–857, 2010.

[5] R. Saxena and K. Singh, "Fractional Fourier Transform: A Novel Tool For Signal Processing," Journal of the Indian Institute of Science, vol. 85, no. 1, pp. 11-26, Jan.-Feb. 2005.

[6] P. Tan and N. C. Beaulieu, "Reduced ICI in OFDM systems using the "better than" raised-cosine pulse," IEEE Communication Letters, vol. 8, pp. 135-137, Mar. 2004.

[7] P. Tan and N. C. Beaulieu, "Improved BER performance in OFDM system with frequency offset

by novel pulse-shaping," IEEE Communications Society, Globecom 2004, pp. 230-236, 2004.

[8] R. Sood and H. Xiao, "Root Nyquist Pulses with an Energy Criterion," IEEE Communications Society the ICC 2007 proceedings, pp. 2711-2716, June 16, 2009.

[9] P. K. Vitthaladevuni, M. S. Alouini and J. C. Kieffer, "Exact BER computation for cross QAM constellations," IEEE Transactions on Wireless Communications, vol. 4, no. 6, pp. 3039-3050, November 2005.

[10] J. K. Gautam, A. Kumar and R. Saxena, "Correction to and Comments on-Minimum Bias Window for High Resolution Spectral Estimates," IEEE Transaction on Information Theory, vol. 42, no. 3, pp. 1001, May 1996.

[11] J. K. Gautam, A. Kumar and R. Saxena, "On the Modified Bartlett-Hanning Windows (Family)," IEEE Transaction on Signal Processing, vol. 44, no. 8, pp. 2098-2102, August 1996.

[12] N. C. Beaulieu and P. Tan, "On the Effects of Receiver Windowing on OFDM Performance in the Presence of Carrier Frequency Offset," IEEE Transactions on Wireless Communications, vol. 6, no. 1, pp. 202-209, January 2007.

[13] R. U. Mahesh and A. K. Chaturvedi, "Closed Form BER Expressions for BPSK OFDM Systems with Frequency Offset," IEEE Communications Letters, vol. 14, no. 8, pp. 731-733, August 2010.

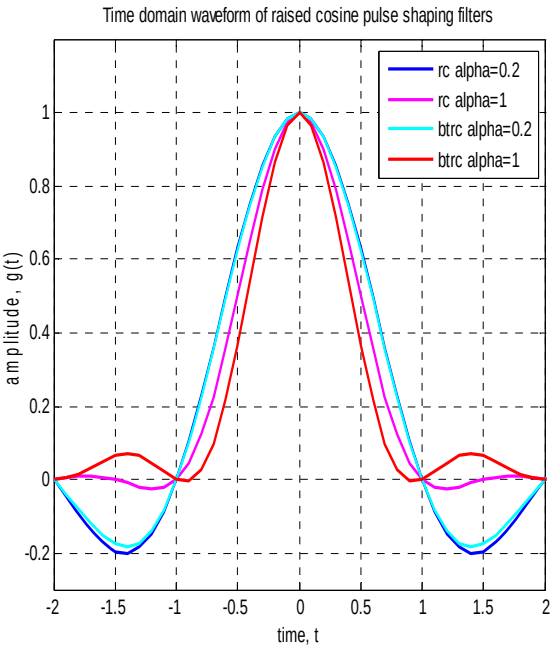


Fig. 1: RC and BTRC plotted in time domain for different value of α

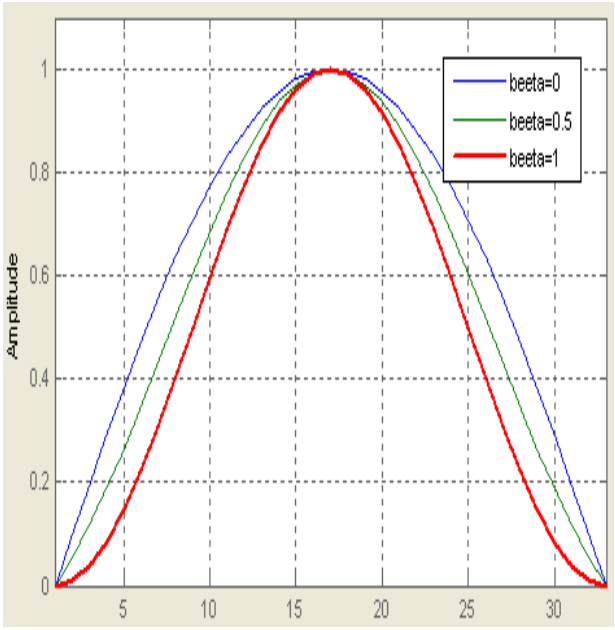


Fig. 2: Time domain of Proposed-I Pulse for different value of β = 0, β = 0.5 & β = 1

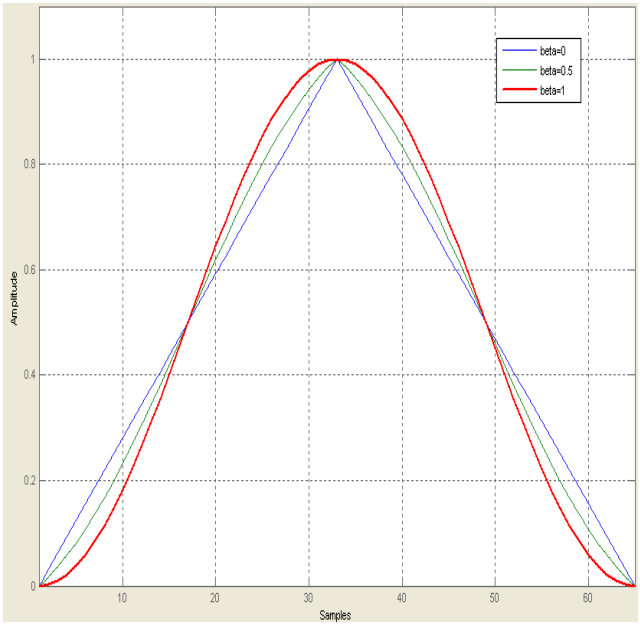


Fig. 3: Times domain Representation of Proposed-II Pulse Shape for different value of β

Prediction of Contextual Sequential Pattern Mining with Progressive Database

Uma N Dulhare[#1], Kavitha Pachika[#2], Prof. P. Premchand[#3] and Prof. K. Sandyarani[#4]

[#1,3]Department of CSE Univesityi,College of Engg Osmania University, Hyderabad, A.P., India
[#,2,4]Department of CSE, S.P.M.V.V University, Tirupathi, A.P., India
E-mail: [1]umarani_puligila@yahoo.co.in[3]drpremchand_p@yahoo.com
[3]kavitha_pachika@gmail.com[4]kasireddy_sanyrani@gmail.com

Abstract
Sequential pattern mining is an important data-mining method for determining time-related behaviour in sequence databases. The information obtained from sequential pattern mining can be used in marketing, medical records, sales analysis, and so on. When sequential patterns are generated, the newly arriving patterns may not be identified as frequent sequential patterns due to the existence of old data and sequences. users are normally more interested in the recent data than the old ones. To capture the dynamic nature of data, addition and deletion, Haung[7] proposed a progressive algorithm Pisa, which stands for Progressive mining of Sequential pAtterns, to progressively discover sequential patterns in defined time period of interest (POI). The POI is a sliding window continuously advancing as the time goes by. Pisa utilizes a progressive sequential tree to efficiently maintain the latest data sequences, discover the complete set of up-to-date sequential patterns, and delete obsolete data and patterns accordingly. An extension of this approach we proposed in this paper, users can select the frequently repeated patterns by allowing the Dynamic Period of Interest(DPOI).Here main focus is on sliding window that changes dynamically, so that the pattern can be extract according to situation.

Keywords: Sequential pattern mining ,progressive database, dynamic period of interest

Introduction

The functionalities of data mining techniques include association rules mining, classification, clustering, mining time series, and sequential pattern mining, to name a few [2], [3], [4].Sequential pattern mining was first addressed in [1] as the problem: "Given a sequence database, where each sequence consists of a list of ordered item sets containing a set of different items, and a user defined minimum support threshold, sequential pattern mining is to find all subsequences whose occurrence frequencies are no less than the threshold from the set of sequences."

The sequential pattern mining with a *static database* finds the sequential patterns in the database in which data do not change over time[5]. On the other hand, the sequential pattern mining with an *incremental database* corresponds to the mining process where there are new data arriving as time goes by (i.e., the sequences database is incremental)[6]. As for the sequential pattern mining with a progressive database, new data are added into the database and obsolete data are removed simultaneously. Therefore, one can find the most up-to-date sequential patterns without being influenced by obsolete data.

The existing algorithms cannot cope with sequential pattern mining with a progressive database efficiently. To remedy this problem, Haung [7] proposed an efficient algorithm Pisa, which stands for Progressive mIning of Sequential pAtterns, corresponding to the mining in a progressive database.

Problem Description

POI is a sliding window, whose length is a user specified time interval, continuously advancing as the time goes by. The sequences having elements whose timestamps fall into this period, POI, contribute to the |Db| for current sequential patterns. On the other hand, the sequences having only elements with timestamps older than POI should be pruned away from the sequence database immediately and will not contribute to the| Db| thereafter.

PS-tree represents elements in the sequence, based on the sequence IDs and timestamps recorded in the nodes and the newly arriving data of the progressive database at each timestamp. PS-tree not only stores the elements and timestamps of sequences in each POI but also efficiently accumulates the occurrence frequency of every candidate sequential pattern at the same time.

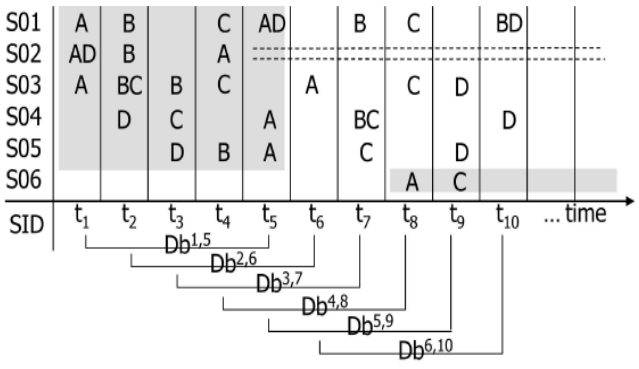


Fig1 sample database

Consider the progressive database in Fig. 1,for example. S01; S02; . . . ; Sn represent different sequence IDs. A, B, C, and D are different items in the database and t1; t2; . . . ; tk represent timestamps. As the time advances, there will be more elements arriving into the progressive database. Every sequence contains a series of elements appearing at different timestamps. Each element consists of a single or multiple items.

For instance, sequence S01 has element A at timestamp t1, element B at timestamp t2, element C at timestamp t4, and element . At the bottom of Fig. 1, Db represents a subset of the database containing the elements from timestamp p to timestamp q. Let the minimum support threshold, min sup, be 0.5 and the POI be five timestamps in this example. There are five sequences having elements in this period. Therefore, the minimum frequency for a frequent sequential pattern is $|Db^{1,5}|$* min sup = 5* 0:5= 2:5.

We can find a frequent sequential pattern AB, whose occurrence frequency is 3 (in S01, S02, and S03) in the first POI. However, after this POI, AB is no longer a frequent sequential pattern in any POI of five timestamps. PS-tree not only contains the information of all sequences in a progressive database but also helps Pisa to generate frequent sequential patterns in each POI. The nodes in PS-tree can be divided into two different types. They are root node and common nodes.They are root node and common nodes. Root node is the root of PS-tree containing nothing but a list of common nodes as its children. Each common node stores two information, say node label and a sequence list, The label is the same as the element in a sequence. The sequence list stores a list of sequence IDs to represent the sequences containing this element. Each sequence ID in the sequence list is marked by a corresponding timestamp

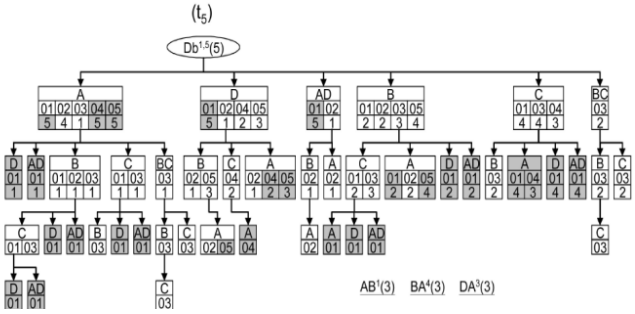


Fig2: PS-tree from t1~t5

If we require frequent patterns in between timestamp 2 and 4 then apply algorithm Dpisa as fig3

Let the minimum support threshold, min_sup be 0.5 and the in this example Dynamic POI is given by startTime **2** and endTime 4. There are three sequences having elements in this period. Therefore, the minimum frequency for a frequent sequential pattern is $|Db^{2,4}|$* min sup =3 * 0:5= 1.5. We can find frequent sequential pattern within that period as $BC^3(2)$ pattern as Fig4.

Algorithm **Dpisa (**support, POI)
Var PS ;//PS-tree
Var *currentTime, startTime, endTime*
Var eleset //used to store elements ele
While(there is new transaction)
eleset=read all elements at currentTime;
While(*endTime<=startTime*)
if (sequences are given) then *currentTime=startTime*;
Traverse(*currentTime*,PS)
currentTime++;
end.

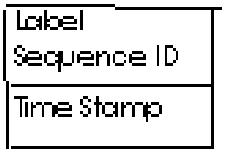


Fig3: Algorithm Dpisa

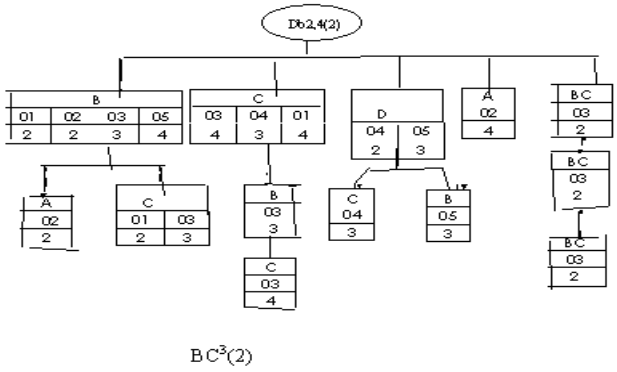


fig4:PS-tree from t2~t4

Conclusions

Our proposed work mine frequent pattern in progressive sequential databases seasonally in market basket analysis. Every time sales analysis is not uniform so user can extract the pattern s on his own interest. The major constraint is it consider only recent items. To achieve this we modified progressive sequential algorithm by using startTime and endTime as dynamic period of interest our goal is according to requirement user can extract patterns dynamically.

References

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, Sept. 1994.

[2] R. Agrawal and R. Srikant, "Mining Sequential Patterns," Proc. 11th Int'l Conf. Data Eng. (ICDE '95), pp. 3-14, Feb. 1995.

[3] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurasamy, Advances in Knowledge Discovery and Data Mining. MIT Press, 1996..

[4] J. Han and M. Kamber, Data Mining: Concepts and

Techniques. MorganKaufmann, 2000

[5] S. Aseervatham, A. Osmani, and *E. Viennet, "Bitspade: A Lattice-Based Sequential Pattern Mining Algorithm Using Bitmap Representation," Proc. Sixth Int'l Conf. Data Mining (ICDM), 2006. .*

[6] G. Chen, X. Wu, and X. Zhu, "Sequential Pattern Mining in Multiple Streams," Proc. Fifth Int'l Conf. Data Mining (ICDM '05), pp. 585-588, Nov. 2005.

[7] Jen Haung,IEEE transaction on knowledge and engineering,vol 20,no.9,sep2008," A General Model for Sequential Pattern Mining with a Progressive Database.

Analysis and Approach: K-Means and K-Medoids Data Mining Algorithms

Dr. Aishwarya Batra

Asst Professor, L. J. Institute of Computer Applications, Ahmedabad, India.
E–mail: batra.aishwarya@gmail.com

Abstract
Clustering is similar to classification in which data are grouped. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters .There exist a large number of clustering algorithms in the literature. The choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. Clustering analysis is one of the main analytical methods in data mining. K-means is the most popular and partition based clustering algorithm. But it is computationally expensive and the quality of resulting clusters heavily depends on the selection of initial centroid and the dimension of the data. Several methods have been proposed in the literature for improving performance of the k-means clustering algorithm. In this research, the most representative algorithms K-Means and K-Medoids were examined and analyzed based on their basic approach. The best algorithm in each category was found out based on their performance. The input data points are generated by two ways, one by using normal distribution and another by applying uniform distribution.

Keywords: K Means, K Medoid, Clustering, Partitional Algorithm

Introduction

Clustering techniques have a wide use and importance nowadays. This importance tends to increase as the amount of data grows and the processing power of the computers increases. Clustering applications are used extensively in various fields such as artificial intelligence, pattern recognition, economics, ecology, psychiatry and marketing. Data clustering is under vigorous development. Contributing areas of research include data mining, statistics, machine learning, spatial database technology, biology, and marketing. Owing to the huge amounts of data collected in databases, cluster analysis has recently become a highly active topic in data mining research. As a branch of statistics, cluster analysis has been extensively studied for many years, focusing mainly on distance-based cluster analysis.

The main purpose of clustering techniques is to partition a set of entities into different groups, called clusters. Cluster analysis tools based on k-means, k-medoids, and several other methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SA.

Categorization of Major Clustering Methods

In general, the major clustering methods can be classified into the following categories:

- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods

Classical Partitioning Methods: K-Means & K-Medoids

The most well-known and commonly used partitioning methods are *k-means*, *k-medoids*, and their variations. Partitional clustering techniques create a one-level partitioning of the data points. There are a number of such techniques, but we shall only describe two approaches in this section: K-means and K-medoid. Both these techniques are based on the idea that a centre point can represent a cluster. For K-means we use the notion of a centroid, which is the mean or median point of a group of points. Note that a centroid almost never corresponds to an actual data point. For K-medoid we use the notion of a medoid, which is the most representative (central) point of a group of points. Partitional techniques create a one-level (un-nested) partitioning of the data points. If *K* is the desired number of clusters, then partitional approaches typically find all *K* clusters at once.

Clustering

k-means (MacQueen'67): Each cluster is represented by the center of the cluster.

k-medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

Centroid-Based Technique: The K-Means Method

Basic Algorithm

The K-means clustering technique is very simple and we immediately begin with a description of the basic algorithm. We elaborate in the following sections.

Basic K-means Algorithm for finding K clusters

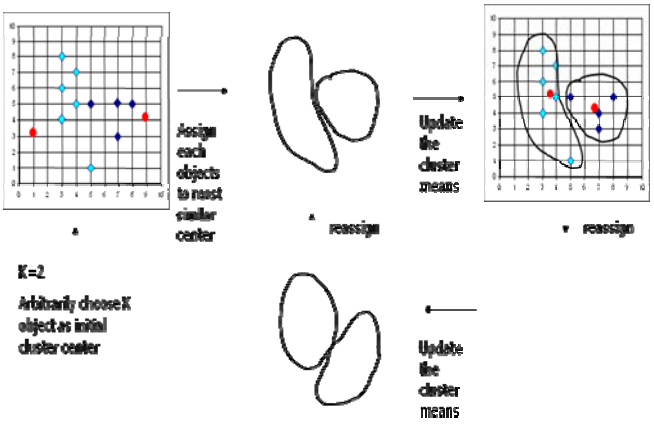
1. Select *K* points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.

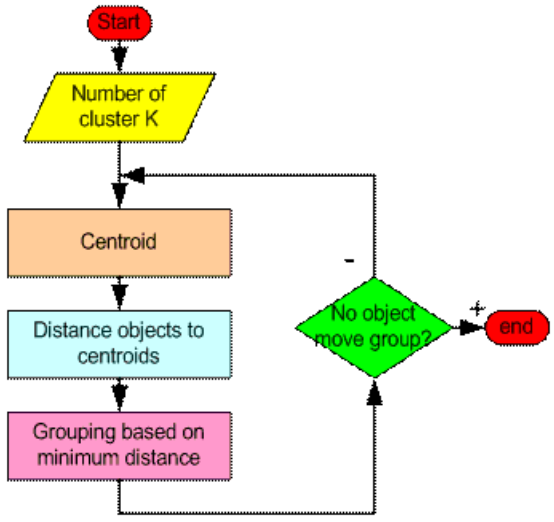
4. Repeat steps 2 and 3 until the centroids don't change.

The *k*-means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intracluster similarity is high but the intercluster similarity is low.

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} |p - m_i|^2$$

where E is the sum of the square error for all objects in the data set; p is the point in space representing a given object; and mi is the mean of cluster Ci (both p and mi are multidimensional). In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible.





Flow Chart of K-Means Algorithm

For example if we consider the following data set, K-Means Algorithm will work like this –

Point		(2,10)	(5, 8)	(1,2)	Cluster
		Dist Mean 1	Dist Mean 2	Dist Mean 3	
A1	(2,10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)	12	7	9	2
A4	(5, 8)	5	0	10	2
A5	(7, 5)	10	5	9	2
A6	(6, 4)	10	5	7	2
A7	(1, 2)	9	10	0	3
A8	(4, 9)	3	2	10	2

The initial cluster centres – means, are (2, 10), (5, 8) and (1, 2) - chosen randomly. Next, we will calculate the distance from the first point (2, 10) to each of the three means, by using the distance function:

Point mean1
x1, y1 x2, y2
(2, 10) (2, 10)

p (a, b) = |x2 – x1| + |y2 – y1|
p (point, mean1) = |x2 – x1| + |y2 – y1|
= |2 – 2| + |10 – 10|
= 0 + 0
= 0

point mean2
x1, y1 x2, y2
(2, 10) (5, 8)

p (a, b) = |x2 – x1| + |y2 – y1|
p (point, mean2) = |x2 – x1| + |y2 – y1|
= |5 – 2| + |8 – 10|
= 3 + 2
= 5

point mean3
x1, y1 x2, y2
(2, 10) (1, 2)

p (a, b) = |x2 – x1| + |y2 – y1|
p (point, mean2) = |x2 – x1| + |y2 – y1|
= |1 – 2| + |2 – 10|
= 1 + 8
= 9

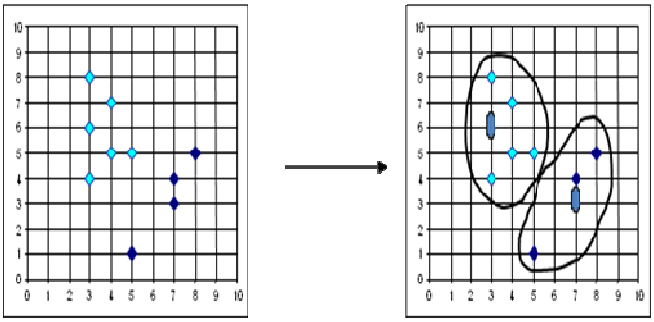
- Next, we need to re-compute the new cluster centers. We do so by taking the mean of all points in each cluster.
- The k-means algorithm is sensitive to outliers, since an object with an extremely large value may substantially distort the distribution of the data.
- In k-means algorithm (MacQueen, 1967), the prototype, called the center, is the mean value of all objects belonging to a cluster.
- Furthermore it requires several passes on the entire dataset, which can make it very expensive for large datasets as the dataset in our application.
- Often found in data relating to classified images.

- The k-medoids approach is more robust in this aspect.

K-Medoid Algorithm

K-medoid (The PAM-algorithm)(Kaufman and Rousseeuw, 1990),a partitioning around Medoids was one of the first k-Medoids algorithms introduced. It attempts to determine k partitions for n objects. After an initial random selection of k medoids, the algorithm repeatedly tries to make a better choice of medoids.

Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.



Distribution of Data Objects

The PAM-algorithm is based on the search for k representative objects or medoids among the objects of the dataset. These objects should represent the structure of the data. After finding a set of k medoids, k clusters are constructed by assigning each object to the nearest medoid. The goal is to find k representative objects which minimize the sum of the dissimilarities of the objects to their closest representative object.

Basic K-medoid Algorithm for finding K clusters.

1. Select K initial points. These points are the candidate medoids and are intended to be the most central points of their clusters.
2. Consider the effect of replacing one of the selected objects (medioids) with one of the non-selected objects.

Conceptually, this is done in the following way. The distance of each non-selected point from the closest candidate medoid is calculated, and this distance is summed over all points. This distance represents the "cost" of the current configuration. All possible swaps of a non-selected point for a selected one are considered, and the cost of each configuration is calculated.

$$E = \sum_{j=1}^{k} \sum_{p \in c_j} |p - o_j|$$

where E is the sum of the distances for all objects in the data set; p is the point in space representing a given object; and Oj is the centroid of cluster Ci (both p and Oj are multidimensional).

3. Select the configuration with the lowest cost. If this is a new configuration, then repeat step 2.

Numerical experiments

Cluster the given data set of ten objects into two clusters i.e. k = 2.

Consider a data set of ten objects as follows:

X_1	2	6
X_2	3	4
X_3	3	8
X_4	4	7
X_5	6	2
X_6	6	4
X_7	7	3
X_8	7	4
X_9	8	5
X_{10}	7	6

Data Set – Ten Objects

Step 1

Initialise k centre, Let us assume c1 = (3,4) and c2 = (7,4). Here c1 and c2 are selected as medoid.

Calculating distance so as to associate each data object to its nearest medoid. Cost is calculated using Minkowski distance metric.

C_1		Data Objects X_i		Cost (Distance)
3	4	2	6	3
3	4	3	8	4
3	4	4	7	4
3	4	6	2	5
3	4	6	4	3
3	4	7	3	5
3	4	8	5	6
3	4	7	6	6

C_1		Data Objects X_i		Cost (Distance)
3	4	2	6	3
3	4	3	8	4
3	4	4	7	4
3	4	6	2	5
3	4	6	4	3
3	4	7	3	5
3	4	8	5	6
3	4	7	6	6

Cost Calculation

Then the clusters become:

Cluster1 = {(3,4) (2,6) (3,8) (4,7)}

Cluster2 = {(7,4) (6,2) (6,4) (7,3) (8,5) (7,6)}

Since the points (2,6) (3,8) and (4,7) are closer to c1 hence they form one cluster whilst remaining points form another cluster.

So the total cost involved is 20.

Where cost between any two points is found using formula where x is any data object, c is the medoid, and d is the dimension of the object which in this case is 2.

Total cost is the summation of the cost of data object from its medoid in its cluster so here:

Total cost =

{ cost((3,4),(2,6)) + cost ((3,4),(3,8)) + cost ((3,4),(4,7)) }

+

{ cost ((7,4) , (6,2)) + cost ((7,4) , (6,4)) + cost ((7,4) , (7,3)) + cost ((7,4),(8,5)) + cost ((7,4) , (7,6)) }

= (3+4+4)+(3+1+1+2+2)

= 20

Cluster1 = {(3,4)(2,6)(3,8)(4,7)}

Cluster2 = {(7,4)(6,2)(6,4)(7,3)(8,5)(7,6)}

Step 2

Selection of nonmedoid O′ randomly. Let us assume O′ = (7,3), so now the medoids are c1(3,4) and O′(7,3)

If c1 and O′ are new medoids, calculate the total cost involved by using the formula in the step 1

Total cost = 3+4+4+2+2+1+3+3

= 22

So cost of swapping medoid from c2 to O′ is

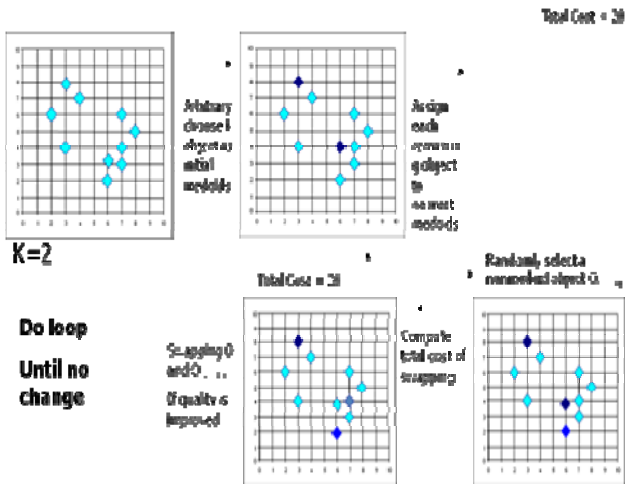
S = current total cost – past total cost

=22 - 20

=2 > 0

So moving to O′ would be bad idea, hence the previous choice was good and algorithm terminates here (i.e. there is no change in the medoids).

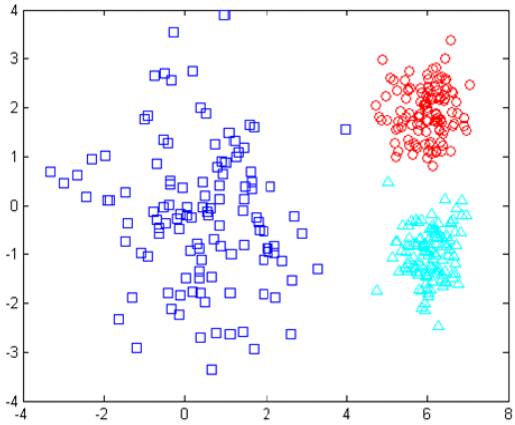
It may happen that some data points may shift from one cluster to another cluster depending upon their closeness to medoid.



Typical K-Medoids Algorithm (PAM)

Artificial Data for Comparison & Handling Outliers

In order to evaluate the performance of the clusters the artificial data will be generated and clustered using K-means clustering and PAM.We generate 120 objects having 2 variables for each of three classes shown in Fig. 1. We call the first group marked by square as class A, the second group marked by circle as class B and third group marked by triangle as class C for the sake of convenience.



Artificial Data for Comparison (Fig 1)

Data is generated from multivariate normal distribution, whose mean vector and variance of each variable (variance of each variable is assumed to be equal and covariance is zero) are given in Table 1. In order to compare the performance when some outliers are present among objects, we add outliers to the class B. The outliers are generated from a multivariate normal distribution which has the equal mean with class B but larger variance as shown in Table1.

Table1 - Mean and variance when generating objects

	Class A	Class B	Class C	Outliers (Class B)
Mean Vector	(0 , 0)	(6 , 2)	(6,-1)	(6 , 2)
Variance of each Variable	1.5^2	0.5^2	0.5^2	2^2

The adjusted Rand index will be used as the performance measure, which is proposed by Hubert and Arabie (1985) and is popularly used for comparison of clustering results. The adjusted Rand index is calculated as

$$RI_{ab} = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}$$

Where

a = number of pairs which are in the identical cluster of compared clustering solution for pairs of objects in certain cluster of correct clustering solution

b = number of pairs which are not in the identical cluster of compared clustering solution for pairs of objects in certain cluster of correct clustering solution

c = number of pairs which are not in the identical cluster of correct clustering solution for pairs of objects in certain cluster of compared clustering solution

d = number of pairs which are not in the identical cluster of both correct clustering solution and compared clustering solution.

Features of K-Medoid Algorithm:

It operates on the dissimilarity matrix of the given data set or when it is presented with an nxp data matrix, the algorithm first computes a dissimilarity matrix.

It is more robust, because it minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances

It provides a novel graphical display, the silhouette plot, which allows the user to select the optimal number of clusters.

However, PAM lacks in scalability for very large databases and it present high time and space complexity.

Unsupervised classification, or clustering, as it is more often referred as, is a data mining activity that aims to differentiate groups (classes or clusters) inside a given set of objects , being considered the most important unsupervised learning problem. The resulting subsets or groups, distinct and non-empty, are to be built so that the objects within each cluster are more closely related to one another than objects assigned to different clusters. Central to the clustering process is the notion of degree of similarity (or dissimilarity) between the objects.

Let O = fO1;O2;::: ;On be the set of objects to be clustered. The measure used for discriminating objects can be any metric or semi-metric function The distance expresses the dissimilarity between objects. The partitioning process is iterative and heuristic; it stops when a “good” partitioning is achieved. Finding a “good” partitioning coincides with optimizing a criterion function defined either locally (on a subset of the objects) or globally (defined over all of the objects, as in kmeans).

These algorithms try to minimize certain criteria (a squared error function in K-Means); the squared error criterion tends to work well with isolated and compact clusters .In k-medoids or PAM (Partitioning around medoids) algorithm, each cluster is represented by one of the objects in the cluster. It finds representative objects, called medoids, in clusters. The algorithm starts with k initial representative objects for the clusters (medoids), then iteratively recalculates the clusters (each object is assigned to the closest cluster - medoid), and their medoids until convergence is achieved. At a given step, a medoid of a cluster is replaced with a non-medoid if it improves the total distance of the resulting clustering.

It is understood that the average time for normal distribution is greater than the average time for the uniform distribution. This is true for both the algorithms K-Means and K-Medoids.

If the number of data points is less, then the K-Means algorithm takes lesser execution time. But when the data points are increased to maximum the K-Means algorithm takes maximum time and the K-Medoids algorithm performs reasonably better than the K-Means algorithm. The characteristic feature of this algorithm is that it requires the distance between every pair of objects only once and uses this distance at every stage of iteration.

Conclusion

Since measuring similarity between data objects is simpler than mapping data objects to data points in feature space, these pairwise similarity based clustering algorithms can greatly reduce the difficulty in developing clustering based pattern recognition applications. The advantage of the K means algorithm is its favourable execution time. Its drawback is that the user has to know in advance how many clusters are searched for. It is observed that K means algorithm is efficient for smaller data sets and K medoids seems to perform better for large datasets.

Bibliography

[1] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2006 (2nd Edition), http://www.cs.sfu.ca/~han/dmbook

[2] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Morgan, Kaufmann, 2005,http://www.cs.waikato.ac.nz/~ml/weka/book.html

[3] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining (First Edition), Addison Wesley, (May 2, 2005).

[4] Margaret H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2003,http://lyle.smu.edu/~mhd/book

[5] A K-means-like Algorithm for K-Medoids Clustering and Its Performance, Hae-Sang Park, Jong-Seok Lee and Chi-Hyuck Jun, Department of Industrial and Management Engineering, POSTECH, South Korea, shoo359@postech.ac.kr, jongseok@postech.ac.kr, chjun@postech.ac.kr

[6] Computational Complexity between K-Means and K-Medoids Clustering lgorithms for Normal and Uniform Distributions of Data Points T. Velmurugan and T. Santhanam ,Department of Computer Science, DG Vaishnav College, Chennai, India

[7] Best of Both: A Hybridized Centroid-Medoid Clustering Heuristic, NizarGriragrira@nii.ac.jp, Michael E.Houlemeh@nii.ac.jp, National Institute of Informatics, Japan

[8] K-Medoids: CUDA Implementation, Douglas Roberts, May 21, 2009

[9] International Journal of Database Management Systems (IJDMS), Vol.3, No.1, February 2011

[10] Top 10 algorithms in data mining, XindongWu · Vipin Kumar · J. Ross Quinlan

[11] Comparison between K-Means & K-Medoids clustering Algorithms, T SoniMadhuLatha, International Journal of Advanced Computing, April 2011.

[12] Clustering Parallel Data Streams, Yixin Chen,Department of Computer Science, Washington University. St. Louis, Missouri
[13] A Partitional Clustering Algorithm for Crosscutting Concerns Identification,Gabreilla Czibula Grigoreta, Sofia Cojocar, Istvan Gergely Czibula
[14] Data Mining and Soft Computing Research Group on Soft Computing and Information Intelligent Systems (SCI2S), Dept. of Computer Science and A.I., University of Granada, Spain
[15] Variance enhanced K-medoid clustering qPor-Shen Lai , Hsin-Chia Fu Department of Computer Science, National Chiao-Tung University, Hsin-Chu 300, Taiwan, ROC

Clustering Dynamic Class Coupling Data using K-Mean and Cosine Similarity Measure to Predict Class Reusability Pattern

Jitender Kumar Chhabra and Anshu Parashar

Department of Computer Engineering, National Institute of Technology, Kurukshetra, 136 119, India
E-mail: conf.ppr@gmail.com

Abstract
Data clustering techniques can be applied to cluster set of software components having similar characteristics like dependence to the same set of components and coupling with each other etc. This paper is an attempt to identify clusters of classes having dependence amongst each other and observe a common coupling pattern existing in the same repository. We explore both document clustering technique based on tf-idf weighing and cosine similarity measure to cluster classes from the collection of class coupling data for particular java application. For this purpose firstly dynamic analysis of java application is done using UML diagrams to collect class import coupling data. Then in second step, this coupling data of each class is represented in N-dimensional vector space. Further class coupling frequency and inverse class frequency is calculated using tf and idf. Then finally in the third step basic K-mean clustering and cosine similarity techniques are applied to find clusters of classes. Further each cluster is ranked for its goodness based on some user specified criteria and threshold. The proposed approach has been applied on simple Java application and our study indicates that by browsing such clusters a developer can discover class's reusability patterns and behaviour.

Keywords: Coupling, Data Clustering, Software Reusability.

Introduction

Identification of reusable components during the process of software development is an essential activity it helps to develop, identify and store reusable components [4]. Software Reuse is defined as the process of building or assembling software applications from previously developed software [23] and to increase productivity, quality, maintainability etc [20,5]. Object oriented development [15] offers many features above the traditional development approaches to improve software reusability through encapsulation and inheritance [17]. In object-oriented paradigm, coupling between classes is well-recognized structural attribute and plays a vital role in measuring the reusability. One can define a class C_a related to class C_b if C_a must use C_b in all future reuse. These highly coupled groups of classes should be reused together for ensuring the proper functioning of the application [8, 24].Hence it is always desirable to find out the classes along with their associated classes [15]. So to discover the clusters of classes that should be reused in combination, clustering approach can be used. By using clustering, one can find frequently used classes in the same cluster and can know their coupling behaviour in a particular application.

Data Clustering & Reusability

Data mining is the process of extracting new and useful knowledge from large amount of data. Mining is widely used to solve many business problems such as customer behavior modeling, product recommendation, fraud detection etc [27]. Data mining techniques can be used to analyze software engineering data to better understand the software and assist software engineering tasks. Clustering plays an important role in mining task like in data analysis and information retrieval [12] .Clustering means to clusters, where a *cluster* is a collection of objects, which are similar and closer to each other and dissimilar/distant objects belong to different clusters. Clustering can be applied to cluster the documents in the process of information retrieval. The Vector Space Model (VSM) is the basic model for document clustering. In this model, each document, d_j, can be represented as a term-frequency vector in the term-space:

$$d_{jtf} = (tf_{1j}, tf_{2j}, \ldots . tf_{vj}) \quad j=1,2,\ldots D$$

where tf_{ij} is the frequency of the i_{th} term in document d_j , V is the total number of the selected vocabulary, and D is the total number of documents in the collection [25]. One can weight each term based on its Inverse Document Frequency (idf) [25,2]. After having VSM representation, K-mean algorithm can be applied to cluster the documents. Basic K-means Algorithm for finding K clusters involves following steps [22]:

1. Select *K* points as the initial centroids.
2. Assign points to their closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids don't change.

Clustering technique can be applied to cluster the Classes/components that may often be reused in combinations [27]. Due to the popularity of open source concept large amount of source code of classes is available on internet as software repositories. Some also exists in large software companies where developer in one group may reuse classes written by other groups. For this reason, it is desirable to have clustering mechanism that forms clusters of classes based on their association or coupling patterns/behaviour. By searching for class patterns with high probability of repetitions

developer can correlate one set of classes with other set of classes. Also he will know that classes belonging to same cluster are likely to be reused together.

In this paper, we explore N-dimensional vector space model and document clustering technique based on tf-idf weighing scheme [2] to cluster classes from the collection of class import coupling data for particular java project/program. By browsing such clusters a developer can discover patterns for reusing classes and predict their coupling behaviour. For this purpose firstly dynamic analysis of java application is done using UML diagrams to collect class import coupling data. Then in second step, these collected coupling data of each class are represented as N-VSM and treated as document (using tf and idf). Then finally in the third step basic K-mean clustering technique and cosine similarity measure are applied to find cluster of classes. Further each cluster is ranked for their goodness based on some user specified criteria and threshold.

The rest of the paper is organized as follows. Section 2 discusses the related works. Section 3 describes the proposed methodology. Section 4 shows example to illustrate our approach. Section 5 presents results and discussions. Finally, Section 6 concludes this paper.

Related Works

For object-oriented development paradigm, class coupling has been used as an important parameter effecting reusability. Efforts have been made by the researchers to measure reusability through coupling and cohesion of components [18]. ISA [19] methodology has been proposed to identify data cohesive subsystems. Gui et al [10] proposed a new static measure of coupling to assess and rank the reusability of java components. Arisholm et al[3] have provided a method for identifying import coupled classes with each class at design time using UML diagrams. Data mining is focused on developing efficient techniques to extract relevant information from very large volumes of data that may be exploited, for example, in decision making, to improve software reliability and productivity [26]. Cluster formation from large databases is an important data mining task. Few algorithms like CLARANS [21], BIRCH [29] has been proposed for clustering large data sets. Li et al [16] devised a set-theoretic clustering method called PCS (Pair-wise Consensus Scheme) for high-dimensional data. Abrantesy et al [1] described a method for the segmentation of dynamic data. Document clustering has been an interesting topic of study since a long time. Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [13] is one of the best algorithms for agglomerative clustering [30]. K-Mean's and its family of algorithms have also been extensively used in document clustering[14]. Kiran et al [14] proposed a hierarchical clustering algorithm using closed frequent item sets that use Wikipedia as an external knowledge to enhance the document representation. Fung et al [9] proposed to use the notion of frequent itemsets and applied TF-IDF to define the score function Score ($C_i \leftarrow doc_j$) for measuring the goodness of a cluster C_i for a document doc_j. Alzghool et al [2] also proposed a technique based on clustering the training topics according to their tf-idf (term frequency-inverse document frequency) properties. Some researchers have been using clustering in software development tasks. Czibula et al [7] introduced a search based approach for identifying instances of design patterns in a software system. For Evaluation of cluster quality authors in [28] proposed cluster ranking to quickly single out the most significant clusters. Also to measure goodness and quality of cluster Entropy, F-measure has been used[22]. There are some distance measures available in literature like Absolute distance, Euclidean distance and cosine similarity/distance [31,32,33,6]. We find the basic K-mean algorithm and cosine similarity measure very simple and go well with our initial idea.

Proposed Methodology

In our approach we propose to cluster class import coupling data for a particular java application. To collect class import coupling data dynamic analysis of java application is done using UML. Then collected import coupling data of each class is represented using N- VSM and tf-idf. Then basic K-mean clustering technique and cosine similarity measures are applied to find cluster of classes. Our approach consists of three steps:

1. Collection of Class import coupling data through UML.
2. Representation of Collected Data.
3. Clustering of class import coupling data

The steps are described in section 3.1 to 3.3.

Collection of Class Import Coupling Data through UML

Dynamic analysis of a program is one way of finding the coupling between classes. During this step, the existing application is analyzed through UML diagrams [11] as described by Erik Arisholm [3] in order to extract import coupling of its classes. They used following formula for calculating class import coupling IC_OC (C_i) to measure dependency of one class to other classes.

$$IC\_OC(c_1) = \{(m_1, c_1, c_2) \mid (\forall (o_1, c_1) \in R_{oc})(\exists (o_1, c_2) \in R_{oc} \mid \in N) c_1 \neq c_2 \wedge (o_1, m_1 \mid o_2, m_2) \in ME\}$$

IC_OC (C_i) counts the number of distinct classes that a method in a given object uses.

Representation of Collected Data using N-dimensional import coupling vector and tf-idf weighing scheme

Data collected in step 1 is then represented as N-dimensional import coupling vector and 2-D import coupling vector using tf-idf weighing scheme. For an application A, the class set of A is represented as Class_Set(A)={ $C_1,C_2,C_3 \ldots .C_n$} where n is total number of classes in an application A. Each class C_i is represented as N-dimensional import coupling vector $NIC\_V(C_i)$ where N is the total number of classes in an application *A.* So for each class C_i this vector is represented as $NIC\_V(C_i)=[x_{i1}, x_{i2}, \ldots x_{iN}]$. The value x_{i1} (also called as cf_{1i}) of class C_i represents import coupling usage frequency of class C_1 in class C_i. Next inverse class frequency (ICF) weighing is used to weight each class C_i based on idf. So we calculate ICF of each class using idf formula 1:

$$ICF(C_i)=\log\left(\frac{n}{ICoupF(C_i)}\right) \quad (1)$$

Where n is total number of classes, $ICoupF(C_i)$ is number of classes using C_i . Then finally import coupling $ICoup(C_i,C_j)$ of class C_i with C_j is represented as 2D-point (cf_{ji} , $ICoupF(C_i)* cf_{ji}$).

Clustering of class import coupling data

Clustering using K-Mean approach

In this step clustering technique is to be applied on 2D representation of each pair of classes $ICoup(C_i,C_j)$. Each cluster will have set of classes that can be reused together. So, K-mean algorithm as described in section 1.1 is used to find out such clusters. Here K is the pre assumed required number of clusters and value of K is decided by the user. Once we decide on what are the K clusters and their initial centroids then K-mean algorithm starts as per section 1.1. The absolute distance function (formula 2) is used to measure the closeness.

$$d_A(x,y)=\sum_{i=1}^{n}|x_i-y_i| \quad (2)$$

In every iteration centroids are recalculated. Each class pair (C_i, C_j) is assigned to the cluster with the nearest centroid point. By iterating K-mean algorithm (until there is no movement of points), it will discover clusters of the form e.g {(C_1, C_3), (C_2, C_3)}. Their union {C_1, C_2, C_3} will form the final cluster. One can interpret this as classes in a cluster are coupled with each other and will be reused together.

After having K clusters each cluster is ranked by taking average and sum of its *x,y* points , e.g. Cluster I ={(a,b), (c,d)} then RankC(Cluster I)= (a+c/2)+(b+d/2). This RankC(k) should pass the threshold th_c specified by the user. Threshold is the lowest possible permissible rank that will be used to classify cluster as good or bad. If RankC(k) < th_c then cluster k is discarded otherwise k is retained.

Clustering Using Cosine Similarity

The N-Dimensional Class vector representation $NIC\_V(C_i)$ of classes is used to compute cosine similarity between classes on a scale of [0, 1]. The Cosine Similarity [7] of two vectors NIC_V (C_i) & NIC_V (C_j) is defined as:

$$Cos\_Sim(C_i,C_j)=\frac{\mathbf{C_i}.\mathbf{C}_j}{\|\mathbf{C_i}\|\|\mathbf{C}_j\|} \quad (3)$$

Where C_i . C_j = $C_i[0]$* C_j [0] + $C_i[1]$* $C_i[1]$ ... and $\| C_i\|$ = sqrt($C_i[0]$^2 + $C_i[1]$^2 ...)

The $Cos\_Sim(C_1,C_2)$ =1 can be interpreted as C_1 & C_2 are coupled with exactly same set of classes. $Cos\_Sim(C_1,C_2)$ =0 can be interpreted as coupling set of C_1 & C_2 do not have any common class. So for each k-cluster we have to decide its permissible similarity scale *Sim_Scale* . If any class pair $Cos\_Sim(C_1,C_2)$ satisfy this *Sim_Scale* then (C_1,C_2) is included in cluster k.

This similarity scale *Sim_Scale* of each cluster also reveals its rank. So after computing $Cos\_Sim(C_i,C_j)$ for each pair of classes we place that pair in nearby cluster as per its cosine similarity value. In next section, we demonstrate our methodology of clustering class import coupling data using K-Mean and Cosine Similarity approaches.

Example

We are using a small example to illustrate our approach for clustering of class import coupling data to have K- clusters of classes. Let application ***A*** having classes Class_Set(A)={ **C_1 , C_2 , C_3, C_4** }. As per the first step we assume to have import coupling data collected for application ***A*** using UML approach. The next sections 4.1 & 4.2 show the second and third step of our approach.

Representation of Collected Data

In second step the collected coupling data is represented as N-dimensional Class Import Coupling Vector $NIC\_V(C_i)$ of all classes in an application A as shown in table 1. Then Inverse Class Frequency (ICF) of each class is calculated using formula 1 and represented in table 2.

Table 1. N-dimensional Class import coupling Vector.

	x_1	x_2	x_3	x_4
$NIC\_V(C_1)$	0	0	0	3
$NIC\_V(C_2)$	0	0	0	0
$NIC\_V(C_3)$	6	1	0	0
$NIC\_V(C_4)$	1	0	0	0

Table 2. ICF value of each class

Class	$ICF(C_i)$
C_1	.30
C_2	.60
C_3	0
C_4	.60

Table 3. Point representation and K-mean clustering iterations

$ICoup(C_i, C_j)$	Point		Iteration I			Cluster	Iteration II			Cluster
			(0,0)	(1,.60)	(6,1.8)		(0,0)	(1,.45)	(4.5,1.8)	
	Cf_{ij}	$ICF(C_j)* cf_{ij}$	Dist 1	Dist 2	Dist 3		Dist 1	Dist 2	Dist3	
(C_1,C_2)	0	0	0	1.60	7.8	I	0	1.45	6.3	I
(C_1,C_3)	0	0	0	1.60	7.8	I	0	1.45	6.3	I
(C_1,C_4)	3	1.8	3.8	3.2	3.0	III	3.8	3.35	1.5	III
(C_2,C_1)	0	0	0	1.60	7.8	I	0	1.45	6.3	I
(C_2,C_3)	0	0	0	1.60	7.8	I	0	1.45	6.3	I
(C_2,C_4)	0	0	0	1.60	7.8	I	0	1.45	6.3	I
(C_3,C_1)	6	1.8	7.8	6.2	0	III	7.8	6.35	1.5	III
(C_3,C_2)	1	.60	1.6	0	6.2	II	1.6	.15	4.7	II
(C_3,C_4)	0	0	0	1.60	7.8	I	0	1.45	6.3	I
(C_4,C_1)	1	.30	1.3	.30	6.5	II	1.3	.15	5	II
(C_4,C_2)	0	0	0	1.60	7.8	I	0	1.45	6.3	I
(C_4,C_3)	0	0	0	1.60	7.8	I	0	1.45	6.3	I

Table 6. NIV_V(C_i) and Cluster formation using *Cos_Sim(C_i,C_j)*

N-Dimensional import coupling representation					**Cosine similarity values & assigned clusters**		
NIC_V (C_i)	***x_1***	***x_2***	***x_3***	***x_4***	***Cos_Sim(C_i,C_j)***	***Sim_Scale***	***Assigned Cluster***
NIC_V (C_1)	0	0	0	3	Cos_Sim(C_1,C_2)	0	**I**
NIC_V (C_2)	0	0	0	0	Cos_Sim(C_1,C_3)	0	**I**
NIC_V (C_3)	6	1	0	0	Cos_Sim(C_1,C_4)	0	**I**
NIC_V (C_4)	1	0	0	0	Cos_Sim(C_2,C_3)	0	**I**
					Cos_Sim(C_2,C_4)	0	**I**
					Cos_Sim(C_3,C_4)	.98	**III**

After this import coupling ICoup(C_i,C_j) between all classes is represent as 2D-points (cf_{ij} , $ICF(C_j)$* cf_{ij}) as shown in table 3.

Clustering of Class Import Coupling Data Using K-Mean Approach

In this step K-mean algorithm is applied on the import coupling data represented as points in the table 3.We assume to have 3 clusters as output of K-mean algorithm (K=3) and th_c = 3. The initial cluster-means (centroid) of three clusters are (0, 0), (1,.60) and (6,1.8), chosen randomly. Next, in Iteration I the distance of all the points to each of the three centroids are calculated by using the distance function: *dist(a, b) = |x2 – x1| + |y2 – y1| where a=(x1,y1) and b=(x2,y2).* Then each point (C_i,C_j) is placed in its nearest distanced cluster . Then, in Iteration II re-compute the new cluster centers (centroids) by taking the mean of the points in each cluster. So new centroids become (0, 0), (1,.45) and (4.5,1.8) for cluster I,II and III respectively. Then again the distance from all points to each three centroids are calculated and each point (C_i,C_j) assigned to its nearest cluster. After iteration II we found that no point movement is there i.e. all points remain in their previously assigned clusters. It means, these are the final three clusters as shown in table 3. After obtaining three clusters I, II and III, their ranks are calculated and compared with th_c =3 as shown in below table 4.

Cosine Similarity Approach

The N-Dimensional import coupling vector (table 1) is used to find coupling behaviour of all classes. For each K cluster we decide permissible *Sim_Scale*. Clusters I will have class pairs having *Sim_Scale* < .20 , cluster II will have class pairs having *Sim_Scale* < .60 and cluster III will be have class pairs having *Sim_Scale* < .80 as shown in table 5.

Results & Discussion

After applying the clustering step using K-Mean as in section 4 we obtain these clusters I, II and III as shown bellow in figure 1.

Cluster I={ (C_1,C_2) , (C_1,C_3) , (C_2,C_1) , (C_2,C_3) , (C_2,C_4), (C_3,C_4), (C_4,C_2) , (C_4,C_3) **}**
Cluster II= {(C_3, C_2), (C_4, C_1)**}**
Cluster III= {(C_1, C_4), (C_3, C_1)**}**

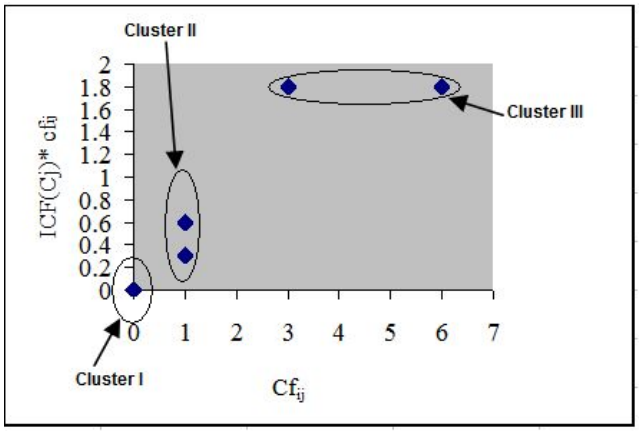


Figure 1: Final Clusters Formation using K-mean

Table 4. Rank of Clusters formed by K-mean and status

Cluster	***Sim_Scale***	**Cluster Rank**
I	< .30	Class pair having least import coupling similarity
II	< .60	Classes pair having some import coupling similarity
III	<.80	Classes pair having significant import coupling similarity

Table 5. Cluster ranks based on *Sim_Scale*

Cluster	**RankC(k)**	**Comparison with th_c=3**	**Cluster status**
I	0	< th_c	Discarded
II	1.45	< th_c	Discarded
III	6.5	> th_c	**Retained**

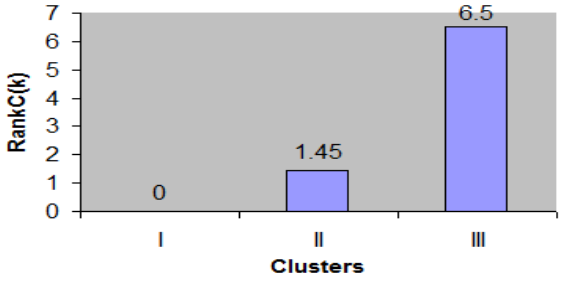


Figure 2. Clusters formed by K-mean and their Rank

Then *Cos_Sim(C_i ,C_j)* between two classes are measured by cosine similarity formula (formula 3) and pair (C_i ,C_j) is assigned to appropriate cluster according to table 5. The

bellow mentioned table 6 shows NIV_V(C_i) and cluster formation using cosine similarity approach. 3 will be have class pairs having *Sim_Scale* < .80 as shown in table 5.

Table 4 and Figure 2 show the rank of these clusters and it has been observed that cluster I and cluster II are not satisfying the required threshold th_c=3. Only cluster III is retained. So the union of cluster III {C_1, C_3, C_4,} will form the final cluster. We interpret this as classes in a cluster are coupled with each other and will be reused together. Ranking of cluster III suggests that these classes are highly used in application *A* and mostly used together.

Further cluster formed by cosine similarity approach in section 4.2.2 are shown below:

Cluster I={ (C_1,C_2) , (C_1,C_3) , (C_1,C_4) , (C_2,C_3), (C_2,C_4) }
Cluster II= { }
Cluster III= {(C_3, C_4)}

For an application *A* Cluster I suggests that import coupling set of classes C_1,C_2 and C_3 are not having any common class. Their import coupling behavior is entirely different. Cluster III suggested that class C_3 and C_4 have some common classes in their import coupling set and their import coupling behavior is significantly similar. So developer can measure the coupling behavior of classes to predict its reusability pattern by browsing these clusters.

Conclusions

In this paper, an attempt has been made to determine class reusability pattern and behavior from dynamically collected class import coupling data of java application. We have explored the idea of document clustering (using tf-idf weighting scheme) and N-dimensional Vector space to represent the coupling between two classes. Our initial study indicates that basic technique of K-mean clustering can be constructive to place of most frequently reusable classes together in a same cluster. It means classes in a cluster are coupled with each other and will be reused together. Further cluster formation using cosine similarity measure is also helpful to know which classes have similar/different coupling behaviour and also to know whether classes are coupled with some common classes or not. So reuse issues like deciding what group of classes should be incorporated into repository and identifying exact set of classes to reuse, can be addressed through these clustering mechanism. Currently, we have applied our approach on a simple example .However the approach can also be applied on larger java applications. Moreover, other mining and clustering algorithms can be

References

[1] Abrantesy, A. J., Marquesz J.S., "A Method for Dynamic Clustering of Data", In Proceedings of the British Machine Vision Conference ,pp. 154-163,1998

[2] Alzghool , M., Inkpen,D. "Clustering the Topics using TF-IDF for Model Fusion", In ACM **Proceeding of the 2nd PhD workshop on Information and knowledge management,** pp. 97--100,2008

[3] Arisholm, E., "Dynamic Coupling Measurement for Object-Oriented Software", In IEEE Transactions on Software Engineering, vol. 30, no. 8, pp 491—506,2004

[4] Bhatia, P., K.,Mann, R. "An Approach to Measure Software Reusability of OO Design", In Proceedings of the 2nd National Conference on Challenges & Opportunities in Information Technology,pp 26—30 ,2008

[5] Caldiera, G., Basili, V. R.," Identifying and Qualifying Reusable Software Components", In IEEE Journal of Computer, vol. 24 , No.2, pp 61—70,1991

[6] Cosine Similarity, http://en.wikipedia.org/wiki/Cosine_similarity andhttp://www.appliedsoftwaredesign.com/cosineSimilarity Calculator.php

[7] Czibula, I.G., Serban, G., "Hierarchical Clustering Based Design Patterns Identification", In Int. J. of Computers, Communications & Control, vol. 3, pp. 248—252, 2008.

[8] Eickhoff, F. Ellis, J., Demurjian, S., Needham, D., "A Reuse Definition, Assessment, and Analysis Framework for UML", In International Conference on Software Engineering, http://www.engr.uconn.edu/~steve/Cse298300/ eickhofficse2003 submit.pdf, 2003

[9] Fung, B.C.M ., Wang, K., Esterz, M. "Hierarchical Document Clustering Using Frequent Itemsets", In Proceedings of the third SIAM International Conference on Data Mining, 2003

[10] Gui, G., Scott, P. D., "Ranking reusability of software components using coupling metrics", In Elsevier Journal of Systems and Software, vol.80, pp. 1450 -- 1459. 2007

[11] Gupta, V., Chhabra, J. K. "Measurement of Dynamic Metrics Using Dynamic Analysis of Programs", In Proceedings of the Applied Computing Conference, pp. 81—86,2008

[12] Henry, S., Lattanzi, M. "Measurement of Software Maintainability and Reusability in the Object Oriented Paradigm", In ACM Technical Report, 1994

[13] Kaufman, L., Rousseeuw, P.J., "Finding Groups in Data, An introduction to Cluster Analysis ", John Wiley & Sons, Inc ,1990

[14] Kiran, G. V. R. , Shankar, K.R., Pudi, V. ,"Frequent Itemset based Hierarchical Document Clustering using Wikipedia as External Knowledge", In Proceeding of Intl Conference on Knowledge-Based and Intelligent Information Engineering Systems,2010 http://www.stanford.edu/class/cs229/ projects2010. html.

[15] Lee, Y., Chang, K.H., "Reusability and. Maintainability Metrics for Object-Oriented Software", In ACM 38th annual Southeast regional conference, pp. 88—94,2000

[16] Li, W., Chen, C., Wang, J., "PCS: An Efficient Clustering Method for High-Dimensional Data", In Proceedings of the 4th International Conference on Data Mining (DMIN'08). July 14-17, 2008.

[17] Michail, A. "Data Mining Library Reuse Patterns in User-Selected Applications", In 14th IEEE

International Conference on Automated Software Engineering, pp. 24--33. 1999
[18] Mitchell, A., Power, F. "Using Object Level Run Time Metrics to Study Coupling Between Objects", In ACM Symposium on Applied Computing, pp 1456 -- 1462 ,2005
[19] Montes C., Carver, D. L., "Identification of Data Cohesive Subsystems Using Data Mining Techniques", In IEEE International Conference on Software Maintenance, pp.16 --23.1998
[20] Negandhi, G., "Apriori Algorithm Review for Finals", www.cs.sjsu.edu
[21] Ng,R.T., Han, J., "Efficient and effective clustering methods for spatial data mining". In Proceeding of VLDB conference, pp. 144-155,1994
[22] Rao, I.K.R., "Data Mining and Clustering Techniques", In Proceeding of DRTC Workshop on Semantic Web,2003
[23] Shiva, S. J., Shala, L., A. "Software Reuse: Research and Practice", In Proceedings of the IEEE International Conference on Information Technology, pp. 603--609. 2007
[24] Taha, W., Crosby, S., Swadi, K., "A New Approach to Data Mining for Software Design", In 3rd International Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications, 2004
[25] Xiao, Y. "A Survey of Document Clustering Techniques & Comparison of LDA and moVMF", In CS 229 Machine Learning Final Projects,2010
[26] Xie, T., Acharya, M., Thummalapenta, S., Taneja, K., "Improving Software Reliability and Productivity via Mining Program Source Code", In IEEE International Symposium on Parallel and Distributed Processing, pp 1--5. 2008
[27] Xie, T., Pei, J. "Data mining for Software Engineering", http://ase.csc.ncsu.edu/dmse/ dmse. pdf
[28] Yossef, Z.B.,Guy,I., "Cluster Ranking with an Application to Mining Mailbox Networks", In ACM Proceedings of the Sixth International Conference on Data Mining,2006.
[29] Zhang,T., Ramakrishnan, R., and Birch,L. M., "An efficient data clustering method for very large data-bases", In ACM SIGMOD, pp. 103-114,1996
[30] Zhao, Y., Karypis, G. "Evaluation of Hierarchial Clustering Algorithms for Document Datasets", In Proc. of Intl. Conf. on Information and Knowledge Management,2002
[31] http://en.wikipedia.org/wiki/Distance
[32] http://en.wikipedia.org/wiki/Euclidean_distance
[33] http://en.wikipedia.org/wiki/Metric_(mathematics)

A Proposed Quartile Clustering Algorithm to Detect Outliers for Large Data Sets

Mamta Malik[1], Dr. A.K. Sharma[2] and Dr. Parvinder Singh[3]

[1, 3] DCRUST, Murthal, Sonepat, India
[2]Professor &Dean, YMCA University of Science & Technology, Faridabad, India
E-mail: 1malik.mamta@gmail.com

Abstract
The applications of spatial database are more and more popular in the computer technologies. It is the reason why information retrieving becomes is an important issue. So clustering is an important task in spatial data mining and spatial analysis. Knowledge Data Discovery (KDD) schemes may include way to find solution of many unanswered questions. The purpose of this paper is to proposed novel quartile clustering algorithm to detect outliers to enhance the knowledge discovery of spatial database. This algorithm is more appropriate to define clusters by detecting outliers in large scale spatial datasets. Combining the concepts of polygon reduction and quartile, a new spatial database clustering algorithm is proposed. We measure the compactness of the clusters produced using quartile clustering algorithm and compare it with the clusters formed using DBSCAN algorithm.

Index Terms: Spatial database Mining; Quartile; Polygon Reduction

Introduction

Currently, most of geospatial data, both spatial and non-spatial properties, are managed by geospatial databases. With the growing need of integration and sharing of geospatial data located in different places on the network, distributed geospatial database technology has become a hot research field. The research topics on distributed geospatial database typically include global spatial data catalogue, global spatial indexing, global query processing and optimization, transaction management [19] etc. These objects are similar to each other in the same cluster, and are different from the objects in other clusters. At present, there are a large number of clustering algorithms, such as partitioning method, density-based method, hierarchical method, grid-based method, model method and fuzzy-based method. Also, there are new ones which are properly combined by the methods mentioned above.

However, with the development of information technology and the appearance of Internet and mass database, present clustering methods cannot adapt to the clustering of large-scale data. According to the characteristics of large-scale data, a new method is pointed out in this paper.

So far, the traditional cluster algorithms which can manage spatial constraint include COD-CLARANS [1], DBCluC [2], AutoClust+ [3], and DBRS+ [4] and so on; each algorithm has its own superiority and drawback. Based on CLARANS, COD-CLARANS algorithm was the first algorithm to solve the problem of clustering in the presence of physical obstacles. It defines obstacles by building visibility graphs to find the shortest distance among data objects in the presence of obstacles. The visibility graph is expensive to build.

The algorithm does not consider facilitator that connects data objects. A simple modification of the distance function in COD_CLARANS is inadequate to handle facilitators because the model used in pre-processing for determining visibility and building the spatial join index would need to be significantly changed. AutoClust+ builds a Delaunay structure to cluster data points considering obstacles, it is expensive to construct and is not flexible to combine a different kind of constraints. Since the points connecting by facilitators usually does not share a boundary in Voronoi regions. Based on DBSCAN, DBCLuC is the first approach that handles both obstacles and facilitators. It determines the visibility through obstruction lines [5]. To allow for facilitators, entry points and entry edges are identified. The lengths of facilitators are ignored. But reach-ability between any two points is defined by not intersecting with any obstruction lines, the algorithm will not work correctly when the shortest distance between them is actually less than the radius. DBRS+ is a modified version of DBRS adapted for clustering in the presence of obstacles and facilitators. It proposed “Chop and Conquer” to handle obstacle. For facilitator, it uses special polygon with access point which are classified three types: entrance, Exit, primary entrance. But sometimes it is not proper to use polygon to model the obstacle when the obstacle is a river or a road.

In the paper we propose concept of quartile clustering to detect outliers, which has a new idea to model a spatial cluster. The new algorithm quartile clustering to detect outliers using the concept of convex and concave polygon by redefining the concepts of the neighbourhood of a point, seed point, outliers, and noise points. The clustering is done based on the distance between two data points leading to the rest of each other being clustered together, and thus resulting in spatially compact clusters. Note that a key component of our quartile clustering algorithm is the calculation of the distance function between data sets and using first, second and third quartile.

The rest of the paper is organized as follows. Section 2 presents the related work giving a background on spatial

clustering and density-based spatial clustering. Section 3 defines the polygons, our methodology for computing the distance between two data points, finds it's neighboured and explains our algorithm in detail. Section 4 presents our quartile clustering algorithm with experimental results. Section 5 explains complexity of algorithm. Finally, our conclusion is given in Section 6.

Related Work

Spatial Clustering Algorithms

Clustering algorithms can be categorized into five main types: Partitional, Hierarchical, Density-based, Grid-based, and Model-based clustering algorithms. In Partitional algorithms, partitions of a database D are developed, and a set of clusters are formed. The number of clusters generated has to be specified in advance. The cluster similarity is measured with respect to the mean value (cluster centre) of the objects in a cluster. Examples are PAM [6], CLARA [6], and CLARANS [6].

Hierarchical algorithms create a hierarchical decomposition of the database. This hierarchical decomposition is represented as a dendrogram. Each level of the dendrogram represents a set of clusters. Thus, a set of nested clusters organized as a hierarchical tree are produced. As a result the initial knowledge of the number of clusters is no longer required. However, a termination condition needs to be specified. Examples of hierarchical clustering are CURE [7] and BIRCH [8].

Density-based clustering algorithms are based on the idea that objects which form a dense region should be grouped together into one cluster. These algorithms search for regions of high density in a feature space that are separated by regions of lower density. Thus, density-based methods can be used to filter out noise, and discover clusters of arbitrary shape. Examples of density-based clustering algorithms are DBSCAN [9], DENCLUE [10], and OPTICS [11].

Grid-based algorithms are based on multiple level grid structure. The entire space is quantized into a finite number of cells on which operations for clustering are performed. Summarized information about the area covered by each cell is stored as an attribute of the cell. The main advantage of this approach is its fast processing time. However, the summarized information leads to loss of information. Examples of grid-based clustering algorithms are STING [12], Wave Cluster [13], and CLIQUE [14].

In model-based algorithms a model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each cluster. They are often based on the assumption that the data are generated by a mixture of underlying probability distributions. COB-WEB [15] is an example of this approach. We select the density-based approach for clustering polygons since there is no need to know the number of clusters in advance as required in Partitional algorithms, nor is there a need to store summarized information as in grid-based algorithms.

Moreover, polygons in geographic space and in many other domains naturally respond to the density-based approach. For example, in geographic space, we have a set of contiguous polygons, and another set of polygons located far away from the first set. At a larger scale, these two sets will belong to a cluster each, thus corresponding to clusters formed where the object density is high.

Polygon Reduction

In any clustering algorithms, when obstacles are considered, the visibility of data objects with each other is checked via the line segments or edges of the obstacle. The number of line segments to check is the number of edges of the polygons, which is large in number for a large data space. The number of lines to check can be reduced to actual one by our polygon-edge reduction method but memory would be much less then the matrix approach contains reduction lines. We are here to going to use set of reduction lines. Let us call the reduced number of lines as reduction lines. The algorithm assumes various definition of a polygon will be discussed in detail.

Concept of Polygon Reduction

Before explaining the concept of polygon reduction let we discuss some definitions:

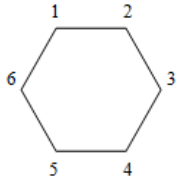
Definition: Polygon

A simple polygon is denoted by an undirected graph P(V,E) where V is a set of k vertices: V = {v1, v2…vk} and E is a set of k edges: E = {e1, e2…ek} where ei is a line segment joining vi and vi+1, 1≤i≤k. i+1=1 if i+1>k.

First all the convex vertices of the polygon are extracted because only convex vertices are considered to find the visibility between two data objects. Assume that a polygon P(V,E) of n convex vertices is stored in the form of adjacency matrix A of order n×n where A [I,J]=1 if edge (I,J) exists between vertices I and J i.e.(I,J) ε E.

A[I,J]=0 if (I,J) not ε E.

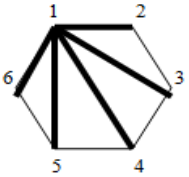
The algorithm returns the output half matrix O of order n×n where O[I,J]=1 denotes the presence of a reduction line and O[I,J]=0 denotes the absence of line.



(a) Input polygon

	1	2	3	4	5	6
1	0	1	0	0	0	1
2	1	0	1	0	0	0
3	0	1	0	1	0	0
4	0	0	1	0	1	0
5	0	0	0	1	0	1
6	1	0	0	0	1	0

(b) Input matrix A



(c) Output matrix

Figure 1 (a) Original Input polygon with six data points 1 to 6, (b) Input matrix of six data points, (c) Neighboured of data point 1 corresponding to others.

Definition: Neighbourhood

The neighbourhood of point p in denoted by $N_{Nb}(p)$ and defined as $N'_b(p)=\{ q\varepsilon D \mid dist(p,q)<=Nb\}$. Its size is denoted by $| N'_{Nb}(p)|$.

As shown in figure 1(c) we can find out the neighbourhood of a seed point and calculate distance between all the neighbour data points of seed pixel. On these data points we have to use the concept of quartile to eliminate outliers in a cluster.

Definition: Outliers

In statistics, an outlier is an observation that is numerically distant from the rest of the data. Grubbs [16] defined an outlier as: An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.

Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy-tailed distribution. In the former case one wishes to discard them or use statistics that are robust to outliers, while in the latter case they indicate that the distribution has high kurtosis and that one should be very cautious in using tools or intuitions that assume a normal distribution. A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations, or may indicate 'correct trial' versus 'measurement error'; this is modeled by a mixture model.

In larger samplings of data, some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated an assumed family of probability distributions, or it may be that some observations are far from the center of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers are to be expected (and not due to any anomalous condition). Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

Definition: Quartile

Just as the median divides a set of observations (when arranged in ascending or descending order of magnitudes), into two equal parts, similarly quartile divides the observations into four equal parts. The value of the item midway, between the first item and the median is known as first or lower quartile and is denoted be Q_1. The value of the item midway between the last item and the median is known Third or Upper Quartile and is denoted Q_3. The median is known as the Second Quartile and denoted by Q_2.

Case I: For ungrouped data.

Q_1 = n+1/4th item, Q_3 = 3(n+1)/4 the item.

Case II: For a frequency distribution

Q_1 = L + (n/4-C)/f x i, Q3 = L + (3n/4-C)/f x i, where L = lower limit of the class in which a particular quartile lies, f = Frequency of the class-interval in which a particular quartile lies. i = Class-interval of the class in which a particular quartile lies, C = Cumulative frequency of the class preceding the class in which the particular quartile lies.In general, Q_h = L + [(nh/4) - c]/f x i, h = 1, 2, 3, 4.

Quartile Algorithm

The concept of quartiles is that we have arranged the data in ascending order and divide it into four roughly equal parts. The upper quartile is the part containing the highest data values; the upper middle quartile is the part containing the next-highest data values, the lower quartile is the part containing the lowest data values, while the lower middle quartile is the part containing the next-lowest data values. Here's where it starts to get confusing. The terms 'quartile', 'upper quartile' and 'lower quartile' each has two meanings. One definition refers to the subset of all data values in each of those parts. But the terms can also refer to cut-off values between the subsets. The 'upper quartile' (sometimes labelled Q3 or UQ) can refer to a cut-off value between the upper quartile subset and the upper middle quartile subset.

In descriptive statistics, a *quartile* is one of three points that divide a data set into four equal groups, each representing a fourth of the distributed sampled population. It is a type of quintile. In epidemiology, the four ranges defined by the three values discussed here.

- first quartile (designated Q_1) = lower quartile = cuts off lowest 25% of data = *25th percentile*
- second quartile (designated Q_2) = *median* = cuts data set in half = *50th percentile*
- third quartile (designated Q_3) = upper quartile = cuts off highest 25% of data, or lowest 75% = *75th percentile*

The difference between the upper and lower quartiles is called the *inter-quartile range* that helps in to calculate the neighbour edges. We can consider the datasets of any range between {p1 to pn}. We can choose any data point between the range p1 to pn as a seed point and compute the distance [17] between individual. Like we had taken a data point 1 as shown in figure 1(c). We have to calculate distance between its neighbourhoods. In the diagram one is connected with {2, 3, 4, 5, and 6}. Similarly if we consider one as a seed pixel and find its all neighbour in a large dataset, it will gives related value of all connected points to it.

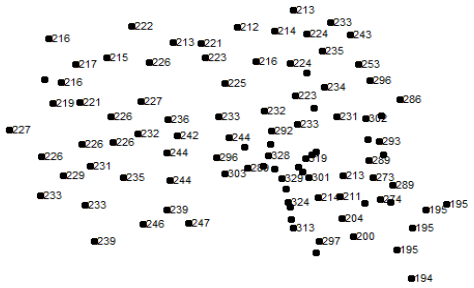
Algorithm: clustering based quartile algorithm

***Input*:** Dataset D of n Data Points, Number of seeds k. Select k seeds .Initialize clusters: each seed is assigned to a cluster. Grow clusters from seeds until all points have been assigned to a cluster and all clusters achieve their target state.

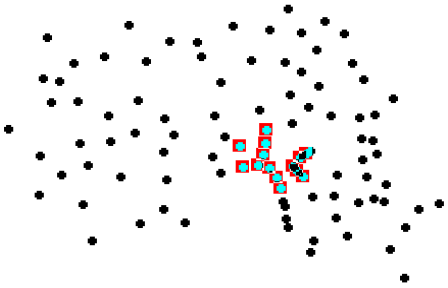
***Step 1*:** Import spatial database file.

Step 2: Call Polygon Reduction method for given data points because it is bi-directed. It will reduce the size of memory by taking total number of directed edges from seed point to its neighbours.

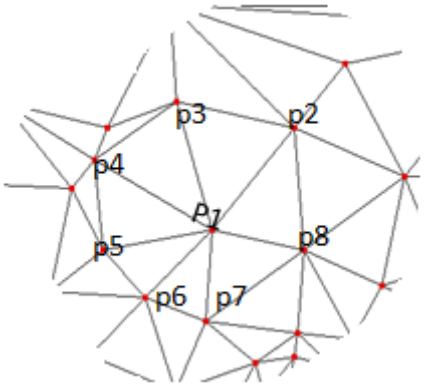
Step 3: Calculate distance of all neighbours from the seed point and stored in a stack as shown in figure 2 (a), (b), (c).



2 (a)



2 (b)



2 (c)

Figure 2(a) original spatial dataset (b) Selected cluster from where we have to precede polygon reduction (c) calculating distance from seed pixel p1 to its neighbours.

Step 4: Sorted calculated distance from seed pixel to its neighbours either in ascending or descending order by applying any sorted algorithm.

Step 5: Compute first quartile

QI= QI=0.25(total edges+ 1)th element

Step 6: Compute third quartile

QIII=0.75(total edges+1)th element

Step 7: Calculate Interquartile Range i.e. IQR

IQR= QIII-QI

Step 8: Detect outliers from IQR. If any Edge is greater than IQR are declared as outliers.

If the edges greater than calculated value are regarded as extreme edges and corresponding data points are stated as outlier.

Experimental Results

We evaluate this algorithm according the three major requirements for clustering algorithms on large spatial databases, Firstly; it reduced polygon lines to calculate distance of neighbours from seed data point. Secondly, this algorithm saves memory by reducing polygon edges. Third it handles efficiently outliers. If we compare this algorithm with the simple DDT clustering algorithm [18] in terms of effectiveness and efficiency. It discover outliers as shown in figure 3.

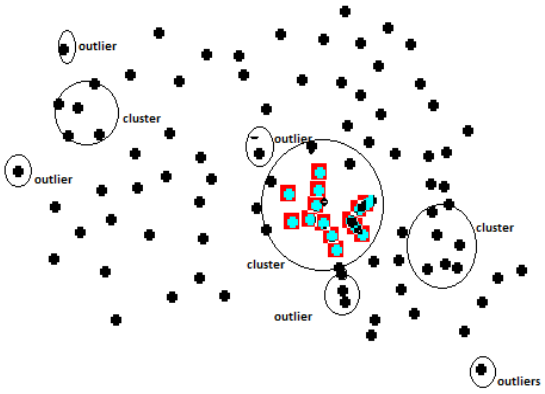


Figure3. Clusters and Outliers are detected

It is very efficient because we used Euclidian distance metrics which reduced processing time and gives better performance.

Complexity of Algorithm

Complexity always based on two parameters time and space. Here we are working on spatial high dimensional datasets. Due to the reason, complexity becomes core issue. Complexity of the algorithm used we need to highlight which components are dependent on the number of points and therefore play an important part in the definition procedure.. The time needed to reconnect the points once the elements forming the void are found can be considered constant. If we maintain a topological relationship among the elements so that each polygone also stores the reference to all its neighbours we can consider T1k as proportional to the number of elements in the void which is independent from k. With the exception of the first search all other operations are of local nature and may be carried out in a time independent of the number of points currently in the structure. Therefore we can estimate the time complexity of the algorithm as roughly proportional to Tk. The search time Tk will be proportional to k leading to an overall time complexity of O (N2). Using the data structure described, the cost of the first search can be reduced to O (logN) giving an overall time complexity for the quartile algorithm of O (NlogN).

Conclusion

The proposed clustering and outlier detection system has been implemented using ArcGIS, MapInfo and calculated distance with Euclidian metric. The application of clustering algorithms to large spatial databases raises the following requirements 1) minimal number of input parameters, 2) discovery of clusters using polygon reduction method and 3) efficiency on large databases. The well-known clustering algorithms offer no solution to the combination of these requirements. In this paper we presents simple and effective algorithm using polygon reduction and concept of quartile for large spatial dataset, which may be useful for a wide variety of applications in clustering spatial datasets. We have taken the concepts of quartile, inter-quartile range to calculate the neighbor edges to detect outlier. Our notion of a cluster is based on the distance of the points of a cluster to their neighbours. The neighbouring region formed in our algorithm reflects the neighbour's distribution. Experimental results demonstrated that our clustering algorithm can provide significant improvement of accuracy of the cluster detecting, especially for objects with arbitrary and linear distribution.

Acknowledgment

The authors would like to thank the YMCA University of Science and Technology and DCRUST Sonepat to support us in our research and allow us to use laboratory for various software and allow accessing various research reports.

References

[1] Zaïane, O. R.and Lee, C. H.: "Clustering spatial data when facing physical constraints". In the Second IEEE International Conf. on Data Mining, Maebashi City, Japan, pp. 737–740, 2002.

[2] Tung, A.K.H., Hou, J., and Han, J. "Spatial clustering in the presence of obstacles", in Proc. 2001 Intl. Conf. On Data Engineering, Heidelberg, Germany, pp. 359~367,2001.

[3] Estivill-Castro, V. and Lee, I. J., "Autoclust+: automatic Clustering of point-data sets in the presence of obstacles", in Proc. 2000 of Intl. Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, Lyon, France, pp. 133–146, 2000.

[4] Xin Wang, Camilo R,."Density-based spatial clustering in the presence of obstacles and facilitators". Technical Report CS-2004-9, May 2004.

[5] Zaïane, O. R.and Lee, C. H., "Polygon reduction: an algorithm for minimum line representation for polygons", In Submitted to 14th Canadian Conf. on Computational Geometry, 2002.

[6] G. Rote, "Computing the minimum Hausdorff distance between two point sets on a line under translation," in *Information Processing Letters*, 38, 1991,123-127.

[7] R.T. Ng, and J. Han, "Efficient and effective clustering methods for spatial data mining," in *Proceedings of 20th International Conference on Very Large Databases*, Santiago, Chile, 1994, 144 – 155.

[8] W. Wang, J. Yang, and R. Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining," in *Proceedings of the 23rd Very Large Databases Conference (VLDB 1997)*, Athens, Greece. 1997

[9] M. Ester, H.-P Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise",in *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96)*, Portland, OR, 1996, pp. 226-231.

[10] A.Hinneburg, and D.A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in *Proceedings of 4th International Conference on Knowledge Discovery and Data Mining*, New York City, NY, 1998, 58 – 65.

[11] D. Joshi, J. Zhang , A. Samal, and L.-K. Soh, "A distance function for polygon-based spatial cluster," *International Journal of Geographical Information System*, submitted for publication.

[12] J. Schwartzberg, "Reapportionment, gerrymanders, and the notion of compactness." Minnesota Law Review 50 (1966) 443 – 452.

[13] J. Sander, M. Ester, H.-P Kriegel, and X. Xu, "Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications," *Data Mining and Knowledge Discovery*, 2, 1998, pp. 169-194.

[14] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications" in *Proceedings of the ACM SIGMOD Conference*, Seattle, 1998, pp 94-105.

[15] D. Fisher, "Knowledge acquisition via incremental conceptual clustering", *Machine Learning 2(2)* ,2007, pp 139 – 172.

[16] Grubbs, F. E.: Procedures for detecting outlying observations in samples. Technometrics 11, 1–21, 2005.

[17] Mamta Malik, Dr. Parvinder Singh, Dr. A.K.Sharma", A Novel Spatial Clustering approach for Outlier Detection & Cluster Generation by probing various Distance Matrices & Delaunay Triangulation", in IJCST Vol. 2, Iss ue 2, with ISSN : 2229 - 4333(Print) | ISSN : 0976 – 8491, June 2011.

[18] Mamta Malik, Dr. Parvinder Singh, Dr. A.K.Sharma," Proposed Dynamic Delaunay Triangulation (DDT) based Clustering Algorithm for Spatial Datasets", IJSTM Vol. 2, Issue 2, ISSN: 2229-6646, April 2011.

[19] J. Lin; Y. Fang, B. Chen, and P. Wu: *Analysis of Access Control Mechanisms for Spatial Database*. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B8, (2009).

GIS and RS for Sustainable Development Integrating with Data Clustering Techniques

Mamta Malik[1], Dr. Parvinder Singh[2] and Dr.A.K. Sharma[3]

[1, 2] DCRUST, Murthal, Sonepat, India
[3]Professor &Dean, YMCA University of Science & Technology, Faridabad, India
E-mail: malik.mamta@gmail.com

Abstract
A Geographical Information System (GIS) is a system of hardware, software and procedures to facilitate the management, manipulation, analysis, modelling, representation and display of geo-referenced data to solve complex problems regarding planning and management of resources. Whereas Remote Sensing (RS) is the technique of deriving information about objects on the surface of the earth without physically coming into contact with them. Clustering is a tool used by GIS and RS who specialize in the field of spatial analysis for sustainable development. Most of the applications of clustering till date have been confined to the field of epidemiology and web application, though of late. Applications have been found in crime data analysis, Infrastructure development, demographic data analysis, rural and urban planning for sustainable development using GIS. In the general context, a cluster is an 'aggregation of same kind of useful information'. Any data that shows geographic (spatial) variability can be subject to cluster analysis. In this paper concept of clustering integrating with GIS and RS techniques for sustainable development are introduced.

Index Terms- Geographical Information System (GIS); Remote Sensing (RS); Data Clustering; Sustainable development

Introduction

Geographic Information Technology has developed at a remarkable pace over the past two decades and will play a key role in development of nations in the 21st Century; there upon many countries have already prepared their strategic development plans for application of GIS Technology with gigantic financing endeavours. Now time has come for all decision makers to discuss the appropriateness of GIS and RS technology and its applications to rural development, forest management, urban development planning, land information systems and agricultural development. This will also provide a suitable solution for the use of GIS and RS for educational infrastructure development with special emphasis on rural sector in India.

Educationists, planners, researchers, decision makers, administrators, communication professionals and officials from different departments and some reputed NGOs should be invited to discuss the role of GIS Technology and should implement the same outcome immediately for ensuring sustainable development and socio-economic and educational uplifting of the country.

Information Technology has emerged as an inevitable phenomenon influencing every walk of life of people in all sections of this society. With the ease of availability of enormous computing power and convenient access to large volume and variety of data and information, the structure and functions of all human organisations will undergo profound transformation in this century. Nations are engaged in exploiting this phenomenon for many of their socio-economic requirements. One area, which is engaging serious attention, relates to use of Information Technology in the governance systems, especially the tools and techniques for acquisition and management of data relating to Geographic Information System (GIS) for sustainable development.

The greatest challenge for sustainable development in the less developed world is to bring the scientific advances of the information technology and geographic information revolutions to bear on problems of severe environmental degradation while at the same time improving livelihoods. Although evaluating impact of information technology is complicated and difficult to carry out, we speculate on the impact or potential impact of geographic information technology on sustainable development. We conclude with some recommendations about using GIS and RS integrating with clustering techniques at local scales in future work.

This paper is organized as follows. Section 2 explains GIS, RS techniques giving a background on spatial clustering for sustainable development. Section 3 defines how clustering techniques are useful for planning and development. Section 4 presents some applications with experimental results. Finally, our conclusion and future scope is given in Section 5.

Sustainable Development

Sustainable development aims at maintaining the equilibrium between the human needs and economic developments within the parameters of environmental conservation through efficient use of natural resources to ensure trade off between desired productions - consumption levels [7]. The well-known Brundtland Commission defined sustainability as a "development that meets the 11needs of the present without compromising the ability of future generations to meet their own needs. In essence, the sustainable development is a process of change in which the exploitation of resources, the direction of investments, the orientation of technological

development and instrumental changes, all are in harmony". The sustainable development of natural resources is based on maintaining the fragile ecosystem balance between the productivity functions and conservation practices through monitoring and identification of problem areas, agricultural practices, crop rotation, use of bio-fertilizers, energy efficient farming methods and reclamation of underutilized lands [8]. Sustainable development requires a holistic approach towards natural resources after taking into account the precarious environmental condition. We are taking three parameters to sustainable development; GIS, RS, Clustering techniques. With the help of first two terms we are managing spatial databases and clustering technology used for decision making.

Remote sensing

Remote sensing is the acquisition of information about an object or phenomenon, without making physical contact with the object. In modern usage, the term generally refers to the use of aerial sensor technologies to detect and classify objects on Earth (both on the surface, and in the atmosphere and oceans) by means of propagated signals (e.g. electromagnetic radiation emitted from aircraft or satellites[1], [2]. There are two main types of remote sensing: passive remote sensing and active remote sensing [3]. Passive sensors detect natural radiation that is emitted or reflected by the object or surrounding area being observed. Reflected sunlight is the most common source of radiation measured by passive sensors. Examples of passive remote sensors include film photography, infrared, charge-coupled devices, and radiometers. Active collection, on the other hand, emits energy in order to scan objects and areas whereupon a sensor then detects and measures the radiation that is reflected or backscattered from the target. RADAR and LiDAR are examples of active remote sensing where the time delay between emission and return is measured, establishing the location, height, speeds and direction of an object.

Data acquisition techniques

The basis for multispectral collection and analysis is that of examined areas or objects that reflect or emit radiation that stand out from surrounding areas.

Geographical Information System

A Geographical Information System (GIS) is a system of hardware, software and procedures to facilitate the management, manipulation, analysis, modelling, representation and display of geo-referenced data to solve complex problems regarding planning and management of resources. Functions of GIS include data entry, data display, data management, information retrieval and analysis. The applications of GIS include mapping locations, quantities and densities, finding distances and mapping and monitoring change [9].

Function of an Information system is to improve one's ability to make decisions. An Information system is a chain of operations starting from planning the observation and collection of data, to store and analysis of the data, to the use of the derived information in some decision making process. A GIS is an information system that is designed to work with data referenced to spatial or geographic coordinates. GIS is both a database system with specific capabilities for spatially referenced data, as well as a set of operation for working with data. There are three basic types of GIS applications which might also represent stages of development of a single GIS application i.e.

Inventory Application

Many times the first step in developing a GIS application is making an inventory of the features for a given geographic area. These features are represented in GIS as layers or themes of data. The emphasis at this stage of application development consists of updating and simple data retrieval.

Analysis Application

Upon completion of the inventory stage, complex queries on multiple layers can be performed using spatial and spatial analysis techniques.

Management Application

More advanced spatial and modelling techniques are required to support the decisions of managers and policy makers. This involves shifting of emphasis from basic geographic data handling to manipulation, analysis and modelling in order to solve real world problems.

Clustering Techniques

Clustering is the process of grouping a set of objects into classes or clusters so that objects within a cluster have similarity in comparison to one another, but are dissimilar to objects in other clusters. So far, many clustering algorithms have been proposed. They differ in their capabilities, applicability and computational requirements. Based on a general definition, they can be categorized into five broad categories, i.e., hierarchical, Partitional, density-based, grid-based, model-based [4] and Partitional clustering methods [6]. As spatial databases are a new research area to work with, data mining is being applied to spatial databases as well. Mining in spatial databases is called Spatial Data Mining. Spatial databases like medical image databases, databases of mine fields, databases of marketing surveys, and satellite databases etc. has a large size in terms of terabytes - more than 1,000,000,000,000 bytes of data. It becomes important to manage this large amount of data. This can be done with the help of data mining process through which useful information can be extracted from these large databases. The information to be extracted is in the form of weather forecasting, optimal placement of ATM machines, population evaluation, determination of diseases etc for sustainable development. Clustering techniques can be better applied to spatial databases to find meaningful groups and hidden patterns in the databases. So spatial data mining and spatial clustering are used to predict future trends in spatial databases.

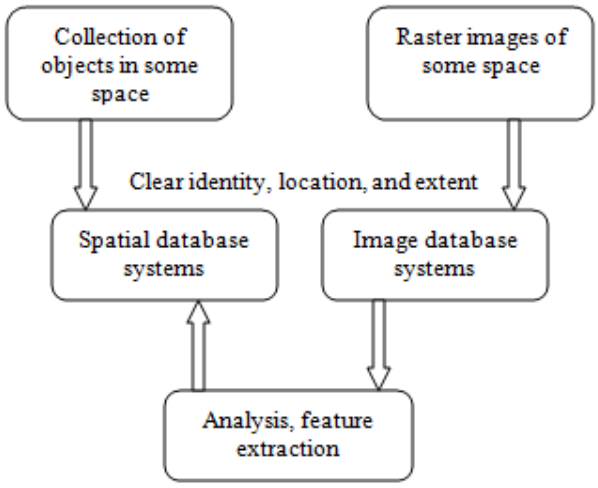
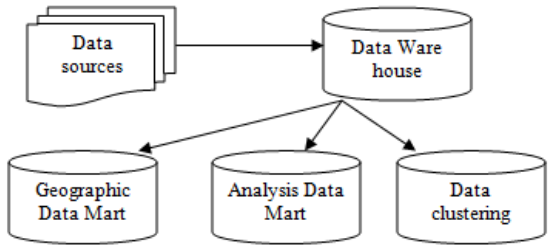
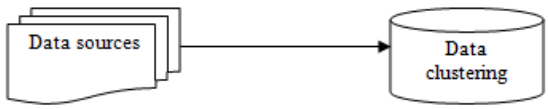


Figure 1: Spatial and image database system

Spatial DBMS provides the underlying database technology for geographic information systems (GIS) and other applications as shown in figure 1. GIS is software to visualize and analyze spatial data using spatial analysis functions such as Search, Location analysis, Terrain analysis, Flow analysis, Distribution, Spatial analysis/Statistics, and Measurements. It is not claimed that a spatial DBMS is directly usable as an application-oriented GIS rather GIS uses SDBMS to store, search, query, and share large spatial data sets explain in figure 2(a), (b).



2 (a)



2 (b)

Figure 2: (a) Generation of Spatial Database Management System as a source data with the help of RS and GIS technology (b) Data Clustering techniques used on database system for knowledge discovery

Spatial dependency and heterogeneity can reflect the nature of the geographic process. Central to spatial data mining is clustering, which seeks to identify subsets of the data having similar characteristics. Two-Dimensional clustering is the non- trivial process of grouping geographically closer points into the cluster. Therefore, a model of spatial proximity for a discrete point-data set P = {p1 to pn} must provide robust answers to which are the neighbours of a point pi and how far the neighbours relative to the context of the entire data set P. A cluster is a group of objects, which are homogeneous among themselves. Clustering has been identified as one of the fundamental problems in the area of knowledge discovery and data mining, and it is of particular importance for spatial data sets. A distinct characteristic of spatial clustering for data mining applications is the huge size of the data files involved. As Tobler's famous proposition states: "Everything is related to everything else, but near things are more related than distant things" [4]. Thus proximity is pretty critical to spatial analysis and in spatial settings; clustering criteria almost invariably makes use of some notions of proximity, usually based on the Euclidean metric, as it captures the essence of spatial autocorrelation and spatial association.

How Clustering Techniques Are Useful In Sustainablity

Modeling of Spatial Data

What needs to be represented?

The main application driving research in spatial database systems are GIS. Hence some modeling needs in this area is considered, which are typical also for other applications. Examples are given for 3-dimensional space, but almost everywhere, extension to the three- or more-dimensional case is possible [11]. There are two important alternative views of what needs to be represented:

1. Objects in space: We are interested in distinct entities arranged in space each of which has its own geometric description.
2. Space: We wish to describe space itself, that is, say something about every point in space.

The first view allows one to model, for example, cities, forests, or rivers. The second view is the one of thematic maps describing e.g. land use or the partition of a country into districts. Since raster images say something about every point in space, they are also closely related to the second view.

We can reconcile both views to some extent by offering concepts for modeling

1. Single objects, and
2. Spatially related collections of objects.

For modeling single objects, the fundamental abstractions are point, line, and region. A point represents an object for which only its location in space, but not its extent, is relevant. For example, a city may be modelled as a point in a model describing a large geographic area (a large scale map). A line is the basic abstraction for facilities for moving through space, or connections in space (roads, rivers, cables for phone, electricity, etc.). A region is the abstraction for something having an extent in 3d-space, e.g. a country, a lake, or a national park. A region may have holes and may also consist of several disjoint pieces.

Figure 3: Representation of point, line and polygon in spatial databases

Thematic map preparation from satellite data using visual interpretation techniques.

- Generation of spatial framework in GIS environment for perspective and development plans.
- Integration of thematic maps using GIS techniques for rural sprawl analysis and urban land use change analysis.
- Area required for ruralisation will be determined on the basis of population projection of the city and its growth centres
- Calculation of land requirements for rural development based on the carrying capacity of the region.
- Projection of rural land use suitability analysis.
- Rural environmental sensitivity analysis based upon both physical as well as air quality parameters.
- Determination of composite functionality index to setup various amenities such as educational, medical, population, sex ratio, recreational etc.

Process of Clustering Integrating with GIS and RS for Sustainable Development

The problem of clustering of Sustainable Development or World Development Indicators data of Countries of the world is considered in this paper. This is with the aim of providing answers to two major questions that are of primary importance to the manner and ways that countries of the world are classified by different organizations. These are:

1. To understand the structure or cluster of countries based on these available data and (2) to investigate the dynamics- that is, how these structure in countries varies from year to year.

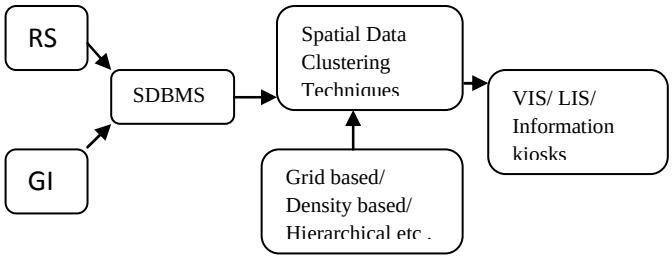


Figure 4: Way to apply clustering over Spatial Database Management System (SDBMS)

The clustering of the data is achieved with the unsupervised and supervised, parameter free approach based on Maximum Likelihood principle, hierarchical clustering, partioning algorithm etc as shown in figure 4. Where Remote Sensing Data and Geographical Information System leads to a spatial database in which different spatial database mining algorithm are applied. Data clustering is important mining tool used for knowledge discovery like Village Information System (VIS), Land Information System (LIS), and various information kiosks according to the requirement of individual for sustainable development.

The clustering of features [11], which are the variables on which the country classification is based, is performed with the aim of observing if they are fully independent or possess some associations. This may be of interest in the investigation of the variables needed to fully classify countries. Since this is not the aim of this study, I simply report my results in the succeeding paragraphs.

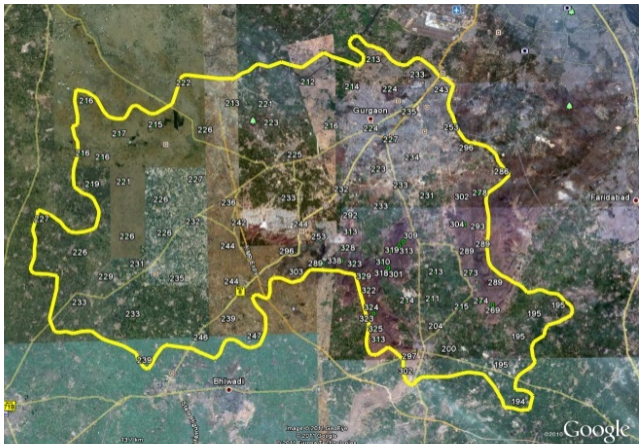


Figure 5 (a): Map showing ground elevation point and district boundary overlay on Landsat satellite imagery Source: Landsat Imagery (30m Resolution), www. **glcf**.umiacs.umd.edu

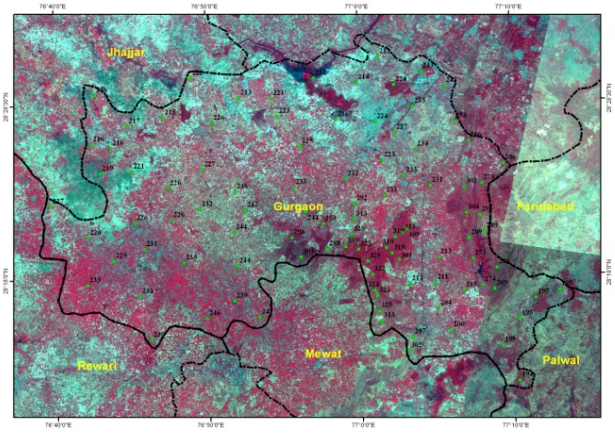


Figure 5 (b): Map showing ground elevation point and district boundary overlay on satellite imagery Source: Google Earth

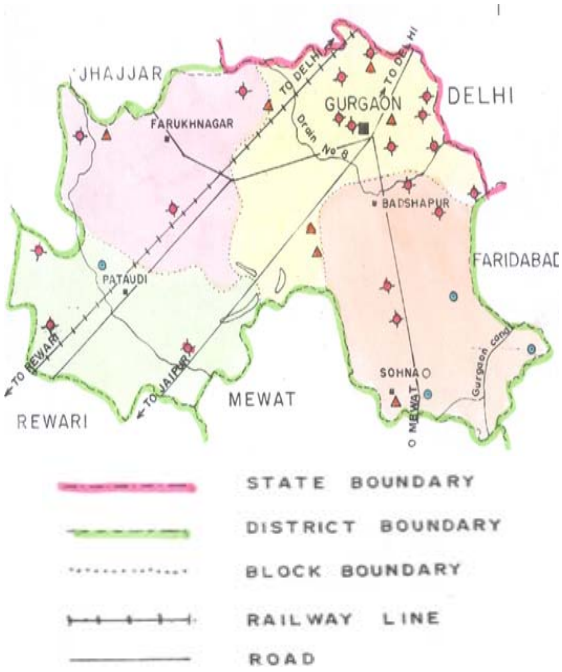


Figure 5 (c): Map showing state, districts, block boundary with railways and road network

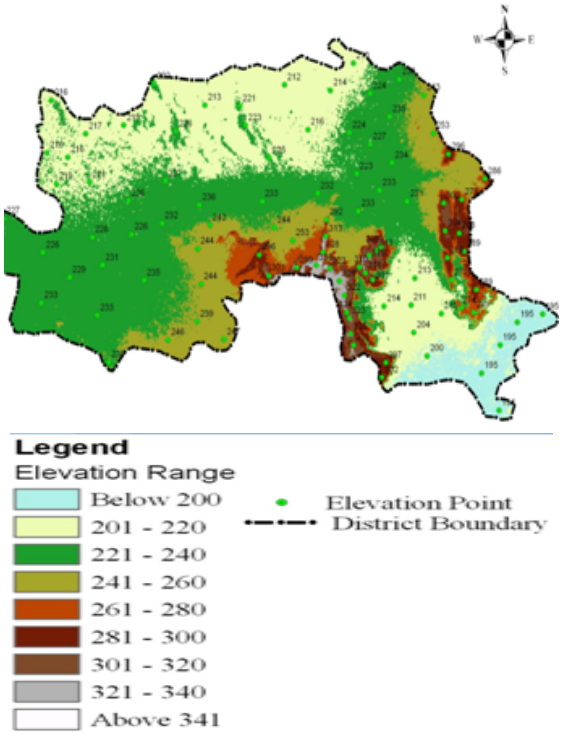


Figure 5 (d): Gurgaon district ground elevations from mean sea level

Figure 5 (e): Clustering data according to our requirement.

Clustering is basically used for spatial database analysis [11]. These results can be used by administrator level, technocrats for planning, managing recourses for sustainable development as show in the figure 5 (a) where Map showing ground elevation point and district boundary overlay on Landsat satellite imagery Source: Landsat Imagery (30m Resolution), with RS and GIS techniques. Figure 5(b) showing ground elevation point and district boundary overlay on satellite imagery Source. Figure 5 (c) Showing states, districts; block boundary with railways and road network and 5 (d) explains Gurgaon district ground elevations from mean sea level. Finally figure 5 (e) clustering data according to our requirement i.e. we want to study demographic data, infrastructure development, Land information system etc. Where information can be retrieved textual as well graphical.

Applications of Clustering In Sustainable Development

Data clustering has immense number of applications in every field of life [4] [11]. One has to cluster a lot of thing on the basis of similarity either consciously or unconsciously. In computer field also, use of data clustering has its own value. Especially in the field of information retrieval, data clustering plays an important role. Some of the applications [7] [8] [9] [10] are listed below.

2. *Marketing*: finding groups of customers with similar behaviour given a large database of customer data containing their properties and past buying records.
3. *Libraries*: book ordering;
4. *Insurance*: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
5. *City-planning*: identifying groups of houses according to their house type, value and geographical location;
6. *Earthquake studies*: clustering observed earthquake epicentres to identify dangerous zones;
7. *WWW*: document classification; clustering web log data to discover groups of similar access patterns.
8. Similarity searching in *Medical Image Database:* This is a major application of spatial clustering technique.
9. *Forest Resources*. Forests are important natural

resources necessary for survival of mankind.

10. *Water Resources*: The major point of concern is pollution of surface water, groundwater and lagoon by wastes from agricultural and urban runoff, domestic and industrial sources.
11. *Disaster Management*: There are many disasters include cyclones/storm surges, tsunamis, flooding due to torrential rains, drought, oil spills, climate change and related consequences, etc need to be managed in advance
12. *Coastal Ecosystem Management* tool to support them in taking management decisions.
13. *Land Information System*
14. *Village Information System* and many more.

All application can be used for sustainable development [9].

Conclusion

This paper focuses on how GIS, RS techniques integrated with data clustering helps in sustainable development, example of Gurgaon District, Haryana taken to explain related issue. Sustainable development is the balance of meeting humankind's present needs while protecting the environment to ensure the fulfilment of future generations' needs. The growing human population and its demands on the earth's resources generate a need for sustainable practices. Implementing these practices often requires collaboration between different organizations and technologies. GIS, RS and Clustering techniques[11][12][13] collectively allows users across the globe to share ideas on how to meet their resource needs, plan efficient land use, and protect the environment to guarantee the survival of future generations

References

[1] Schowengerdt, Robert A. Remote sensing: models and methods for image processing (3rd Ed.). Academic Press. p. 2. ISBN 9780123694072. http://books.google.com/books?id=KQXNaDH0X-IC&pg=PA2, (2007).

[2] Schott, John Robert " Remote sensing: the image chain approach" (2nd ed.). Oxford University Press. p. 1. ISBN 9780195178173, (2007).

[3] Liu, Jian Guo & Mason, Philippa J. Essential Image Processing for GIS and Remote Sensing. Wiley-Blackwell. p. 4. ISBN 978-0-470-51032-2. http://books.google.com/books?id=ae9VPgAACAAJ. (2009).

[4] J. Han and M. Kamber, "Data mining: Concepts and Techniques," Academic Press, 2001.

[5] I. Atsushi and T. Ken, "Graph-based clustering of random point set," Structural, Syntactic and Statistical Pattern Recognition, Springer Berlin, pp. 948–956, 2004.

[6] R. T. Ng and J. Han, "CLARANS: A method for cluster-ing objects for spatial data mining," IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 5, pp. 1003–1016, 2002.

[7] Goreau T.J., 2005, http://globalcoral.org/Mauritius%20Marine%20Management%20Notes.htm

[8] MENDU Report, Mauritius: Staking Out the Future, A Report by Ministry of Environment and National Development Unit, Mauritius. 244p.,2005.

[9] Payet R, Indian Ocean Island- Summary, In Souter, D. And Lindén, O. (Eds) Coral Reef Degradation in Indian Ocean (CORDIO): Status Report. pp.128-132, CORDIO, Stockholm, 2005.

[10] Turner J.R., Klaus R., Hardman E., Fagoonee I., Dabee D., Bhagooli R. and Persands S., The reefs of Mauritius. In D. Souter, D. Obura, O. Linden (eds.) Coral Reef Degradation in the Indian Ocean: Status Report. pp.94-107, CORDIO, Stockholm, 2000.

[11] "Spatial database systems" by Ralf Hartmut Güting, Fernuniversität Hagen, Praktische Informatik IV D-58084 Hagen Germany.

[12] "Density-Based Spatial Clustering in the Presence of Obstacles and Facilitators": Xin Wang, Camilo Rostoker & Howard J. Hamilton Technical Report CS-2004-9.

[13] "AUTOCLUST: Automatic clustering via boundary extraction for mining massive point-data sets": Vladimir Estivill-Castro and Ickjai Lee Department of Computer Science & Software Engineering, The University of Newcastle, Callaghan, NSW 2308, Australia.

Risk Clustering for Diagnosing the Falling Risks in Elderly People Using Self-Organizing Map and Motion Capture Technology

W. Rueangsirarak[1, 2, 3], A.S. Atkins[1], B. Sharp[1], N. Chakpitak[2] and K. Meksamoot[2]

[1]Faculty of Computing, Engineering and Technology, Staffordshire University, Staffordshire, ST18 0AD UK
[2]College of Art Media and Technology, Chiang Mai University, Chiang Mai, 50200, Thailand
[3]School of Information Technology, Mae Fah Luang University, Chiang Rai, 57100, Thailand

Abstract
A Self-Organizing Map was used to classify the risk level of falling based on the criteria of a Risk Assessment Matrix in order to assess the risk in the elderly. This screening system adopts input data collected from elderly Thai people, using Motion Capture Technology. The classification of the screening system based on the result of SOM validation in this study showed 80% accuracy which suggest that the clustering technique is adaptable and useful in falling risk management.

Keywords: Decision Support System; Motion capture technology; Elderly falling risk; Self-Organizing Map

Introduction

The National Survey (National Statistics Office of Thailand) in 2001 indicated that there were 5,969,000 elderly people who are over 59 years old (9.4% of population). In 2007, they had increased to 7,020,000 (10.7% of population). The elderly population is evidently increasing within the last ten years. According to the United Nations online database (2009), the elderly population (≥60 years) is currently 11% and it is expected to increase to 22% by 2045 [1]. While people are getting older, they cannot avoid illnesses such as acute conditions, accidents, general ailments, etc. A recent survey of the elderly population in Thailand indicated that most of the elderly are struggling with the Musculo Skeletal System problems. In Maharaj Nakorn Chiang Mai Hospital, Chiang Mai, Thailand, there are many geriatric patients who are living with an invalid musculo skeletal system [2]. They have difficulty moving their bodies as a healthy person would move and therefore need to have treatment for their disorders. Most of the accidents in geriatric patients are caused by falling. However, this effect accelerates the geriatric patient's risk in breaking a skeletal bone [3].

Falls and fall-related injuries are among the "most serious and common medical problems experienced by elders" [4]. During everyone's life; a person will have at least two bad falls possibly causing severe problems later on in life and approximately 43% of institutionalized elders have trouble with falling each year [5]. The main cause of falls and fall-related injuries is tripping and this is a major contributor to hip fractures in elderly people. Also, walking velocity is one of the many variables that have been associated with falls by elders. The elder who walks slowly has a significantly higher risk of falling [6] and hip fracture [7] and 90% of these falls relate to three issues: gait, balance and mobility. In other cases, they are affected by acute disease and adverse medication [8].

As a result of the evident sickness, physiotherapists are needed to diagnose elderly patients and the number of medical experts is not sufficient for the increasing current numbers of elderly population and this could have serious consequences in the future. Nowadays, visual analysis of human motion is currently an active research topic in computer vision. Human motion analysis concerns the detection, tracking and recognition of people's movements. Understanding human behaviours from image sequences involving humans has been included in the research and Artificial Intelligence (AI) techniques are necessary in this domain, e.g. Decision Support System (DSS), Clustering, and Self-Organizing Map (SOM) to support physiotherapists etc.

This paper explains and discusses the necessity of a DSS to classify the risk of falling in elderly people, using Self-Organizing Map and Motion Capture technology.

Previous study of risk assessment

Risk Assessment Matrix (RAM), by Rueangsirarak and Pothongsunun [9], was created and used as the screening tool as shown in Fig. 1. This enables an elderly fall screening test to be constructed by capturing the experts' heuristic knowledge using Common Knowledge Acquisition and Design System (CommonKADS) model suite. The CommonKADS as well as mining the expert knowledge also consists of constructing different aspect models of human knowledge. CommonKADS's assessment template for falling risk assessment in this study focuses on two issues: firstly, possibility/likelihood refers to biomechanics such as physical gait parameters etc.; secondly, seriousness/consequent concerns daily living activities. Each case is considered through heuristic criteria based on RAM and is disseminated to the physiotherapists to carry out appropriate treatment.

Fig. 1 is classified into two main groups: a likelihood factor and severity factor. This matrix was applied to the outpatient setting, in order to diagnose elderly falling patterns and it consists of nine possible pairs of risk. Heuristic matrix of categories (l, l); (m, l) or (l, m); (l, h) or (m, m) or (h, l) is a result without further need for expert diagnosis. However exception case of (m, h) or (h, m); and (h, h) needs to be referred to an expert for analysis and treatment recommendation [10].

>60%	>50%	>50%	>75%	>75%	>90%	>30%	<10%	<10%	<25°	<25°	>30°	>30°	> 70% of (BSW/2)	> 70% of (BSW/2)	H (3)	(l,h) = 3	(m,h) = 6	(h,h) = 9
41% - 60%	36% - 50%	26% - 50%	61% - 75%	61% - 75%	51% - 90%	16% - 30%	11% - 30%	11% - 30%	26° - 30°	26° - 30°	21° - 30°	21° - 30°	30% - 70% of (BSW/2)	30% - 70% of (BSW/2)	M (2)	(l,m) = 2	(m,m) = 4	(h,m) = 6
<40%	<35%	<25%	<60%	<60%	<50%	<15%	>30%	>30%	31° - 40°	31° - 40°	0° - 20°	0° - 20°	< 30% of (BSW/2)	< 30% of (BSW/2)	L (1)	(l,l) = 1	(m,l) = 2	(h,l) = 3
Base support length	Base support width	Cadence	Stride length (L)	Stride length (R)	Step width	Double support phase	Swing phase (L)	Swing phase (R)	Hip flexion (L)	Hip flexion (R)	Knee flexion (L)	Knee flexion (R)	Centre of Gravity (L)	Centre of Gravity (R)		L (1)	M (2)	H (3)
Possibility/Likelihood (Biomechanics)															Percentage of walking	<10%	10%-50%	>50%
															Type of work	Officer	Deliverer	Heavy carrier
Seriousness/Consequent (Behavior)															Type of exercise	Stretch	Aerobic Exercise	Sport

Fig. 1. Risk Assessment Matrix.

The preliminary result from the Risk Assessment Matrix is used as the data for the screening system to analyze the falling risks by the application of SOM together with motion capture technology as explained in Section III.

Case Review-Related Work

Motion Analysis in Biomechanics

Three-dimensional motion capture or biomechanical evaluation has fast become an indispensable tool in the medical assessment of the neuromuscular and musculoskeletal systems. These opto-electronic motion analysis systems are one of the most accurate means of measuring human motion [11]. The basis of motion capture is the electronic tracking of markers, either passive or active, placed over defined bony landmarks on the body (see Fig. 2 (a)).

Fig. 2. Marker System (a) a participant was installed with markers (b) the participant was walking during motion data collection.

Mathematical algorithms generate 3D joint angles from gait data which helps to analyze the information that is not visible to the eyes e.g. the rate of motion at the knees. Motion analysis is widely used in sport science [12]. The objective of this application is to obtain raw positional data of a segment point (marker) that can be filtered and used to calculate various kinematic derivable variables. These variables are applied to quantify and experimentally validate descriptions of sport techniques, and also to provide biomechanical explanations of the motion patterns observed in sports, to improve the quality and clarity of coaching instruction.

Beside the standard methodology for using a motion analysis system, such as defining the capture volume, completing calibrations and developing appropriate marker systems, which are a preliminary method, there are other procedures that need to be completed for biomechanical analysis. Firstly, the motion analysis positional data is exported to numeric computational software. Secondly, the cut-off frequency needs to calculate and to smooth the data using an algorithm with appropriate filters.This allows the full range of kinematic to be calculated automatically by defined algorithms. Finally, the inverse dynamic is calculated and exported to numeric computational software, in order to perform a kinetic analysis. The motion analysis is applied not only to evaluate an individual performance, but also to suggest methods of optimizing technique for enhanced performance and injury risk reduction, such as fall risk assessment.

Self-Organizing Map (SOM)

Another technique which is used to classify the characteristics of data is Clustering. Clustering is a type of data mining (DM) technique with an unsupervised learning approach. When the clustering is performed, the supporting information helps in the DM process. The intention is to use unsupervised learning and therefore it is not required to know the number of clusters and an attributes in advance [13].

One well known clustering technique is Self-Organizing Map (SOM), which has been proposed by Kohonen in the early 1980's [14]. It is an extremely popular artificial neural network model based on unsupervised learning. SOM models are mostly used for visualization of nonlinear relation of multi-dimensional data. The multi-dimensional data is drawn on to map units, which form the plane of a two-dimensional lattice. SOM will cluster similar data patterns together in the output space while preserving the topology of input space. Network architecture of SOM consists of two layers; input layer and output layer. The input layer is connected to each vector of the dataset (training vector) and the output layer forms a two-dimensional array of nodes (see Fig. 3). The output space often results in the reduction of the dimensionality of the input space which is not shown in Fig. 3.

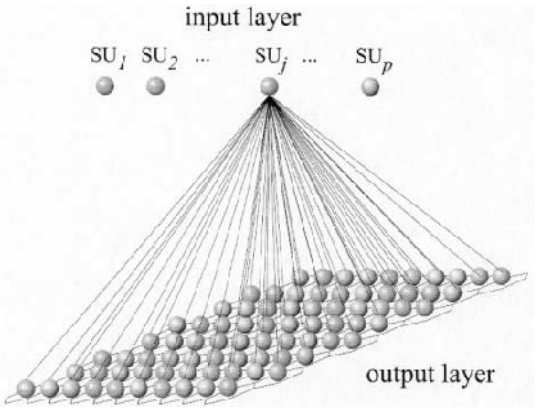


Fig. 3. A two-dimensional Self-Organizing Map [15].

The SOM training process is briefly summarized as an algorithm by Mehotra [16] as follows:

Step 1: Select output layer network topology; initialize current neighborhood distance, *D(0)*, to a positive value

Step 2: Initialize weights from inputs to outputs to small random values

Step 3: Let $t = 1$

Step 4: While computational bounds are not exceeded, do:

1. Select an input sample i_l to the network
2. Compute the distance of i_l from weight vectors (w_j) associated with each output node by using the Euclidean distance equation.

$$\sum_{k=1}^{n}(i_{l,k} - w_{j,k}(t))^2 \quad (1)$$

3. Select output node j^* that has weight vector with minimum value from step 2)
4. Update weights to all nodes within a topological distance given by *D(t)* from j^*, using the weight update rule:

$$w_j(t+1)=w_j(t)+\eta(t)(i_l-w_j(t)) \quad (2)$$

5. Increment *t*

Step 5: End While.

In this algorithm, *D(t)* is the neighborhood function, *t* is the iteration step, *i* is the input vector node, *w* is the weight of the output node, and $\eta(t)$ is a learning rate which ranges in (0, 1). Learning rate generally decreases with time (t):

$$0< \eta(t) \leq \eta(t-1) \leq 1 \quad (3)$$

The most widely used SOM for data analysis has appeared in many areas, for example, the medical domain [17], geographic information systems [18], Pattern recognition [19], Ecological [15], Intrusion detection [20], Expert system [21], Initialization [22], Computation [23] and Seismic [24].

Consequently, an application of motion analysis and self-organizing map can be combined to provide best practice in a medical domain.

Falling risk clustering using SOM

In this section, the procedure of a risk clustering called screening system to diagnose the falling risk in elderly people using Motion Capture technology has been proposed. This screening system investigates the design of the decision support system which applies risk assessment matrix and data clustering techniques. Procedures of combination techniques within the system, and methodologies, are explained as follows:

Firstly all the motion data were collected from the participants by using a Motion Analysis System [25]. These participants are 35 elderly people who are over 60 years old. The participants were asked to wear a motion capture suite and install a marker set on their bodies as show in Fig. 2 (a). They then were asked to walk along naturally to capture different values from the motion capture system (see Fig. 2 (b)). This motion capture system generated a positional data of each elderly person in the form of a three-dimensional coordinate system (x, y, z). The positional data cuts-off the noise within itself in order to prepare the data for biomechanical calculation which is an input data for the fall risk clustering process. Then, the positional data was imported from the motion capture system into the screening system which can calculate the biomechanical parameter of RAM as outlined in Fig. 1. This screening system was developed with Visual C# 2010 Express [26].

The second step of the system is to gather the participants' daily living activities (behavioral data). The data collected from returned questionnaires was analyzed, summarized and stored in the screening system in order to acquire the behavioral parameters of the elderly as a parameter of RAM. Then, both biomechanical and behavioral parameters were used as input data for clustering procedure in order to classify a falling risk.

In the third step, the biomechanical and behavioral parameters were fed into the Java SOMToolbox, [27], by the screening system. The SOM toolbox is an open-source implementation in Java, which was developed at Vienna University of Technology and licensed under the Apache License, Version 2.0. Eighteen input nodes were used corresponding to the parameters of RAM. For the output nodes, the size of map is usually experiment-dependent [28]. Pölzlbauer [29] suggested that the number of output nodes should be in the range of $\sqrt{N}$ (with *N* the number of samples), in order to obtain good mapping results. However, using too few output nodes may cause the congestion of input vectors over an output node, which may make in difficult to distinguish the characteristics of the output space [28]. It is therefore better to use a large number of output nodes. In this study, a map sized *3x3* was used as output nodes, which equates to the Risk Assessment Matrix (RAM) as depicted in Fig. 1. The map is a rectangular lattice which is the default of the Java SOMToolbox.

The weighting of each connection between an input node and an output node was initialized in a random value automatically generated by the Java SOMToolbox. For the training cycle (iteration) decisions, there is no definitive stopping point [28]. The preliminary trial uses enough training cycles so that the network approach is in a stable state. In this study the recommended iteration of five times the number of input vectors as the default of the Java SOMToolbox.

Finally, the screening system classified all elderly data into a group of similarity called a cluster. The elderly person data was clustered into falling categories of risk between low-moderate and high level. The screening system then finalized the diagnostic result for the individual in order to give supported information to physiotherapist. The physiotherapist can then diagnose recommended procedures and appropriate treatment.

Experimental result

SOM presented the clusters of all the data as shown in Fig. 4. As visualization is an advantage of SOM, the group of clusters was represented for further data analysis. The *3x3* rectangular lattices were analyzed and Fig. 4 shows the result of the

clusters. SOM classified the input data into seven clusters as shown in Fig. 4 (a), which represented the amount of members in each cluster. In Fig. 4 (b), (c) and (d), SOM labeled the clusters with pie-chart representation in order to present the classification of data based on criteria of RAM [13]; highest-risk-level rule for (b), risk's weighting score shown as risk-level for (c), and risk's weighting score shown as a number for (d). Each cluster was named depend on the greatest amount of class's member contained in that group.

A performance outcome was measured in terms of the number and percentage of correct classifications and the number and percentage of misclassifications. These classifications were compared to the risk level of fall for each participant as outlined in RAM. When evaluating the results with highest-risk-level classification, Fig. 4 (b), SOM gives an 82.86% correct classification. It could perform better using a risk's weighting score and Fig. 4 (c) shows the modified classification rated at 91.42%. Table 1 portrays the SOM experiments results.

Table I: Performance result of Self-Organizing Map.

RAM Criteria	Type	#	%
Highest-risk-level	Correct classification	29	82.86
	Error	6	17.14
Risk's weighting score	Correct classification	32	91.42
	Error	3	8.58

# denotes the number of members classified in each cluster and % shows the percent of classified member in each cluster

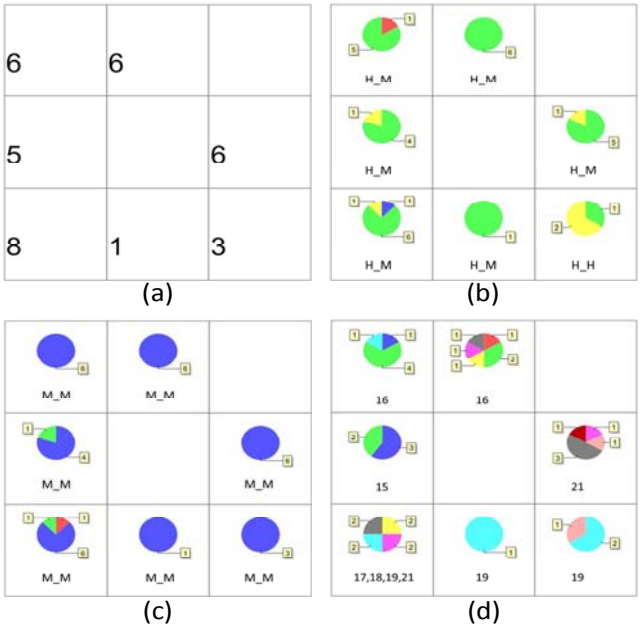


Fig. 4. Results of SOMon rectangular lattices; (a) map unit shown clusters' member; (b) pie-chart for highest-risk-level rule; (c) pie-chart for risk level of weighting score; (d) pie-chart for number of weighting score

As well as indicating the precision of classification, SOM can also present the actual position of the input data in the output map. The layout of clusters was illustrated in Fig. 5 (a). In order to measure the distance between clusters, SOM provided a visualization called U-Matrix to calculate the distance of adjacent prototype vector (closed map units) by Euclidian Distance. Fig. 5 (b) shows the U-Matrix of the falling risk cluster. The map was split into two clouds of clusters, left and right by the border and demonstrated through the lighter color in U-Matrix.

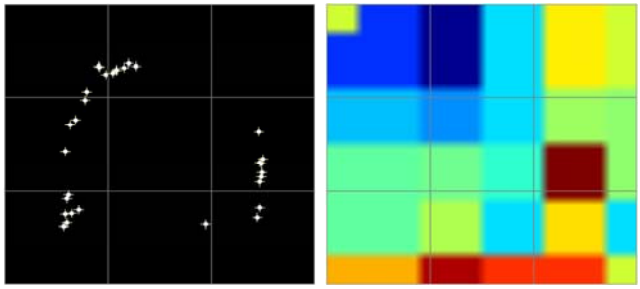


Fig. 5. (a) An exact placement of input data (b) U-Matrix for7 risk clusters.

These SOM's features can help the physiotherapist to make a better decision for diagnosing the risk of falling in elderly people especially the component plane of each input vector, which structured a density of value for each vector as shown in Fig. 6. The physiotherapist could use this component plane to analyze each risk factor in more detail such as feature extraction scheme etc.

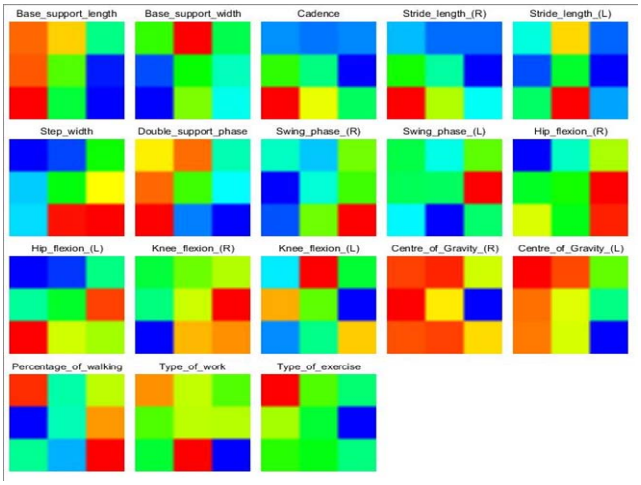


Fig. 6. SOM component plane for each input vector.

Conclusion

The Musculo Skeletal System is the main health problem of elderly people. It is related to gait, balance and mobility and is affected by falling. A new proposed framework of a falling risk screening system was designed to help physiotherapists to make an accurate diagnostic decision. The idea is to combine SOM technique and motion capture together as a decision

support system. This DSS was created by the application of a screening system with a data clustering approach. The classification derived from the screening system process provides the results to physiotherapists so that they can determine how serious the falling risk of the elderly person is likely to be, and this assists in ensuring that the patients will receive the best medical treatment. This screening system can also shorten the health check- up duration because the patients do not have to wait in long queues in the hospital for falling risk analysis.

The advantage of applying SOM together with motion capture technology in the screening system is that the user only needs the input vector to feed into the SOM. Therefore, extra information about risk-level is not required to be embedded in the unsupervised learning process. Based on the validation of SOM in this research, the rate of correct classification of risk of falling is well over 80%. Also, in the post processing, SOM visualization data also provides enhanced information to support physiotherapist's diagnosis.

However, the obstacle in using SOM as a clustering tool during the research is searching for suitable amounts of cluster and the identification of each cluster in the output space. Therefore, another clustering technique should be selected to classify the output space of SOM with a larger size map in order to avoid the congestion of input vectors.

If the cluster of elderly people shows a falling risk between low and moderate level, the result can be simply diagnosed from the knowledge of physiotherapist and appropriate treatment provided. However, in elderly people whose data derived from the cluster appears higher than moderate/high (> (m, h)), as mention in Section II, these groups of patients need to be referred to an expert for a detailed medical analysis and appropriated treatment schemes. Case based reasoning provides a beneficial decision support system to store and retrieve knowledge from an expert to provide solutions for the future.

Acknowledgment

This study has been conducted with the College of Art Media and Technology at Chiang Mai University, Thailand in collaboration with the Faculty of Computing, Engineering and Technology, Staffordshire University, United Kingdom. The research is supported by a grant from Mae Fah Luang University on behalf of the Office of the Higher Education Commission, Thailand.

The achievement of my study has been due to the assistance of a number of people. I am grateful to my parents for the support and love that they have always offered during the difficult times of my study and research. Special recognition and thanks are extended to Asst. Prof. Dr. Prapas Pothongsunun, Mrs. Lyn Atkins and Ms. Nipawan Mantalay for useful recommendations during the procedure of the study.

References

[1] United Nations. (2009), *World Population Ageing, 2009*, Population Division, Department of Economic and Social Affairs.

[2] Tongsawad, T. (1998), *Job Analysis: Elderly Medicine Service*, Department of registrar and statistic, Maharaj Nakorn Chiang Mai Hospital, Faculty of Medicine, Chiang Mai University.

[3] Tunmukkayakul, A. (1983), An Elderly Accident, *The Sarnsiriraj, 35*, 153-9.

[4] Hayes, et al. (1996). In A. J. Van den Bogert, M. J. Pavol, and M. D. Grabiner, *Response time is more important than walking speed for the ability of older adults to avoid a fall after a trip*, Department of Biomedical Engineering, ND-2, The Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195, USA.

[5] Welsh, L. R. (2006), *Preventing Falls from Unpredictable Balance Disturbances*, Unpublished, Graduate School, Oregon State University.

[6] Bath & Morgan (1999). In A. J. Van den Bogert, M. J. Pavol, and M. D. Grabiner, *Response time is more important than walking speed for the ability of older adults to avoid a fall after a trip*, Department of Biomedical Engineering, ND-2, The Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195, USA.

[7] Dargent-Molina, et al. (1996). In A. J. Van den Bogert, M. J. Pavol, and M. D. Grabiner, *Response time is more important than walking speed for the ability of older adults to avoid a fall after a trip*, Department of Biomedical Engineering, ND-2, The Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195, USA.

[8] Tideiksaar, R. (2005). Fall risk factors, (Online) Available at: http://www.seekwellness.com/fallprevention/fall-risk-factors.htm, Cited: 14-Dec-2010.

[9] Rueangsirarak, W., and Pothongsunun, P. (2009). Risk Assessment Matrix for Diagnostic Knowledge of the Elderly Falling Patterns, *3rd International Conference on Software, Knowledge, Information Management and Applications*, 148-153.

[10] Rueangsirarak, W., Atkins, A. S., Sharp, B., Chakpitak, N., and Meksamoot, K. (2011). Application of CBR Techniques in Elderly Falling Risk for Physiotherapist Assessment and Support, *5th IEEE International Conference on Software, Knowledge, Information Management and Applications*.

[11] Gibbs, S. (2008). Clinical applications of motion analysis, *Clinical Nursing & Patient Care, PHHE*.

[12] Ferdinands, R. (2010). ADVANCED APPLICATIONS OF MOTION ANALYSIS IN SPORTS BIOMECHANICS, *XXVIII International Symposium of Biomechanics in Sports*.

[13] Free Books Online (2005), Supervised Vs. Unsupervised Learning. (Online) Available at: http://free-books-online.org/computers/data-warehousing/supervised-vs-unsupervised-learning/, Cited: 28-Mar-2011.

[14] Kohonen, T. (1990). The Self-Organizing Map, *Proceedings of the IEEE, 78* (9).

[15] Giraudel, J.L., and Lek, S. (2001). A comparison of

self-organizing map algorithm and some conventional statistical methods for ecological community ordination, *Ecological Modeling, 146*, 329-339.

[16] Mehotra, K., Mohan, C. K., and Ranka, S. (1997). *Elements of Artificial Neural Networks*. MIT Press.

[17] Basara, H.G., et al. (2008). Community health assessment using self-organizing maps and geographic information systems, *International Journal of Health Geographics 2008, 7* (67).

[18] Lobo, V., et al. (2004). The Self-Organizing Map and it's Variants as Tools for Geodemographical Data Analysis: the Case of Lisbon's Metropolitan Area, *7th AGILE Conference on Geographic Information Science*.

[19] Zhang, J., et al. (2009). Self-Organizing Map Methodology and Google Maps Services for Geographical Epidemiology Mapping, *Digital Image Computing: Techniques and Applications*.

[20] Patole, V.A., et al. (2010). Self-Organizing Maps to Build Intrusion Detection System, *International Journal of Computer Applications, 1* (8).

[21] Chu, X., et al. (2010). An expert system using rough sets theory and self-organizing maps to design space exploration of complex products, *Expert Systems with Applications, 37*, 7364-7372.

[22] Temi, K., et al. (2009). Intrusion Detection with Self-Organizing Map and Learning Classifier System, *CIT2009*.

[23] Zhu, G., et al. (2010). The Growing Self-organizing Map for Clustering Algorithms in Programming Codes, *International Conference on Artificial Intelligence and Computational Intelligence*.

[24] Smith, T., et al. (2011). Introduction To Self-Organizing Maps In Multi-Attribute Seismic Data, *Geophysical Society Of Houston*.

[25] Motion Analysis (n.d.), Motion analysis system, (Online) Available at: http://www.motionanalysis.com/index.html, Cited: 28-Mar-2011.

[26] Microsoft Visual Studio 2010 Express (2010), Visual C# 2010 Express, (Online) Available at: http://www.microsoft.com/visualstudio/en-us/products/2010-editions/visual-csharp-express, Cited: 13-Aug-2011.

[27] Mayer, R., and Rauber, A. (2010). Data Mining with the Java SOMToolbox, (Online) Available at: http://www.ifs.tuwien.ac.at/ dm/somtoolbox/, Cited: 14-Aug-2011.

[28] Lee, K., Booth, D., and Alam, P. (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms, *Expert Systems with Applications, 29*, 1–16.

[29] Pölzlbauer, G. (2004). *Application of Self-Organizing Maps to a Political Dataset*, Unpublished, Master Thesis, Vienna University of Technology.

Database Ownership Issues in Networking – A Roadmap

Dhowmya Bhatt

Department of Information Technology, SRM University, NCR campus, Modinagar, Uttarpradesh, India
E-mail: doveme@rediffmail.com

Abstract
Networks may be classified depending on the vastness and usage of any particular network and taking it to mind the purpose for which has been mainly implemented. In general the Current issues include, local systems the linking standards and database ownership (and its related issues). This paper deals with database owner. The database owner (dbo) is a user that has implied permissions to perform all activities in the database. Any member of the system admin fixed server role who uses a database is mapped to the special user inside each database called dbo. Also, any object created by any member of the system.admin fixed server role belongs to dbo automatically. Also discussed are the related issues like database ownership chaining and the problems related to undetermined ownership. I intend to provide the paper readers get a view about database owner, chaining users and problems in situations where the owner is not known.

Keywords: dbo, svchost, ownership chaining

Introduction

A network is defined as a group two or more of systems such as Windows desktop and server platforms that connect together for the purpose of sharing resources common resources include printers, storage devices and folders that include files and other data. A network, is a collection of computer and devices interconnected by communications channels that facilitate communications and allows sharing of resources and information among interconnected devices. more simply, a computer network is a collection of two or more computers linked together for the purposes of sharing information, resources, among other things. Networks may be classified according to a wide variety of characteristics such as medium used to transport the data, communications protocol used, scale, topology, organizational scope and they are used to give centralized and secure access to networked resources and generally, the entire network is connected to the World Wide Web.

The Current Scenario of Networking

The current issues in networking may be classified depending on the vastness and usage of any particular network and taking it to mind the purpose for which has been mainly implemented. In general the current issues include,

- local systems
- the linking
- linking standards and
- database ownership (and its related issues).

The local system issues mainly popped up when windows-vista was put into use and as years have passed by, when numerous system like apple are also in usage this problem has not solved. The main issues dealing with this would be the lack of speed and occupying too much memory space. This had to be addressed by disabling the SVChost .exe-k through the “superfetch service” and then reboot the system. But this is not practically possible because it takes lot of time and when there are a number of systems commonly disabling SVChost may not work out and solve the problem.

This is one way of dealing with the problem which is simple. This superfetch features can be enhanced so as to meet out the needs of a vast environment. When this features are implemented in wider range and tested, then there will be some remedy to overcome the local system issues. But when SVChost format differs then there shall be some problem in implementing the same exe-k files of the superfetch.[3]

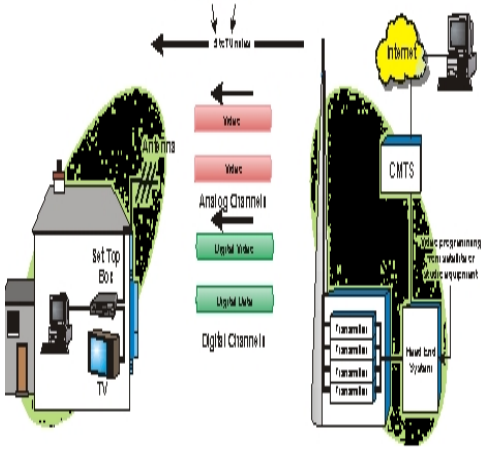


Fig 1. A simple network system showing flow of information

In linking, the usage of a hypertext to specify a particular page or part of a document rather than the main homepage of any website is done so as to save time and the load to the network [6]. It is very useful when the information to be accessed is very vast and the number of pages to be read in the memory is more than few [6]. But the problem arises when the hypertext does not get generated when the linking is done, like when there are two fourms.

Linking standards again refer to hyperlink creation with in a program or any content using a single language (XML) or two languages (ASP, .NET) that can be linked. It is easier to create an hyperlink in a particular language and it consumes less time than creating a link within two different languages. For example in XML, the Xpointer and Xlink are generated as hyperlinks to establish a hypertext connectivity to access information from a network. But there is risk in implementing this and Xpointer version has not been fully implemented in XML leading to linking issues. For a link to be established the linker version of the fourms to be linked has to be defined clearly when it is made,

Database owner(dbo) is the one who creates the d-base or has the authorization to its access. This seems to be simple but the issues arise when the owner does not set or specify the ownership or give the authentication permission. In this case the database becomes open to access and leads to security issues. If the owner wants to set priorities at a later point of time rather than the time of creation, then only encryption-decryption methods can be used by the people sharing the information or content.

In this paper, I shall deal with the issues relating to database owner (dbo) and challenges in the process of authorization by a dbo and the scenario in which the dbo has not set permissions and priorities.

The Database Owner

The database owner (dbo) is a user that has implied permissions to perform all activities in the database. Any member of the system admin fixed server role who uses a database is mapped to the special user inside each database called dbo. Also, any object created by any member of the system admin fixed server role belongs to dbo automatically. The dbo user cannot be deleted and is always present in every database. If the priorities are not set by the owner at the time of creation, then all users have same access to the information.

Only objects created by members of the system admin fixed server role or by the dbo user belong to dbo. Objects created by any other user who is not also a member of the system admin fixed server role,

- Belongs to the user creating the object, not to the dbo.
- Are qualified with the name of the user who created the object and not the dbo's name.

The most common question that arises in our mind is that is it possible to Grant all permissions on a database without granting ownership. This does not have a direct answer but instead few methods have been tried so far which have worked under certain cases and failed in some. The following are the steps,

1. Remove database ownership from user U(the dbo)
2. Grant all permission to user U on the same database (no priorities)
3. Deny delete and drop permissions to U over some specific tables on the database (access to all users using the database).

The Significance of Database Owner

Database ownership is important from a security perspective because the owner account is mapped to the built-in "dbo" user. The "dbo" user, system admin role members and db_owner role members all have full database permissions and can also "DROP" the database. The database owner is also used as the authorization of the "dbo" schema, which comes into play with ownership chaining. With cross-database chaining, the databases involved must have the same owner in order to provide an unbroken chain for "dbo" schema objects. [4]

A difference between the database owner and db_owner role members is that there is exactly one "dbo" user (the database owner) but there may be many users that are db_owner role members. The owner's account cannot be explicitly added to the database because the owner is already implicitly mapped to the "dbo" user and an account can be mapped to no more than one user per database. If any attempt to add the owner as a database user, error message "The proposed new database owner is already a user or aliased in the database" results.

Changing Database Owner

The owner of the current database can be changed. Any user, a SQL Server login or Microsoft Windows user, who has access to connect to SQL Server can become the owner of a database. Ownership of the system databases cannot be changed.

Alter Authorisation can be used to change the ownership of any entity that has an owner. Ownership of database-contained entities can be transferred to any database-level principal. Ownership of server-level entities can be transferred only to server-level principals. In database, change of the owner of system databases master is possible. The model, tempdb, the resource database, or a database that is used as a distribution database cannot be altered. The principal must be a login. If the principal is a Windows login without a corresponding SQL Server login, the principal must have control server permission and take ownership permission on the database. If the principal is a SQL Server login, the principal cannot be mapped to a certificate or asymmetric key. Dependent aliases will be mapped to the new database owner. The DBO SID will be updated in both the current database and in system databases. The schema owner option is only valid when transferring ownership of a schema-contained entity. Schema owner will transfer ownership of the entity to the owner of the schema in which it resides.[7]

Database Ownership Chaining

By default, all database objects have owners. When an object such as a view, a stored procedure, or a user-defined function references another object, an ownership chain is established. For example, when a table is owned by the same user and When the same user owns the source object, the view, stored procedure, or user-defined function, and all target objects underlying tables, views, or other objects, the ownership chain is said to be unbroken. When the ownership chain is unbroken, SQL Server checks permissions on the source object but not on the target objects. The following diagram shows three users

accessing the same information from a database at different level and is chained to share information.

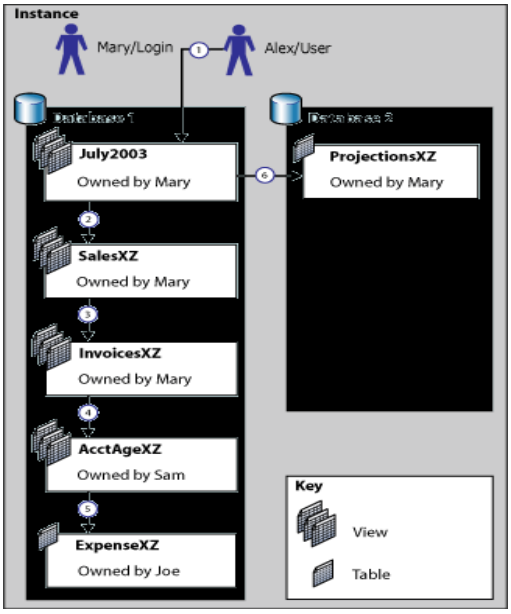


Fig 2. Multiple users and chaining

Cross-Database Ownership Chaining

Cross-database ownership chaining occurs when the source object depends on objects in another database. A cross-database ownership chain works in the same way as ownership chaining in a database, except that an unbroken ownership chain is based on all the object owners being mapped to the same login account. Therefore, in a cross-database ownership chain, if the source object in the source database and the target objects in the target databases are owned by the same login account, SQL Server does not check permissions on the target objects.[9]

If there are more than one database used by an application, and that application calls stored procedures or views in a database that is based on objects in another database, then cross-database ownership chaining is used. Applications that rely on cross-database ownership chaining may generate permission denied errors if cross-database ownership chaining option is turned off.

Database Owner Troubles

To find who owns the database the user has to execute sp_helpdb on SQL Server instances and "owner" column appears. It isn't uncommon to see accounts of people who are no more any part of the database team or the company or moved on to other roles in the organization that don't require privileged database access. Yet these owners still have full database permissions, including the ability to drop the database. To prevent these security issues and other problems, consider establishing an appropriate database ownership standard for any environments.[8]

Database ownership is an often forgotten detail because it is implicitly set to the database creator's account. The owner will initially be a Windows account or SQL login, depending on the authentication method used by the creator. Note that the owner is always an individual account, not a group or role, so a database created by a system admin role member is actually owned by the creator's individual account instead of a built-in security principal (unless the creator logged in using the "sa" account).

A Best Practice is to change the database owner immediately after creating, restoring or attaching a database. Unless there is a reason to do otherwise, specified "sa" as the database owner. This can be done with sp_changedbowner in SQL 2000 or with ALTER AUTHORIZATION in SQL 2005:

For versions of SQL two commands were developed in the year 2005 & 2008 to change ownership from the developer. This worked sometimes and dint many times.

SQL 2000: EXEC MyDatabase..sp_changedbowner 'sa';

SQL 2005 ALTER AUTHORIZATION ON DATABASE::MyDatabase to sa;

SQL 2008

```
ALTER DATABASE database_name
{
 | MODIFY NAME = new_database_name
 | COLLATE collation_name
 | <file_and_filegroup_options>
 | <set_database_options>
}
[;]
```

Risks Associated with Cross-Database Ownership Chaining

Microsoft recommends that disabling the cross-database ownership chaining option because of the actions that highly-privileged users can perform:

- Database owners and members of the db_ddladmin or the db_owners database roles can create objects that are owned by other users. These objects can potentially target objects in other databases. This means that if enabling cross-database ownership chaining, The organization must fully trust these users with data in all databases. To identify the members of the db_ddladmin and the db_owners roles in the current database, execute the following Transact-SQL commands: [5] exec sp_helprolemember 'db_ddladmin'execsp_helprolemember 'db_owner'
- Users with "CREATE DATABASE" permission can create new databases and attach existing databases. If cross-database ownership chaining is enabled, these users can access objects in other databases from newly created or attached databases.

Even though Microsoft recommends that to turn off cross-database ownership chaining for maximum security, there are

some environments where it can be fully trusted highly-privileged users; therefore, can enable cross database ownership for specific databases to meet the requirements of specific applications.

Conclusion

Thus this paper is an attempt to throw light on database ownership and its significance in networking. The risks and troubles related in implementing ownerships of database is discussed. Also ownership chaining and cross ownership chaining is also explained. While changing the owner is not always possible I have tried out certain methods by which under certain circumstances and fulfilling certain criterion, the change of ownership can be done with certain versions of databases. I have also tried out situations where the owner has not set permissions, and under these situations without encryption the users can access the database and alter certain information without changing the basic information. Under situations where there is no ownership, all users are given equal privileges. I hope that the readers will be able to get a outset view about the concepts of database ownership and chaining.

References

[1] C. Rapier and B. Bennett, "High speed bulk data transfer using the database chainingl", MG '08: Proc. of 15th ACM Mardi Gras Conference. pp. 1-7, 2008.
[2] M. Mathis, J. Heffner, P. O'Neil, P. Siemsen, "Pathdiag: Automated SVChosting", PAM 2008.
[3] A. Adams, M. Mathis, "A System for Flexible Network Performance Measurement," Proceedings of INET 2000, July 2000.
[4] V. Paxson, A. Adams, M. Mathis, " Experiences with chaining," Proceedings of the Passive and Active Measurement Workshop 2000, April 2000.
[5] A. Adams, A. J. Lee, and D. Mossé, "Receipt-Mode Trust Negotiation: Efficient Authorization Through Outsourced database Interactions," in Proceedings of the Sixth ACM Symposium on Information, Computer, and Communication Security (ASIACCS 2011), March 2011.
[6] J. C. Honig, D. Katz, M. Mathis, Y. Reckhter and J. Y. Yu, “Applications of database chaining in the Internet”, June 1990, RFC1164 USC/Information Sciences Institute.
[7] R. L. Clay, J. Mahdavi, G. J. McRae, “Scheduling in the Presence of Uncertainty in database chaining. The Linear Assignment Problem,” Proceedings of AICHE National Meeting, August, 1991.
[8] www.quikr.com.
[9] http://dbmmo.com/

Software Reusability-Application through Software Component

Neha Malik[#] and Isha Goel[*]

[#]Assistant Professor, Dronacharya College of Engineering, Gurgaon, Haryana, India
*E-mail: nehag78@gmail.com, *ishagoel06@gmail.com*

Abstract
Reuse engineering is the next generation of information engineering. Most traditional engineering disciplines make far more use of usable components than software Engineering does today. Electronic, mechanical devices use reusability of components from a long time ago, but this concept is little bit new for Software industry. Reusability has brought drastic changes in the field of software. Now instead of making software from starting developers prefer to use already existing parts. This paper tries to put more light on this concept.

Keyword: Reuse, Repositories, Components, Artifacts, COTS

Introduction

Software reuse is the process of creating new software systems from existing software entities rather than building them from scratch. Reuse has been proposed as a key method for overcoming the software crisis and improving the software quality and productivity. Reusability is the degree to which a software component can be reused. Reusability of Component requires some libraries where components can be stored, retrieved and removed in an efficient way. Such libraries are termed as Component Repositories or Libraries.

Classification of Software Reuse

Software reuse has been classified on different basis. The first classification has been done on the basis of Application Domain. There are the two types of Reuse as discussed below.

Horizontal or General Reuse

The reuse of Horizontal assets, sometimes called General reuse, are those which can be reused within other application domains for example library of components, such as linked list class, string manipulation routines Graphical user interfaces (GUI) or databases connections libraries. Horizontal reuse refers to software components used across a wide variety of applications. Horizontal reuse can also refer to the use of a commercial off-the-shelf (COTS) or third party application within a larger system, such as an e-mail package or a word processing program. A variety of software libraries and repositories containing this type of code and documentation exist today at various locations on the Internet.

Vertical or Domain Reuse

Vertical reuse occurs when a component is specific to an application domain for example assets that capture the business knowledge This type of reuse will not work within all domains, for example, it will be difficult to achieve successful reuse for domains which have time and space constraints. The basic idea is the reuse of system functional areas, or domains, which can be used by a family of systems with similar functionality.

The study and application of this idea has introduces another engineering discipline, called Domain Engineering. Domain engineering is "a comprehensive, iterative, life-cycle process that an organization uses to pursue strategic business objectives". It increases the productivity of application engineering projects through the standardization of a product family and an associated production process. This brings us to application engineering, the domain engineering counterpart. Application engineering is" the means by which a project creates a product to meet a customer's requirements."

The second classification has been done on the basis of Software design and mechanism. It classifies the types of reuse into two categories: Conceptual reuse, which is a reuse of ideas, and Program reuse. (Ramel [2])

Conceptual reuse (design reuse) which can further classified as
1) Reuse of models
2) Reuse of architectures

Program reuse which can further classified as
1) Reuse of frameworks
2) Reuse of code
3) Reuse of components

Conceptual Reuse

This type of reuse is also called design reuse. Design reuse is necessary because of three reasons. As one of the initial phases of software development is the design of the system, many errors that affect the whole development process can be eliminated in advance in this early phase. Moreover, the system may be easier to understand and thus easier to develop and to maintain if a familiar design is reused.

1) Reuse of Models: An example for Reuse of Models is so called "design patterns". In programming, especially in object-oriented programming, there are design tasks which occur very often and thus have to be implemented very often. A design pattern is a general repeatable solution to a commonly occurring problem in software design. A design pattern is not a finished design but a practical aid to solve a particular design problem. Design patterns provide a template that can be used

as a solution in many different situations. Object-Oriented design patterns typically show relationships and interactions between classes or objects. The main advantage of design patterns is that they provide solutions which have been developed and permanently improved by many experienced developers over a larger period of time.

2) Reuse of architectures: Reuse of architectures is the reuse of software at the architectural level. The developer can define architecture of an application which may be reused even on other platforms

Program Reuse

A framework provides structures (Ramel [2]) which a developer may reuse in her programs. These are reusable designs that require software components to function. These structures provide certain functionalities like building a web page or the management of database access. An advantage is that different functionalities are combined in a single framework.

1) Reuse of framework: A framework provides standard interfaces and additional configuration files. Programmers can use frameworks directly in their development after providing implementations for the abstract classes. Framework reuse is a mixture of knowledge and code reuse.

2) Code reuse: Code reuse is what every developer does from time to time: Code reuse describes the use of source code that implements the high-level knowledge artifacts. In it a fragment is copied to previously written code. The inserted code, usually just a small amount, is often modified to suit the needs of the new program. The granularity of code reuse can range from 'copying and pasting' a line of code to the reuse of an entire program. Packages and libraries are commonly reused by developers to complete frequent and often mundane tasks, for example, opening a file.

3) Reuse of Components: New software systems may be composed by reusing already existing software components. Libraries typically also contain software components. Software components usually offer predefined services, as opposed to general operations, and are capable of communicating with other components.

The Third classification has been done on the basis of internal visibility of Components. It classifies the types of reuse into three categories:

Black-Box

Black box is concerned with superficial details. Black box reuse means that the developer who assembles the components has no knowledge about how the component is working internally. To be able to add such a component to a software system and thus to reuse its implementation, the component needs to have well-documented interfaces. Reusing a component as a black box means using it without seeing, knowing or modifying any of its internals. The implementation is hidden and cannot be modified by the user. Thus re-users get the information about what a component is doing, but they do not have to worry about how this is achieved. The implementation can be changed without any effects on users. Object-oriented techniques allow modifications of black boxes by making modifications and extensions to a component without knowing its internals.

Commercial-off-the-shelf (COTS) components are examples for black box components.

White-Box

White box is concerned with internal details. The white box approach denotes that the developer knows how the component works internally. The programmer has access to the source code and its methods and functions. If they do not suit her needs the developer may modify them. White-box reuse means reuse of components of which internals are changed for purpose of reuse. They create more opportunities for re-users due to the ease of making arbitrary changes. On the negative side of white box reuse, it requires additional testing and costlier maintenance. Unlike black boxes, a component is thoroughly tested. Additionally, the new component requires separate maintenance. If many copies of a component exist with slight modifications, it becomes burdensome to fix errors that affect all of them. If the changes made to a component are only minor, e.g., a few variable renaming or changes in procedure calls, the term Grey-box reuse is also used. The disadvantage of being able to modify the component is the modification makes it difficult to update or maintain the component. In white-box reuse, a programmer can modify a component before reusing it. Thus the encapsulation property is violated.

Gray-Box

It can also term as glass box. The compromise between the white box approach and the black box approach is called gray box approach. Contrary to the white box extensibility, in this approach the source code is not available but the binary code is available. As a mixture of the white box and black box extensibility the source code is not fully revealed nonetheless details about the implementation of the original system and the extensions are provided. Usually, programmers do not directly reuse components; instead they can access, but not modify the inner workings of a component and use this as an example for their own development. Glass-box reuse contributes in-directly to the quality and productivity of programming because examples can reduce the cognitive load on programmers. Glass box reuse has its negative sides. It may lead to dependencies on certain implementation details that become fatal when the internals of the component are changed. Unfortunately, giving re-users detailed information about a component's internals often serves as compensation of nonexistent or insufficient documentation.

Requirements for effective software reuse

Knowledge contained in the specification of the software is critical for successful reuse of application software. That specification must contain the following:

Documentation of design principles

Software requirements documents contain equations to be

used, but rarely provide insight into how the equations were derived, or how values of constants were determined. This information exists on paper at some point, in the form of informal memos and company internal letters. A basic requirement for successful and safe reuse is having thoroughly documented design underlying principle and design assumptions for both the system and the software design.

Documentation of the assumptions about the operational environment that is implicit in the software
These assumptions include interfaces with other components and other structural features, but also include assumptions about behaviour. Without specification of the assumptions of the environment in which the system was developed, tested, and used, it is not possible to determine what additional testing and analysis needs to be performed or what changes may be necessary to meet the conditions in a different operational environment.

Traceability from high-level system requirements to system design to software requirements to code and vice versa
By traceability, we do not mean simply the mapping between high-level requirements and software modules but instead traceability to system designs features and decisions. Such traceability allows those planning reuse to make sure that the requirements and assumptions about the operation of the component that fit the new use and to determine any interactions with other components that need to be considered. Traceability potentially provides a way to acquire the system knowledge necessary to successfully reuse software.

Documentation of hazard analysis and safety information
A hazard analysis needs to be performed for each safety-critical system. Without information about the original hazard analysis and the specific safety constraints related to the reused software component, it is very difficult to perform this analysis. This difficulty, coupled with lack of documentation of reused software module and common issues with proprietary information, makes reuse very hazardous. Cost effective reuse of safety-critical software requires clear documentation of the assumptions and procedures underlying the original hazard analysis.

Software Reuse Metrics
Metrics are distinctive qualities that help to determine the extent to which reuse should be practiced in system development projects. Software metric is any measurement that relates to a software system, process, or related documentation. The software metrics are grouped into [9] five major categories: general, quality, parameterization, coupling, and cohesion. Quality, parameterization, coupling, and cohesion are software engineering principles that correspond to the reuse attributes. The general category is for attributes that are not in the four software engineering categories.

General Metrics
1) Understandability: Efforts required getting the algorithms, data structures, and control structures of the module. The more understandable the module, the greater the possibility that the module is reusable.

2) Size: Size gives information about the measure of software contained in the module, e.g., lines of code or number.

3) Type of module: Modules can be divided into two major categories: specifications and bodies. Specifications are related to interface, types, exceptions, variables, parameter types, procedures, tasks, and functions. Bodies relates to the algorithms used to implement the declared functions and data.

Quality Metrics
1) Ease of Change: It means the flexibility i.e. difficulties that a developers face to integrate the module into a new system.

2) Comments: The value and accuracy of the comments provided to a designer or programmer who wants to incorporate the module into a new system.

Formatting
The understand ability and readability of the code in the module.
***1) Parameterization Metrics*:** The number *o*f function or data parameters in a module.

2) Coupling Metrics: Coupling *is* the strength of the interconnection and dependency among modules. The higher the coupling, the less independent the module.

3) Cohesion Metrics: Cohesion *met*ric is the strength interconnection with in the component If the components include parts that are not directly related it has a low degree of cohesion. The functional cohesion metric is concerned with the degree to which each part of the module is necessary for performing a single function. The data cohesion metric addresses the degree to which the module has a single-data type associated with it.

4) Complexity Metrics: Software *con*trol structures use graph theory to measure internal procedural characteristics such as number of branches or processing paths Complexity is commonly used to get an entirety of all internal characteristic.

Results and Discussion
Case Study-Application of Reusability of Component named Aurigma Image Uploader Dual
Aurigma Image Uploader Dual is a component intended for bulk file upload. Additionally, it can perform additional operations. Important feature of Image Uploader is that it is embedded into HTML code of your website and does not require end user to download and install it manually.

Aurigma Image Uploader Dual exists as:

- ActiveX Control
- Java Applet

Depending on the browser the end user loads Image Uploader, either ActiveX (for Internet Explorer) or Java version (for Mozilla Firefox, Safari, or other browser supporting Java runtime 1.4) is used. Aurigma Image

Uploader Dual includes special JavaScript file called iuembed.js that makes this choice, as well as unifies client-side API so that you could work with both versions transparently, as if there is the only one component.

To deploy Image Uploader on the server, you should copy the following files along with other files of the web application you are building:

- ImageUploader5.cab
- ImageUploader5.jar
- iuembed.js

Client Side

Java Version

- Any platfrom with Java Runtime Environment (JRE) version 1.4.2_06 or later.
- A browser which supports Java applets with JRE 1.4.
- Internet Explorer (Windows).
- Any Mozilla-family browser, including Firefox, Camino, Netscape and others.
- Safari.
- Konqueror.
- LiveConnect availability (required for some advanced features such as additional form name uploading, event handling, etc).
- Java applets and JavaScript should be enabled in the browser settings.

Server Side

*Any HTTP-com*pliant server platform, including (but not limited to):

- ASP.NET (both Microsoft .NET Framework and Mono are supported)
- ASP
- PHP
- JSP
- ColdFusion
- Perl
- Python
- Ruby

Here we are using it as Java Applet on server side using Apache Tomcat as server as shown in Fig. 1

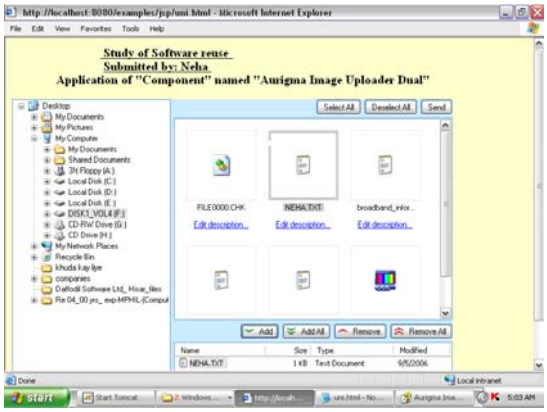


Fig. 1 Implementation of "Augrima ImageUploader"

Image Uploader is designed to be friendly even to untrained users.

- Navigation is familiar to everyone who ever worked with files in Windows.
- Thumbnails are displayed for images, and therefore it is easy to locate necessary image for the upload.
- Ability to upload multiple files and even folders in few clicks. Compare it with a Browse button (when <input type="file" > is used) and feel the difference.
- Progress bar dialog is displayed during upload. The user always knows how long time to wait and they can always stop the upload process if it takes too long time.
- Ability to resume broken uploads without having to send already uploaded files again.
- A number of small but handy features like ability to rotate or remove a file by clicking an icon on the thumbnail, ability to sort items in the upload list, quality meter, tooltips, etc.
- Image Uploader sends data to the server in standard multipart/form-data format. It means that uploaded files can be received with any HTTP-compliant server environment.

Conclusions

This study ponders light on Reusability of Software components. .A software reuse technique is only effective when reusing the components is easier than writing new components. That means that a suitable component has to be found quickly. Therefore components may be stored in repositories, so called component catalogues, with detailed descriptions of what the component does and sophisticated search functions. Thus enabling easy and fast finding of components which fit the developer's needs. Additionally it has to be just as easy and fast to integrate the component into an existing system. The reuse of implementation-level code can enhance the economics of software development.

References

[1] Alan W.Brown ,Kurt C. Brown; *The Current state of CBSE* ; Software Engineering Software for IEEE Software, September 1998.

[2] Ramel S; S/W Reuse in Free Software, State-of-the-Art.pdf, 2005.

[3] Frank McCarey; *Knowledge Reuse for Software Reuse*; College of Engineering, Mathematical and Physical Sciences University College Dublin; September 2007.

[4] Sbeng Zhong; *Software Library for Reuse-OrientedProgram Development*; University of Windsor, Windsor, Ontario, Canada; 2000.

[5] Usa Rungratchakanon, Hisham Haddad; Study of Information Retrieval Systems and Software Reuse Libraries; CSIS Department Kennesaw State University Kennesaw.

[6] Ruben Prieto-Diaz Et.al.; Classifying Software for Reusability; GT Laboratories; 1987.

[7] Rym Mili,Ali Mili, Roland T. Mittermeir, Member IEEE; Storing and Retrieving Software Components; A Refinement Based System; IEEE Transactions On Software Engineering,; July 1997.
[8] Zina Houhamdi; A Formel Language For Software Reuse; Computer Science Department, University of BiskraBP, Algeria.
[9] Steffen Zschaler; Formal Specification of Non-functional Properties of Component-Based Software, Dresden University of Technology.
[10] Oscar Lopez Villegas; Requirements Reuse for Software Development, Technological Institute of Costa Rica. San Carlos Regional Campus. July 2004.

Reliability and Testing Effort Estimation of Web Projects

Anand Singh Rajawat[1], Sangita Tomar[2],Upendra Dwivedi[3] and Dr. Akhilesh R. Upadhyay[4]

[1]JJT University, Jhunjhunu, India
[2]TRUBA College of Science & Technology, CSE Department, Indore, India
[3]JJT University Jhunjhunu, India
[4]Professor and Head, Dept. of Communication Engg., SIRT Bhopal, India.
E-mail: [1]rajawat_iet@yahoo.in, [2]sangitatomar31@gmail.com, [3]ud1985@gmail.com,
[4]akhileshupadhyay@yahoo.com

Abstract
A lot of research has been performed on the issue of reliability of web based project. Reliability is the probability of failure-free web project operation for a specified period of time in a specified environment. It plays a key role in planning and controlling development of projects. It facilitates the organization to deliver a quality product on time and within budget. Nowadays there are large numbers of models available for estimating reliability. Most of the models are complex, expensive and require much effort to estimate the parameters. In this paper, a comparative study of metrics parameters (i.e. defect density, planned efforts, planned time and resources) and defect statistical data (which is measured by testing) is done for measuring and predicting reliability and calculating testing effort. For predicting web project reliability, Moranda Model is used. Through our extended model, we can find out whether our project is reliable or not and at the same time we can also predict its reliability unlike other traditional models which measure only the reliability of any project. Through the proposed model, it is easier to plan the resources requirement. The model is computationally simple and produces fairly accurate results.

Keywords: Web project Reliability, Moranda Model, Web project Reliability Estimation, Defect Statistical Data.

Introduction

In today's technological world many types of project are being developed for many purposes. Now we must know how much reliable these project's are. So in order to test the reliability many web project reliability models have been developed over time. The work on web project reliability models started in 70's, the first model being presented in 1972. This Research paper describes a system that can measure the reliability and calculate testing effort for any web project. For this a comparative study of metrics parameters and defect statistical data is done. Defect statistical data, which is measured by testing, is compared with the metrics, used to predict web project reliability by using the Moranda Model equation. The outcome of this research is presented through graphical representation which facilitates the estimation of web project reliability.

Our system is based on comparison between baseline data and actual data. Baseline data is the collection of standard information from various projects. We have used baseline data provided by reputed software company in this research. We assume baseline data as standard for all our research work. Actual data is collection of testing results of various project for which we are calculating software estimation. Our research compares baseline data and actual data to estimate software reliability. We also define baseline defect density as number of defects for particular size of code. Baseline defect density decides what standard defect density we follow for various project. Project defect density must follow the standard define by baseline defect density to pass our reliability estimation standard. When the baseline effort estimates, revised effort estimates, and actual effort are plotted together for all the phases of SDLC, effort variances are estimated. Schedule variances are calculated at the end of every milestone to find out how well the project is doing with respect to the schedule.

Moranda Model is used to predict the reliability and effort estimation in our research. This model is credited with being the first reliability model [1]. It belongs to a class of exponential order statistic model that assumes that fault detection and correction begins when a program contains faults and all the faults have the same rate. The basic assumptions of the model includes the rate of fault detection is proportional to the current fault content of the web project. The fault detection rate remains constant over the intervals between fault occurrences. A fault is corrected instantaneously without introducing new faults into the software. Every fault has the same chance of being encountered within a severity class as any other fault in that class. The failures, when the faults are detected, are independent.

Web project Reliability Estimation

Nowadays, estimating the reliability of web project is becoming increasingly important. As we know, there are large numbers of models available for estimating reliability. However, most of the models are complex, expensive and require much effort to estimate the parameters. In such a situation, there is a need to develop a model which suits the user's choice, to estimate the reliability of given web project.

Our objective is to derive an approach that produce analytical pictorial reports for actual vs. planned effort

variances, resource, time and cost variances and predict web project reliability through 'Moranda Model'. This system is very easy to use and produces fairly accurate results. It is very useful for monitoring the reliability of many web projects.

Reliability Models

Web project reliability models are used to predict the web project reliability which cannot be estimated unless the development of software is complete. As software reliability model specify the general form of the dependence of the failure process on the principle factors that affect it, namely fault introduction, fault removal and the environment reliability [2] we have analyzed four well-known reliability models. These models include Littlewood-Verall model, Goel-Okumoto NHPP (GO) model, Musa-Okumoto logarithmic execution (MO) model. Our research focuses on Jelinski - Moranda model.

Jelinski-Moranda (JM) Model

The model proposed by Jelinski and Moranda [1] is one of the earliest and the simplest software reliability models. The JM model assumes that times between failures are independent random variables, $T_1,T_2,\ldots$.following an exponential distributions, that there are finite number of faults at the beginning of the test phase, and that the failure rate is uniform between successive failures and is proportional to the current error content of the program being tested. It also assumes that the fault detected is immediately and completely fixed. From these assumptions we have failure rate λ_i

$$\lambda i=\varphi(N-i+1)$$

Where N is the total number of faults in the software at the beginning of the test, i ; is the number of faults detected so far and φ is the reduction in failure intensity per failure per fault. The reliability function is given by $R_i(t)=e^{-\lambda_i t}$ and the current MTTF is given by MTTF $=1/\lambda_i$. The advantage of this model is that it is very simple to use. It is also fairly accurate for some data sets.

Littlewood-Verall (LV) model

Littlewood-Verall model [3] assumes exponential distribution for the random variable T_i representing the failure interval time. But the failure intensity is regarded as a stochastically decreasing function with gamma distribution, implying that the fault fixing process is not considered as perfect, and that faults are of different sizes. A function Ψ (i), which is under the control of the user, determines the nature of the reliability growth. In this model Ψ (i) is taken as $\Psi(i,\beta) = \beta_1 + \beta_2 i$. The current reliability estimate in this model is given by

$$R^p_i(t)=[\Psi(I,\beta^p)/t+ \Psi(i,\beta^p)]\alpha$$

Mean Time To Failure is given by (it does not exist for $\alpha <= 1$) $MTTF=\Psi(i)/\alpha-1 \quad \alpha>1$

Predictions are by maximum likelihood estimation of parameters α_1, β_1, β_2 and use of plug-in rule. The problem with this model lies in complexity involved in determining the parameters. For estimated parameters, MTTF may not be finite.

Goel-Okumoto NHPP (GO) model

The Goel-Okumoto model [4] considers software failure process as a non homogeneous Poisson process. With a mean function μ(t). This model treats initial error contents as a random variable. Time between k - 1^{th} and k^{th} failure depends on the time to k – 1^{th} failure. For the NHPP we have

$$Pr\{n(t)=y\}= ([\mu(t)]^y/y^!) e^{-\mu(t)}$$

for y=0,1,…… With μ(t) considered as

$$\mu(t) = a(1 - e^{-bt})$$

Where a is the expected number of failures in the system and b is the initial failure intensity. Hence, the failure rate can be expressed as, $\lambda(t) = abe^{-bt}$ and $\lambda(\mu)= b(a-\mu)$ is the time between j - 1^{th} and j^{th} failures. From this data cumulative error n(t) can be easily calculated. Estimation of the parameters a and b is by maximum likelihood method.

Musa-Okumoto logarithmic execution (MO) model

The model proposed by Musa and Okumoto [5] views failure process as an NHPP like GO model. But Unlike GO model it assumes that reduction in failure rates are greater for the earlier fixes. MO model assumes failure rate to be an exponential function of the expected number of failures. $\lambda(t) = \lambda_o e^{-\theta\mu(t)}$, where λ_0 and θ are initial failure rate and reduction in the normalized failure intensity per failure respectively. Input to the model is in the form $t_1, t_2, \ldots$, where each t_i represents the execution time. Musa has established the superiority of execution time over calendar time when it comes to software reliability models. However, this model works for calendar times also. The conditional reliability is given as,

$$R(t/t_{i-1})=\{(\lambda_0\theta t_{i-1}+1)/\lambda_0\theta(t+t_{i-1})+1\}^{1/\theta}$$

Estimation of parameter λ_0, θ is by maximum likelihood method. By substituting the estimated values λ_0, θ^P the reliability and other quantities are determined. Execution time is related to calendar time through some suitable assumptions and further computation.

Research Methodology

This research uses the Moranda Model to estimate and predict the reliability. For this we have used the baseline data i.e. metrics parameters. The parameters are defect density, planned efforts, planned time and resources. We also collected the defect statistical data by testing. A comparative study of metrics parameters and defect statistical data has been done for measuring the reliability and calculate testing effort.

Find out metrics

The analysis of metrics relates several data and consolidating the results in terms of charts and pictures simplifies the analysis and facilitates the use of metrics for decision making.

Baseline defect density: Defect density is number of defects for particular size of code. We determine the defect density by using metrics and measurements in our environments.

Defect Density is computed by:

(number of defects)/(1000 line of code)

Effort variances: when the baseline effort estimates, revised effort estimates, and actual effort are plotted together for all the phases of SDLC, it provides many insights about the estimation process.

Effort variances are calculated by:

Variance %= [(actual effort-revised estimated)/revised estimate]*100

Schedule variances: Schedule variances are calculated at the end of every milestone to find out how well the project is doing with respect to the schedule. To get a real picture on schedule in the middle of project execution, it is important to calculate and plot it along with the actual schedule spent.

Defect statistical data

This statistical data is found after the testing process. A comparative study of metrics parameters and defect statistical data is done for measuring the reliability and calculating testing effort. For reliability prediction we use the Moranda Model [1]. The Moranda Model equation is given by this formula:

$\lambda i = \Phi (N - i + 1)$

Where N is the total number of faults in the software at the beginning of the test, 'i' is the number of faults detected so far and Φ is the reduction in failure intensity per failure per fault. The reliability function is given by $Ri(t) = e^{-\lambda i t}$ and the current MTTF is given by $MTTF = 1/\lambda i$ The advantage of this model is that it is very simple to use. It is also fairly accurate for some data sets.

Implementation of the System

Implementing the web based application we used the MVC architecture. In our system, control elements are implemented using servlets or JSP. For measuring and predicting software reliability graphical results are obtained that shows the comparative study of actual vs. planned effort variances. This system is tested and fully implemented. Here, we are enclosing snapshots of running system.

Fig 1. Testing effort estimation tool.

Baseline statistical data list

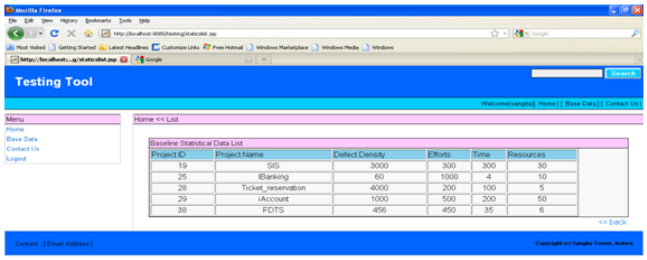


Fig 2. Baseline statistical data list

Import Statistical Data List

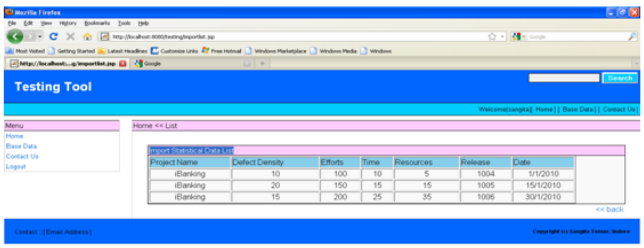


Fig 3. Import Statistical Data List

Actual Statistical Data List

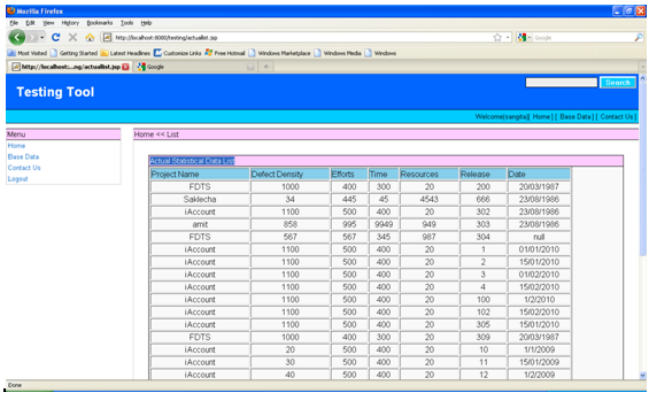


Fig 4. Actual Statistical Data List

Positive factors of proposed System:

- It is very simple to use.
- It produces fairly accurate results.
- It plays a key role in the planning and controlling of software development projects.
- This research is useful for monitoring the reliability of many types of software.
- The model needs extensive comparison with other existing models and reliability statistics of past projects. This will help us to enhance the model.

Result of the System

In this research, we have analyzed and designed a system. This system is based on Moranda Model reliability prediction. We have also assumed statistical data such as baseline data, actual data and baseline defect density. The final outcome of our system is analytical pictorial reports. These reports include

comparison between actual versus planned effort variances and between Resource versus time and cost variances. Moranda Model is a standard model to predict reliability and also produces fairly accurate results.

Fig 5. Base Line Data Graph Representation

Fig 6. Actual Data Graph Representation and reliability prediction through Moranda Model

Conclusion

Through our extended model, we can find out whether our web project is reliable or not and at the same time we can also predict its reliability unlike other traditional models which measure only the reliability of any web project. The model is computationally simple and produces fairly accurate results. Through this model resources can be planned effectively and efficiently in the coming future. It is also very cost effective system which can be implemented by various organizations very easily. This research work is naturally extensible to any similar situation, where there is any need to make reliability prediction web project. There is a possibility to extend this model. The more we learn about past mistakes, the better sure our chances to avoid them in the future and build better web project. In future, more diversity can add to this mode and help organization to maximize their quality efforts.

Acknowledgment

We would like to express our gratitude to all those who gave us the possibility to complete this paper. We want to thank the Department of Engineering of the JJT University for giving me permission to commence this paper in the first instance, to do the necessary research work and to use departmental data. We are deeply indebted to our supervisor Prof. Dr. Akhilesh R. Upadhyay from the JJT University whose help, stimulating suggestions and encouragement helped me in all the time of our research work for our Phd. and writing of this paper.

References

[1] Ahmad, N., Khan, M. G. M., Quadri, S. M. K. and Kumar, M., "Modeling and Analysis of Software Reliability with Burr Type X Testing-Effort and Release-Time Determination", Journal of Modeling in Management, Vol. 4 (1), 28 – 54, 2009.

[2] Huang, C. Y. "Performance analysis of software reliability growth models with testing-effort and change-point", Journal of Systems and Software, Vol. 76, pp. 181-194, 2005.

[3] Huang, C. Y. "Cost-reliability-optimal-release policy for software reliability models incorporating improvements in testing efficiency", Journal of Systems and Software 77(2), pp. 139-155, 2005b.

[4] Stringfellow, C., Andrews, A., "Integrating Defect Estimation Methods to Make Release Decisions," Proc. IASTED Software Engineering Applications, Marina Del Rey, CA, November 2003, pp. 447-452.

[5] Musa, J.D., "Introduction to software reliability engineering and testing." Proc. 8-th International Symposium on Software Reliability Engineering: Case studies, Albuquerque, NM, November 1997, pp. 3-12.

[6] Y. Tohma, K. Tokunaga, S. Nagase, and Y. Murata, "Structural approach to the estimation of the number of residual software faults based on the hyper-geometric distribution," IEEE Trans. Software Engineering, vol. 15, no. 3, pp. 345–355, Mar. 1999.

[7] Phil McMin, Search-based Software Test Data Generation: A Survey, Proceedings in Software

Testing, Verification and Reliability, 2004, 14:105-156.

[8] D. J. Berndt and A. Watkins, High Volume Software Testing Using Genetic Algorithms, Proceedings of the 38th (IEEE) Hawaii International Conference on System Sciences, Waikoloa, Hawaii, Jan 2005.

[9] Nachiappan Nagappan, Laurie Williams, Jason Osborne, Mladen Vouk, Pekka Abrahamsson: Providing Test Quality Feedback Using Static Source Code and Automatic Test Suite Metrics, Proceedings of the 16th IEEE International Symposium on Software Reliability Engineering, Chicago, 2005, pp85 - 94

[10] Z. Jelinski and P.B. Moranda, Software Reliability Research, Statistical Computer performance Evaluation, W. Freibeger (Ed,), New York, Academic,1972.

[11] Lyu, M., Nikora, A.,"Applying reliability Models more effectively, "IEEE software 9(4),July 1992.

[12] B. Littlewood and J.L. Verrall, A Bayesian Reliability Growth Model for Computer Software, J.Royal Statist. Soc., C (App1ied Statistics), Vol. 2, pp 332-346, 1973.

[13] A.L. Goel and K. Okumoto, Time-dependent error-detection rate model for software reliability and other performance measures, IEEE Tr. Reliability, Vol. R-28, No. 3, pp 206-211, 1979.

[14] J.D. Musa and K. Okumoto, A Logarithmic Execution time model for software reliability measurement, Proc. 7th International conference on Software Engineering, Orlando, Florida, March26-29, pp 230-238, 1984

Software Reliability Estimation using *Inflected S-shaped* Model Involving Fault Dependency, Debugging Time Lag and Imperfect Debugging

[1]Dr. Ajay Gupta and [2]Dr. Suneet Saxena

[1]Asstt. Director, Bhagwant Institute of Technology, Muzaffarnagar (U.P.), India
E-mail: ajaydr_gupta@yahoo.co.in
[2]Asstt. Professor, Department of Mathematics, J.P. Institute of Engineering Technology, Meerut, (U.P.), India
E-mail: sunsax75@yahoo.co.in

Abstract
The software responsible for running applications such as internet banking, railway reservation, air-traffic control system, etc are highly complex. The need and the importance for software quality are growing as the functionality of software becomes more complex and critical. The most important attribute in software quality is software reliability, which has attracted an increasing amount of attention in software engineering community. The most important task for software developer is to develop highly reliable software. Over past thirty years, many mathematical models have been proposed for estimation of reliability growth of product during software development process . Such models often referred as software reliability growth models(SRGM).

We have considered *Inflected S-shaped* SRGM and incorporated the fault dependency, debugging time lag and imperfect debugging . Results shows that reliability of software gets improved under imperfect debugging.

Keywords: Software Reliability, Imperfect debugging, Debugging time lag and fault dependency.

Introduction

Software reliability is the probability of failure free software operation for a specified period of time in specified environment. The reliability of software depends on fault detection and correction process. Removing all detected faults will presumably increase the reliability of the software. Ohba [8] conceived that there were two types of faults namely mutually independent and mutually dependent faults. Mutually independent faults can be detected and corrected immediately. There is no time delay between detection and correction. Mutually dependent faults cannot be removed immediately. Goel and Yang [2] analyzed the problem whether detected faults can be corrected immediately or not. Yang [13] reported that detected faults take months to remove for large software system.

Hung and Lin [5] incorporated fault dependencies and debugging time lag into exiting SRGM. They analyzed problem of optimal release time for software system based on reliability and cost criterion. They also assumed detected faults were removed with certainty (perfect debugging). If some of the detected faults are not removed with certainty or new faults introduced during debugging process then it is called imperfect debugging. Yamada, Tokunou and Osaki [11] studied imperfect debugging model models with fault introduction rate. Xie and Yang [13] analyzed imperfect debugging on software development cost.

In this paper we have analyzed the software reliability using *Inflected S-shaped Model* model and generalized it by involving imperfect debugging (b=0.1) and time delay function

$$\phi(t) = \frac{1}{r(1-b)} \ln \left[\frac{\psi - 1}{1 + \psi \exp\{-r(1-b)t\}} \right]$$

Notations

f_0	Expected number of initial faults.
f_i	Total number of independent faults.
f_d	Total number of dependent faults.
r	Fault detection rate of independent faults.
θ	Fault detection rate of dependent faults.
p	Proportion of independent faults.
Ψ	Inflection factor of inflected S shaped model.
$\Phi(t)$	Delay effect factor.
$m(t)$	Mean value function (MVF) of the expected number of faults detected in time (0, t).
$m_d(t)$	MVF of the expected number of dependent faults detected in time (0, t).
$m_i(t)$	MVF of the expected number of independent faults detected in time (0, t).
b	Independent fault introduction rate while removing/fixing a detected fault.

Assumption

a) All detected faults are either independent or dependent. b) The total number of faults is finite. c) The detected dependent fault may not be removed. immediately and it lags the fault detection process by $\Phi(t)$. d) Introduction of new independent faults during debugging. process.

Software Reliability Growth Model

The total detected faults in time (0, t). are given by

$$m(t) = m_I(t) + m_d(t) \quad (1)$$

Independent Faults $m_i(t)$

The rate of independent faults detected is proportional to the remaining faults. We have following differential equation

$$\frac{d}{dt} m_I(t) = r\left[f_I - m_I(t)\right], \quad f_I > 0, \ 0 < r < 1$$

Under imperfect debugging, the differential equation becomes

$$\frac{d}{dt} m_I(t) = r\left[f_I + b m_I(t) - m_I(t)\right], \quad f_I > 0, \ 0 < r < 1$$

Solving equation using initial condition $m_I(0) = 0$ and involving time delay function $\phi(t)$,

we propose

$$m_I(t) = \frac{f_I}{(1-b)}\left[1 - \exp\left\{-r(1-b)(t-\phi))\right\}\right] \quad (2)$$

Dependent Faults $m_d(t)$

The rate of dependent faults detected is proportional to the remaining dependent faults in the system and to the ratio of independent faults removed at time t to the total number of faults. Thus, we have

$$\frac{d}{dt} m_d(t) = \theta\left[f_d - m_d(t)\right] \frac{m_I(t)}{f_0}, \quad 0 < \theta < 1$$

Putting the $m_i(t)$ we get,

$$\frac{d}{dt} m_d(t) = \frac{f_I\,\theta}{f_0\,(1-b)}\left[f_d - m_d(t)\right]$$

$$\left[1 - \exp\left\{-r(1-b)t\right\} \exp\left\{r(1-b)\,\phi(t)\right\}\right] \quad (3)$$

Inflected S-shaped Model

This model was proposed by Ohba. Its underlying concept is that the observed software reliability growth becomes S-shaped if faults are mutually dependent.

Assuming

$$\phi(t) = \frac{1}{r(1-b)} \ln\left[\frac{\psi - 1}{1 + \psi \exp\left\{-r(1-b)\,t\right\}}\right]$$

simplifying equations (2) and (3) , and using $f_1 = p f_0, \quad f_d = (1-p)\ f_0$

we get

$$m_I(t) = \frac{f_I}{(1-b)}\left[\frac{1 - \exp\left\{-r(1-b)t\right\}}{1 + \psi \exp\left\{-r(1-b)t\right\}}\right]$$

$$m_d(t) = f_d\left[1 - \exp\left\{\frac{f_I\,\theta\,t}{f_0(1-b)}\right\}\left\{\frac{\psi+1}{1+\psi\exp\left\{-r(1-b)\,t\right\}}\right\}\right]^A.$$

where,

$$A = \frac{f_I\,\theta(1+\psi)}{r f_0 (1-b)^2\ \psi}.$$

Reliability Analysis

Removing all detected faults will presumably increase the reliability of the software. The software reliability defined as the probability that a software failure does not occur in the time interval $(t, t+\Delta t)$ is

$$R(\Delta t / t) = \exp\left[-\left\{m(t+\Delta t) - m(t)\right\}\right],$$

$$t \geq 0, \ \Delta t \geq 0$$

Assuming $f_0 = 400, \ r = 0.225, \ \theta = 0.0833,$

$p = 0.55, \ \psi = 2.84$

(These numerical constants taken from reference paper [5]) Number of failures m(t) and software reliability R(10/t) have been evaluated under perfect debugging (b = 0) and imperfect debugging (b = 0.1). Further, graphs have also been plotted for m(t) and R(10/t) with respect to testing time t.

Conclusion

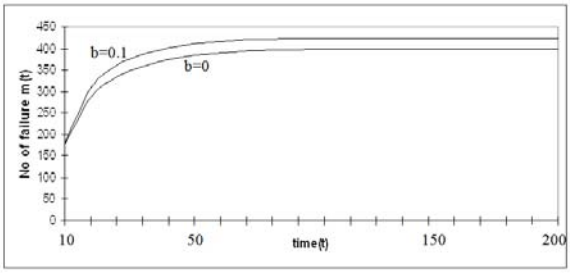
Graph 1 reveals the variation of number of faults detected with respect to testing time. During initial phase of testing time the faults detected are very high and later on becomes constant. The number of faults debugged under imperfect debugging is higher than that in under perfect debugging. This is due to generation of new faults while debugging of detected faults.

Graph 2 shows the variation of software reliability with respect to testing time. Software reliability increases rapidly with testing time during initial phase . Under imperfect debugging (b=0.1) after 140 units of testing time the probability of failure free execution of software in 10 units time interval is 91 % whereas under perfect debugging (b=0) the probability is 85%. This shows that if we incorporate the factors fault dependency, debugging time lag and imperfect debugging into model, prediction of software reliability is more realistic and generalized. Also, we can predict when to stop testing based on reliability of software

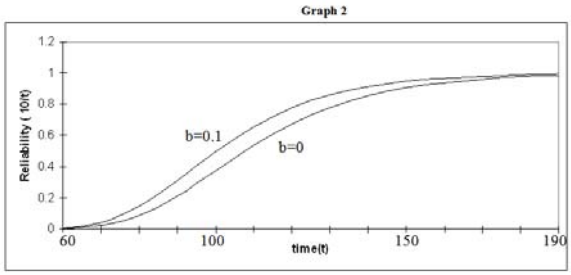
Table 1

No. of Failures under perfect and imperfect debugging			Software Reliability under perfect and imperfect debugging	
Time(t)	$b = 0$	$b = 0.1$	$b = 0$	$b = 0.1$
	$m(t)$		$R(10/t)$	
10	178.0561658	180.9320033	2.92E-48	8.00086E-56
20	287.5088142	307.7972195	1.5514E-20	1.07448E-23
30	333.1213605	360.6848379	1.29972E-11	3.53163E-12

40	358.1876498	387.0540983	2.00858E-07	2.92668E-07
50	373.6083197	402.0983261	6.09271E-05	0.000132109
60	383.3141526	411.030208	0.002169106	0.004730482
70	389.4475927	416.3839362	0.020680304	0.04011762
80	393.3261662	419.5998758	0.086038795	0.144754208
90	395.7791232	421.5325939	0.211956758	0.312969263
100	397.3304962	422.6942442	0.374872523	0.497469835
110	398.3116655	423.3924646	0.537652426	0.657261185
120	398.9322084	423.8121384	0.675390651	0.777050216
130	399.3246724	424.0643887	0.780192274	0.859316212
140	399.5728873	424.216007	0.854717606	0.912897062
150	399.7298715	424.3071392	0.905484578	0.946697031
160	399.8291565	424.3619153	0.939137782	0.967612169
170	399.8919496	424.3948392	0.961064605	0.980405172
180	399.9316632	424.4146286	0.975195809	0.988175815
190	399.9567802	424.4265232	0.984240195	0.992876065
200	399.9726655	424.4336727		



Graph 1



Graph 2

References

[1] American Institute of Aeronautics and Astronautics (AIAA) (1992) : Recommended Practice for Software Reliability, ANSI.AIAA, R-013.

[2] Goel, A. L. and Yang, K. J. (1997) : Software reliability and readiness assessment based on the non-homogenous Poisson process, Advances in Computers, Vol. 45, 197-267.

[3] Hung, C. Y., Lyu, M. R. and Kuo, S. Y. (2003) : A unified scheme of some non-homogeneous Poisson process models for software reliability estimation, IEEE Trans. on Software Engineering, Vol. 29, No. 3, 261-269.

[4] Hung, C. Y. and Lin, C. T. (2006) : Software reliability analysis by considering fault dependency and debugging time lag, IEEE Trans. on Reliability, Vol. 55, No. 3, 436-450.

[5] Hung, C. Y., Lin, C. T., Kuo, S. Y., Lyu, M. R. and Sue, C. C. (2004) : Software reliability growth models incorporating fault dependency with various debugging time lags, Proceedings of the 28th Annual International Computer Software and Application Conference, Hong Kong, China, 186-191.

[6] Hung, C. Y., Lin, C. T., Lo, J. H. and Sue, C. C. (2004) : Effect of fault dependency and debugging time lag on software error models, in Proceedings of the 2004 IEEE Region 10 Conference, Thailand, 243-246.

[7] Lyu, M. R. (1993) : Handbook of Software Reliability Engineering : McGraw-Hill.

[8] Ohba, M. (1984) : Software reliability analysis models, IBM Journal of Research and Development, Vol. 28, No. 4, 428-443.

[9] Ohba, M. and Chou, X. (1989) : Does imperfect debugging affect software reliability growth, in Proceedings of the 11th International Conference on Software Engineering, Pittsburgh, USA, 237-244.

[10] Pham, H. (2000) : Software reliability : Springer Verlag.

[11] Yamada, S., Tokunou, K. and Osaki, S. (1992) : Imperfect debugging models with fault introduction rate for software reliability assessment, International Journal of System Science, Vol. 23, No. 12, 2241-2252.

[12] Yang, K. Z. (1996) : An Infinite Server Queuing Model for Software Readiness and Related Performance Measures, Ph.D. Dissertation, Department of Electrical Engineering and Computer Science, Syracuse University.

[13] Xie, M. and Yang, B. (2003) : A study of the effect of imperfect debugging on software development cost, IEEE Trans. Software Engineering, Vol. 29, No. 5, 471-473.

An Advanced Algorithm for Optimized Scheduling of Hydrothermal Power Systems with Cascaded Reservoirs

M. Manoj Kumar[1], Dr. B. Brahmaiah[2] and Dr. A. Srinivasula Reddy[3]

[1]Associate Professor, Dept of EEE, Velega Nageswararao College of Engg, Guntur Dist., A.P., India
E-mail: manoj_eee7@yahoo.com
[2]Principal, Priyadarshini Institute of Technology, Tirupati, A.P., India
E-mail: bbrahmaiah2007@gmail.com
[3]Professor in Dept. of EEE & Principal, Samskruti College of Engg. & Technology,, Ghatkesar, Hyderabad, A.P.., India,
E-mail: svas_a@yahoo.com

Abstract
An optimization-based algorithm is presented for the short-term scheduling of hydrothermal power systems using the Lagrangian relaxation technique. This paper concentrates on the solution methodology for hydro sub-problems with cascaded reservoirs and discrete hydro constraints. Continuous reservoir dynamics and constraints, discontinuous operating regions, discrete operating states, and hydraulic coupling of cascaded reservoirs are considered in an integrated fashion. The key idea is to substitute out the reservoir dynamics and to relax the reservoir level constraints by using another set of multipliers, making a hydro subproblem unit-wise and stage-wise decomposable. The optimal generation level for each operating state at each hour can be obtained simply by minimizing a single variable function. Dynamic programming is then applied to optimize the operating states across the planning horizon with a small number of well-structured transitions. A modified subgradient algorithm is used to update multipliers. After the dual problem converges, the feasible solution to the hydro power subsystem is obtained by using a network flow algorithm, with operating states obtained in the dual solutions possibly adjusted by heuristics. Numerical testing based on practical system data sets show that this method is efficient and effective for dealing with hydrothermal systems with cascaded reservoirs and discrete hydro constraints.

Keywords: Scheduling of hydrothermal power systems, Scheduling of cascaded reservoirs, Mixed-integer programming

Introduction

Hydrothermal scheduling is an important daily activity for utilities because of its significant economic impact. It aims at determining the commitment and generation of all schedulable power resources over a planning horizon to meet the system demands and reserve requirements. The goal is to minimize the total generation cost. To solve this NP-hard mixed integer programming problem, many algorithms have been developed. Lagrangian relaxation and its extension are among the most successful ([1-5]).

In Lagrangian relaxation, the problem is converted to a two-level optimization problem. The low level consists of a number of subproblems, one for each thermal unit or river catchment, and the high level is to optimize the multipliers. Generally hydro subproblems are more difficult to solve than thermal subproblems. A hydro unit has reservoir dynamics and constraints coupling the hourly generation across time.

Operating in certain regions may not be permitted for security or efficiency reasons, resulting in discontinuous regions or even discrete generation levels. Furthermore, since the reservoirs in a river catchment are hydraulically coupled, the generation of an upstream unit affects the reservoir levels of the downstream units. Finally, to prevent wear-off caused by frequent starting up and shutting down, hydro units may also have minimum up/down time requirements resulting in discrete operating states, and start-up and shut-down costs as described in [17]. Integrated consideration of the above factors within the Lagrangian relaxation framework is a very challenging issue, and is the focus of this paper.

Many methods have been developed to solve hydro subproblems, including dynamic programming (DP), network flow, and standard mixed integer programming methods. DP is flexible and can handle the above mentioned constraints in a straightforward way ([6-7]). However, DP needs to discretize reservoir levels. For systems with cascaded reservoirs and discrete operating states, the state space expands exponentially with problem size, causing DP to suffer from the "curse of dimentionality" for practical applications. Network flow is the most widely used method for hydro power scheduling ([8-12]). Its major limitation, however, is its inability to deal with discontinuous operating regions and discrete operating states, although continuous non-network constraints can be approximated in a network flow formulation as in [8]. General linear and nonlinear programming methods encounter similar limitations as network flow ([13]). Heuristics are sometimes used to obtain hydro commitment, or to post-process solutions obtained by network flow. Recently, a combination of network flow, dynamic programming and heuristic method is reported in [17]. A genetic aided scheduling method is presented in [15], and a multi-pass dynamic programming method with special state approximation is developed in [16]. These two later methods do not consider discontinuous operating regions or discrete operating states. Recently, a commercial mixed-

integer linear programming package is used to generate hydro schedules where integer variables are handled by partial enumeration such as branch-and-bound methods ([14]). When the problem is large and coordination with thermal units is needed, computational requirements may become too large for practical applications. In our pervious work, efficient algorithms for hydro and pumped-storage subproblems were developed and embedded in the daily scheduling package of Northeast Utility Service Company (NU). However, the hydraulic coupling of cascaded reservoirs was not considered.

Relaxing reservoir dynamics and hydraulic coupling by using additional sets of Lagrangian multipliers is an efficient and systematic way to handle discontinuous operating regions, discrete operating states, and hydraulic coupling of reservoirs ([18-21]). By extending this idea, a new algorithm is presented in this paper for solving hydro sub-problems with cascaded reservoirs within the Lagrangian relaxation framework of hydrothermal scheduling. The basic idea is to substitute out the reservoir dynamics and relax reservoir level limits by using another set of Lagrangian multipliers. This relaxation is computationally efficient and numerically stable as compared to relaxing reservoir dynamics ([18]), since reservoir level limits are inequality constraints and many of them may not be active. The sub-Lagrangian associated with a river catchments then becomes unit-wise and stage-wise decomposable, and the cost function of an individual unit involves only the multipliers associated with its own and its direct downstream unit. For each discrete operating state of a unit at an hour, the optimal generation can be obtained by optimizing a single variable function. DP is then used to optimize the discrete operating states of the unit across the scheduling horizon with a very small number of well-structured transitions. The multipliers are updated at an intermediate level. A nonlinear network flow algorithm is then used to generate a feasible schedule at the convergence of the dual problem, with operating states obtained in the dual solution possibly adjusted by heuristics. Numerical testing based on practical size data sets shows that this method is computationally efficient to handle hydro subproblems with cascaded reservoirs and discrete hydro constraints.

Problem Formulation

Consider a hydrothermal power system with *I* thermal units, *J* hydro reservoirs and *K* pumped-storage units. Without loss of generality, assume that there is only one river catchment since with given Lagrange multipliers, hydro subproblems associated with different river catchments are independent. It is required to determine the operating states and generation/pumping levels of all units over a specified period *T*. The goal is to minimize the total generation cost subject to system demand and reserve requirements, and individual unit constraints. The time unit is one hour and the planning horizon may vary from one day to a week.

To formulate the problem mathematically, the notation to be used is first introduced:

I: number of thermal units;

J: number of hydro units;

K: number of pumped-storage units;

$C_{ti}(p_{ti}(t))$: fuel cost of thermal unit i for generating power $p_{ti}(t)$ at time *t*, in dollars;

$P_d(t)$: system demand at time *t*, in MW;

$P_r(t)$: system spinning reserve requirement at time *t*, in MW;

$p_{hj}(w_j(t))$: power generated by hydro unit *j* at time *t*, in MW;

$p_{pk}(t)$: power generated or used for pumping by pumped-storage unit *k* at time *t*, in MW;

$p_{ti}(t)$: power generated by thermal unit *i* at time *t*, in MW;

$r_{hj}(p_{hj}(w_j(t)))$: spinning reserve contribution of hydro unit *j* at time *t*, in MW;

$r_{pk}(p_{pk}(t))$: spinning reserve contribution of pumped storage unit *k*, at time *t*, in MW;

$r_{ti}(p_{ti}(t))$: spinning reserve contribution of thermal unit i at time t, in MW;

$S_{ti}(t)$: start-up cost of thermal unit *i* at time *t*, in dollars;

$S_{hj}(x_{hj}(t))$: start-up cost of hydro unit *j* at time *t*, in dollars;

t: time index, t=1,2,...*T*.

T: time horizon under consideration, in hours;

$v_j(t)$: reservoir level of hydro reservoir *j* at time *t*;

$\overline{v}_j$: maximum reservoir level of hydro unit *j*;

$\underline{v}_j$: minimum reservoir level of hydro unit *j*;

v_j^0: initial reservoir level of hydro unit *j*;

v_j^T: terminal reservoir level of hydro unit *j*;

$w_j(t)$: water discharge of hydro unit *j* at time *t*;

$\overline{w}_j(t)$: maximum water discharge for hydro unit *j* at time *t*;

$\underline{w}_j(t)$: minimum water discharge of hydro unit *j* at time *t*;

$u_{hj}(t)$: discrete decision variable of hydro unit *j* at time *t*, "1" for up, "-1" for down;

$x_{hj}(t)$: state of hydro unit *j* at time *t*, denoting number of hours that the unit has been on (positive) or off (negative);

$\overline{\varsigma}_j$: minimum up time of hydro unit *j* , in hours;

$\underline{\varsigma}_j$: minimum down time of hydro unit *j*, in hours;

$\xi_j(t)$: natural inflow to the reservoir of hydro unit *j* at time *t*;

τ_j: time required for the water discharged from reservoir *j* to reach its direct down stream reservoir, in hours.

B: reservoir connection matrix with element $b_{ij} = 1$ if hydro unit j is a direct up stream of unit i, $b_{ij} = 0$, otherwise.

The scheduling problem can then be formulated as the following mixed integer programming problem

$$\min_{p_{ti}(t), p_{hj}(t), p_{pk}(t)} C, \text{ with}$$

$$C = \sum_{t=1}^{T} \{ \sum_{i=1}^{I} [C_{ti}(p_{ti}(t)) + S_{ti}(t)] + \sum_{j=1}^{J} S_{hj}(x_{hj}(t)) \}, \quad (1)$$

subject to system wide demand and reserve requirements and individual unit constraints to be described below.

System demand:

$$\sum_{i=1}^{I} p_{ti}(t) + \sum_{j=1}^{J} p_{hj}(t) + \sum_{k=1}^{K} p_{pk}(t) = P_d(t),$$
$$t = 1,2,....T. \quad (2)$$

Spinning reserve requirement

$$\sum_{i=1}^{I} r_{ti}(p_{ti}(t)) + \sum_{j=1}^{J} r_{hj}(p_{hj}(w_j(t))) + \sum_{k=1}^{K} r_{pk}(p_{pk}(t)) \geq P_r(t),$$
$$t = 1,2,....T. \quad (3)$$

Thermal and pumped-storage constraints

Detailed descriptions of individual constraints for thermal and pumped-storage units can be found in [21, 22].

Constraints for river catchment and hydro units

Water balance equation

$$V(t+1) = V(t) + \mathbf{B} w_d(t,\tau) - w(t) + \xi(t), \quad (4)$$

where $V(t)$ is the stack vector of $v_j(t)$, and $w(t)$ the stack vector of $w_j(t)$, for all reservoirs in the river catchment, and $w_d(t,\tau)$ the delayed water discharge to downstream reservoirs defined as:

$$w_d(t,\tau) = [w_1(t-\tau_1) \quad w_2(t-\tau_2) \quad \ldots w_J(t-\tau_J)]^{\mathrm{T}}.$$

Equation (4) requires conservation of flow among reservoirs in the river catchment. Without loss of generality, it is implicitly assumed that the time required for water to travel from a reservoir to a reservoir direct downstream is far less than the scheduling horizon.

Reservoir level limits:

$$\underline{V} \leq V(t) \leq \overline{V} \quad (5)$$

Initial and terminal reservoir levels:

$$V(0) = V^0, \quad (6a)$$
$$V(T) = V^T \quad (6b)$$

Operating regions:

$$\underline{w}_j(t) \leq w_j(t) \leq \overline{w}_j(t) \text{ , if } x_j(t) > 0, \quad (7a)$$
$$w_j(t) = 0 \text{ , if } x_j(t) < 0. \quad (7b)$$

Although only two operating regions, (7a) and (7b), are considered, the method developed can be directly used to solve problems with multiple operating regions or even discrete output levels caused by restricted loading bands or operating efficiency.

Minimum up/down time

$$u_{hj}(t) = 1 \quad \text{if } 1 \leq x_{hj}(t) \leq \overline{\varsigma}_{hj}, \quad (8a)$$
$$u_{hj}(t) = -1 \quad \text{if } -\underline{\varsigma}_j \leq x_{hj}(t) \leq -1, \quad (8b)$$

which discourage frequent start-ups and shut-downs.

State transitions

$$x_{hj}(t+1) = x_{hj}(t) + u_{hj}(t) \text{ if } x_{hj}(t)u_{hj}(t) > 0, \quad (9a)$$
$$x_{hj}(t+1) = u_{hj}(t) \text{ if } x_{hj}(t)u_{hj}(t) < 0. \quad (9b)$$

The water-power conversion is approximated by the following concave quadratic function

$$p_{hj}(w_j(t)) = a_j w_j^2(t) + b_j w_j(t) + c_j. \quad (10)$$

The reserve contribution of a hydro unit is calculated as the difference between generation capacity and the generation level

$$r_{hj}(w_j(t)) = \overline{p}_{hj}(w_j(t)) - p_{hj}(w_j(t)), \quad (11)$$

if the unit is up, and is zero if the unit is down.

Solution Methodology

The Lagrangian Relaxation Framework

The basic idea of Lagrangian relaxation is to relax system-wide demand (2) and reserve requirements (3) by using Lagrange multipliers and to form a hierarchical optimization structure. Given the multipliers, the low level consists of individual thermal and pumped-storage units, and hydro river catchments as in [20-22]. The high level dual problem is to update the multipliers. The methods for solving thermal and pumped-storage subproblems have been presented in detail in [20, 22], and only solution methodology for the hydro subproblem is presented here.

Solving hydro subproblem

The hydro subproblem is presented within the Lagrangian relaxation framework:

$$\min_{w_j(t)} L_h, \text{ with}$$

$$L_h = \sum_{t=1}^{T} \{ -\lambda(t) \sum_{j=1}^{J} p_{hj}(w_j(t)) - \mu(t) \sum_{j=1}^{J} r_{hj}(p_{hj}(w_j(t)))$$

$$+\sum_{j=1}^{J} S_{hj}(x_{hj}(t))\} \quad (12)$$

subject to constraints (4)-(9).

The key idea to solve the subproblem is to substitute out the water balance equation (4) and to relax the reservoir level limits of (5) and terminal reservoir level (6b) by using additional sets of Lagrangian multipliers. An intermediate level is thus created. The hydraulic coupling among reservoirs is "cut-off," and the river catchment subproblem then becomes unit-wise and stage-wise decomposable. At the low level, decomposed subproblems for individual units are efficiently solved by first optimizing a series of single variable functions, and then applying DP with a small number of states and well-structured transitions without discretizing the reservoir levels. These multipliers are updated at the intermediate level. At the convergence of the dual problem, a nonlinear network flow algorithm is then applied to generate a near optimal feasible schedule, with the discrete operating states obtained in the dual problem and possibly adjusted by heuristics.

Problem decomposition

After relaxing the reservoir level limits (5) and terminal reservoir level (6b), the sub-Lagrangian (12) becomes:

$$L_h = L_h + \sum_{t=1}^{T-1}[\beta_1^{\mathrm{T}}(t)(\underline{V} - V(t)) + \beta_2^{\mathrm{T}}(t)\ (V(t) - \overline{V})]$$
$$+\beta_3^{\mathrm{T}}[V^T - V(T)], \quad (13)$$

wher $\beta_1(t), \beta_2(t)$ $(t = 1,2,\ldots,T\text{-}1)$ and β_3 are multipliers defined as

$$\beta_1(t) \equiv [\beta_{11}(t),\ \beta_{12}(t),\ \ldots\beta_{1J}(t)]^{\mathrm{T}},$$
$$\beta_2(t) \equiv [\beta_{21}(t),\ \beta_{22}(t),\ \ldots\beta_{2J}(t)]^{\mathrm{T}},\ t = 1,2\ldots T\text{-}1$$

and

$$\beta_3 \equiv [\beta_{31},\ \beta_{32},\ \ldots\beta_{3J}]^{\mathrm{T}},$$

The water balance equation (5) is substituted out to obtain

$$V(t) = V(0) + \sum_{n=1}^{t} \mathbf{B} w_d(n,\tau) - \sum_{n=1}^{t} w(n) + \sum_{n=1}^{t} \xi(n) \quad (14)$$

where terms such as $w_j(n-\tau_j)$ for $n-\tau_j \le 0$ in (14) are water discharges from the previous scheduling cycle, and are considered as given. By substituting (14) into (13), the Lagrangian can be rewritten as

$$L_h = \sum_{t=1}^{T-1}[\beta_1^{\mathrm{T}}(t)\underline{V} - \beta_2^{\mathrm{T}}(t)\overline{V}] + \beta_3^{\mathrm{T}}[V^T - V^0 - \sum_{n=1}^{T}\xi(n)]$$
$$+\sum_{t=1}^{T-1}\{(\beta_2^{\mathrm{T}}(t) - \beta_1^{\mathrm{T}}(t))[V^0 + \sum_{n=1}^{t}\xi(n)]\} + \sum_{j=1}^{J} L_{hj}, \quad (15)$$

where

$$L_{hj} \equiv \sum_{t=1}^{T}[-\lambda(t)p_{hj}(w_j(t)) - \mu(t)r_{hj}(w_j(t))]$$
$$-\sum_{t=\tau_j+1}^{T} w_j(t-\tau_j)b_j^{\mathrm{T}}\beta_3 + \sum_{t=1}^{T} w_j(t)e_j^{\mathrm{T}}\beta_3$$
$$+\sum_{t=1}^{T-1}\{[\sum_{n=1}^{t} w_j(n-\tau_j)]b_j^{\mathrm{T}}(\beta_2(t) - \beta_1(t))\}$$
$$+\sum_{t=1}^{T-1}\{[\sum_{n=1}^{t} w_j(n)]e_j^{\mathrm{T}}(\beta_1(t) - \beta_2(t))\}$$
$$+\sum_{t=1}^{T} S_{hj}(x_{hj}(t)), \quad (16)$$

e_j is a unit vector, i.e., its j th element is one and all the other elements are zeros; b_j is the j th column vector of connection matrix B. With multipliers λ, μ, β_1, β_2 and β_3 given, the sub-Lagrangian in (16) is unit-wise decomposable.

Suppose that reservoir with index j_d is direct downstream to reservoir j, then the subproblem for unit j is described as

$$\min_{w_j(t),u_{hj}(t)} L_{hj}, \text{ with } L_{hj} = \sum_{t=1}^{T}[h_j(w_j(t)) + S_{hj}(x_{hj}(t))]$$

where

$$h_j(w_j(t)) \equiv -\lambda(t)p_{hj}(w_j(t)) - \mu(t)r_{hj}(p_{hj}(w_j(t)))$$
$$+\beta_{3j}w_j(t) + \{\sum_{n=t}^{T-1}[\beta_{1j}(n) - \beta_{2j}(n)]\}w_j(t)$$
$$-\beta_{3j_d}w_j(t) + \{\sum_{n=t+\tau_j}^{T-1}[\beta_{2j_d}(n) - \beta_{1j_d}(n)]\}w_j(t),$$
$$t = 1,2\ldots T-1, \quad (17a)$$

and

$$h_j(w_j(T)) \equiv -\lambda(T)p_{hj}(w_j(T)) - \mu(T)r_{hj}(p_{hj}(w_j(T)))$$
$$+\beta_{3j}w_j(T), \quad (17b)$$

subject to its individual constraints (5-9). Note that (17) contains multipliers associated with unit j and its direct downstream unit only. A hydro unit is thus coordinated with other units within the same river catchment by the multipliers associated with its direct downstream unit.

Solving an individual unit subproblem

To model the discrete dynamics, the concept of "operating state" is introduced following what was used for thermal units with minimum up/down constraints. A state for a hydro unit is defined as the number of hours that the unit has been up (positive) or down (negative). Since the unit can be kept on or shut down after it has been up for $\overline{\varsigma}_j$ hours, the number of up states needed is the minimum up time. Similarly the number of down states is the minimum down time. By combining the above analysis, the state transition diagram can be constructed as in Fig. 1, where each node represents a state, and start-up

and shut-down costs are associated with edges.

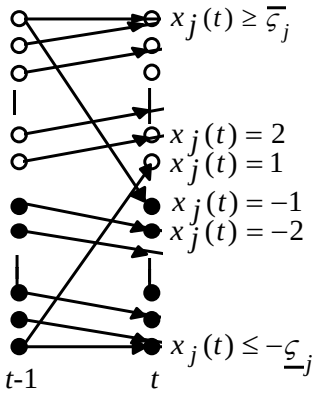


Fig. 1 The state transition diagram

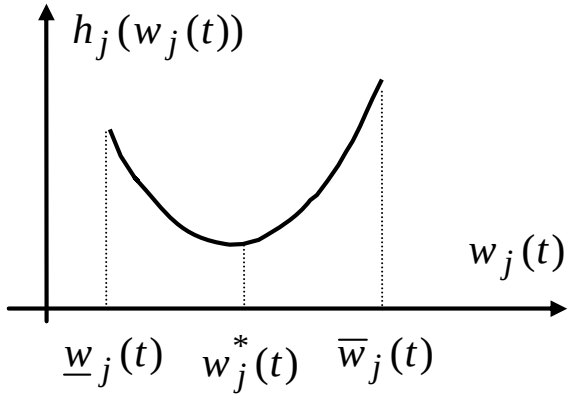


Fig. 2 Function $h_j(w_j(t))$

Based on the water-power conversion function $p_{hj}(w_j(t))$ and reserve contribution $r_{hj}(p_{hj}(w_j(t)))$, the stage-wise cost function $h_j(w_j(t))$ in (17) is depicted in Fig. 2. The optimal water discharge at time t for a particular operating region can then be obtained by

$$w_j^*(t) = \text{argmin } h_j(w_j(t)), \quad (18)$$

subject to the range of the operating region.

After the optimal generation level for each operating region has been obtained for each hour, the associated cost function $h_j(w_j(t))$ can be calculated. Based on the state transition diagram of Fig. 1, dynamic programming can then be applied to optimize states across hours as in [20, 22]. The optimal water discharge over the entire scheduling horizon can thus be obtained without discretizing reservoir levels.

Updating the multipliers

A modified subgradient method with adaptive step sizing is applied to update the multipliers $\beta_1, \beta_2, \beta_3$ associated with reservoir level limits at the intermediate level, and multipliers λ, μ associated with system demand and reserve requirements at the high level. This method has been described in detail in [20-22]. The subgradients for $\beta_1, \beta_2, \beta_3$ are

$$g_{\beta_1(t)} = \underline{V} - V(t), \quad (19)$$

$$g_{\beta_2(t)} = V(t) - \overline{V}, \quad (20)$$

and

$$g_{\beta_3} = V^T - V(T), \quad (21)$$

respectively. It should be noted that these multipliers are needed for the hydro subproblem only.

Obtaining Feasible Solutions

The subproblem solutions obtained from Lagrangian relaxation are generally infeasible, i.e., the relaxed constraints (2, 3, 5, 6b) may not always be satisfied. To obtain a good feasible solution, a feasible hydro schedule is first obtained by using a nonlinear network flow algorithm with the operating states obtained in the dual solution. However, too many idle hours of a unit may cause forced spillage, and too many up hours may result in no feasible schedule even with minimum generation. In this case, a heuristic is developed to adjust operating states as in ([23]). The cost increase of changing the state for each hour is calculated to provide quantitative information. For example, if changing from a down state to an up state is necessary, the hour with the minimum cost increase while satisfying the minimum down time will be selected.

With the feasible hydro schedule fixed, the thermal and pumped-storage units are adjusted to meet the system demand and reserve requirements as in [20, 22].

Implementation and Testing Results

The algorithm is implemented in FORTRAN on HP 715/33 workstations. Numerical testing is performed using modified billing data sets from Northeast Utility Service Company (NU). There are about 70 thermal units, 1 large pumped-storage unit, and 7 hydro units belonging to one river catchment as shown in Fig. 3. The system features have been presented in [20]. Two sets of data were selected from the NU billing database and modified to form a cascaded river catchment.

Testing results are summarized in Table 1. Cases 2 and 4 are obtained by using the method developed in this paper (Lagrangian relaxation + network flow or LNF), and are compared with Cases 1 and 3 obtained by using a pure network flow algorithm (NF). As stated in Section I, the NF algorithm cannot handle hydro units' minimum down time (MDT), and the minimum water discharge has to be set to 0. It can be seen that the costs of the LNF are no more than 0.7% of the corresponding NF costs, which in fact are lower bounds to optimal costs in view of the absence of MDT and zero minimum water discharge. The schedules obtained by LNF are therefore near optimal. The CPU times are in the range of a few minutes on a low-end workstation, efficient for daily scheduling.

The generation levels of Reservoir A in Case 1 and Case 2 are presented in Fig. 4. For a better view, only one day is

shown. Violations of MDT and minimum discharge by using the NF algorithm are observed at hour 7 and hour 8. It can also be observed that the schedule produced by LNF algorithm does satisfy the MDT and minimum discharge constraints.

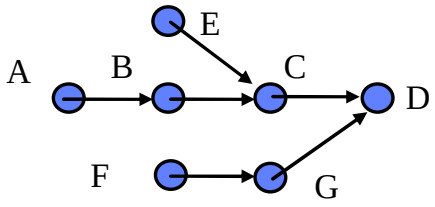


Fig. 3 Reservoirs in the river catchment

Table 1 Testing Results

Date set	Jan. W3, 1991		Oct. W2, 1991	
Case	1	2	3	4
MDT(hr)*	1, 1, 1	6,4,2	1, 1, 1	6,4,2
Cost($)	4,654,993	4,663,640	4,441,612	4,450,913
IT	47	54	45	65
CPU(s)	460	502	426	540

* MDT for Reservoirs A, D and F.
IT: number of high level iterations

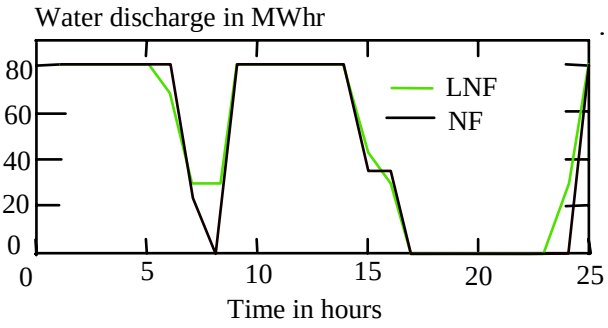


Fig. 4 Water discharge of reservoir A

Conclusions

An optimization-based algorithm has been presented for scheduling hydro units with cascaded reservoirs and discrete hydro constraints within the Lagrangian relaxation framework. The algorithm can systematically deal with discontinuous operating regions and discrete operating states without discretizing reservoir levels. Numerical testing results based on a practical system show that the algorithm is computationally efficient, and near optimal schedules are obtained.

References

[1] A. Cohen and V. Sherkat, "Optimization-Based Methods for Operations Scheduling," *Proceedings of IEEE*, Vol. 75, No. 12, 1987, pp. 1574-1591.

[2] J. J. Shaw and D. P. Bertsekas, "Optimal Scheduling of Large Hydrothermal Power Systems," *IEEE Transactions on Power Apparatus and Systems*, Vol. PAS-104, 1985, pp. 286-293.

[3] L. A. F. M. Ferreira, T. Anderson, C. F. Imparato, T. E. Miller, C. K. Pang, A. Svoboda, A. F. Vojdani, "Short-Term Resource Scheduling in Multi-Area Hydrothermal Power Systems," *Electric Power & Energy Systems*, Vol. 11, No. 3, 1989, pp. 200-212.

[4] A. Renaud, "Daily Generation Management at Electricite de France: From Planning Towards Real Time," *IEEE Transaction on Automatic Control*, Vol. 38, No. 7, 1993, pp. 1080-1093.

[5] S. J. Wang, S. M. Shahidehpour, D. S. Kirschen, S. Mokhtari, and G. D. Irisarri "Short-Term Generation Scheduling with Transmission Constraints Using Augmented Lagrangian Relaxation," *IEEE Transactions on Power Systems*, Vol. 10, No. 3, Aug. 1995, pp. 1294-1301.

[6] A. I. Cohen and S. H. Wan, "An Algorithm for Scheduling a Large Pumped Storage Plant," *IEEE Transaction on Power Apparatus and Systems*, Vol. PAS-104, No. 8, Aug. 1985, pp. 2099-2104.

[7] W. J. Trott and W. Yeh, "Optimization of Multiple Reservoir Systems," *Journal of Hydraulics Division (ASCE)*, Oct. 1973.

[8] H. Brannlund, J. A. Bubenko, D. Sjelvgren and N Anderson, "Optimal Short Term Operation Planning of a Large Hydro-Thermal Power System Based On a Nonlinear Network Flow Concept," *IEEE Transactions on Power Systems*, Vol. 1, No. 4, 1986, pp. 75-82.

[9] R. E. Rosenthal, "A Nonlinear Network Flow Algorithm for Maximization of Benefits in a Hydroelectric Power System," *Operation Research*, Vol. 29, No. 4, pp. 763-786, July 1981.

[10] C. Li, P. Jap and D. Streiffert, "Implementation of Network Flow Programming to the Hydrothermal Coordination in an Energy Management System," *IEEE Transactions on Power Systems*, Vol. 8, No. 3, Aug. 1993, pp. 1045-1053.

[11] J. L. Kennington and R. V. Helgason, *Algorithms for Network Programming*, John Wiley & Son, 1980.

[12] H. Habibollahzadeh, D. Frances and U. Sui, "A New Generation Scheduling Problem at Ontario Hydro," *IEEE Transactions on Power Systems*, Vol. 5, No. 1, Feb. 1991, pp. 65-73.

[13] D. Sjelvgren and T. S. Dillion, "Optimal Operations Planning in a Large Scale Hydro-Thermal Power System," *IEEE Transaction on Power Apparatus and Systems*, Vol. PAS-102, No. 11, Nov. 1983.

[14] M. Christoforidis, B. Awobamise, R. J. Frowd and F. A. Rahimi, "Short-Term Hydro Generation and Interchange Contract Scheduling for Swiss Rail," *Proceedings of the 1995 IEEE Power Industry Computer Applications Conference*, Salt Lake City, UT, USA, 1995, pp. 143-149.

[15] P. Chen and H. Chang, "Genetic Aided Scheduling of Hydraulically Coupled Plants In Hydro-Thermal

Coordination," 1995 IEEE/PES Summer Meeting Portland, OR, USA, July 1995, 95 SM 570-2.

[16] Slobodan Ruzic, N. Rajakovic, "A Flexible Approach to Short-Term Hydrothermal Coordination Part I & Part II," 1995 IEEE/PES Summer Meeting, San Francisco, CA, USA, July 1995, 95 SM 626-2 and 95 SM 625-4.

[17] C. Li, E. Hsu, A. Svoboda, C. Tseng and R. Johnson, "Hydro Unit Commitment in Hydro-Thermal Optimization," 1996 IEEE/PES Summer Meeting Denver, CO, July 1997, 97 SM 497-8.

[18] O. Nilsson and D. Sjelvgren, "Mixed-Integer Programming Applied to Short-Term Planning of a Hydro-Thermal System," *Proceedings of the 1995 IEEE Power Industry Computer Applications Conference*, Salt Lake City, UT, USA, 1995, pp. 158-163.

[19] L. A. F. M. Ferreira, "Short-Term Scheduling of a Pumped Storage Plant, "*IEE Proceedings*, Part C, Vol. 139, No. 6, 1992, pp. 521-528.

About Authors

Mr. Manoj Kumar, has been working as an Associate Professor at Velega Nageswar Rao College of Engineering, Ponnur, Guntur (Dist.), A.P., India. He is having about 7 years of Teching and research experience in Electrical Engineering with power system specialization. His research interests are power system optimization, power system operations and control.

Dr. B. Brahmaiah, has been working as Principal at Priyadarshini Institute of Technology, Srirama Chandrapuram, Tirupati, A.P., India. He has over three decades of Teaching and Research experience.

Dr. A. Srinivasula Reddy is working, presently as a Professor in the Dept. of EEE & Principal, Samskruti College of Engineering and Technology, Ghatkesar, Hydarabad. He has about 17 years of teaching and research experience. His fields of interest are Power Systems, Field Computations in Electrical Machines and High Voltage Engineering.

Dynamic Decompression of Text File

Amit Jain

Sir Padampat Singhania University, Udaipur, Rajasthan, India

Abstract
Compression algorithms reduce the redundancy in data representation to decrease the storage required for text data. Various algorithm available in lossless compression, such as Huffman encoding, arithmetic encoding, the Lempel-Ziv (LZ) family, Dynamic Markov Compression (DMC), Prediction by Partial Matching (PPM), and Burrows-Wheeler Transform (BWT) based algorithms. Decompression is also required to retrieve the original data by lossless means. A compression scheme for text files coupled with the principle of dynamic decompression, which decompresses only the section of the compressed text file required by the user instead of decompressing the entire text file. Dynamic decompressed files offer better disk space utilization due to higher compression ratios compared to most of the currently available text file formats.

Keywords : Compression, Text file format, Dynamic Decompression, Portable Document Format, Compression Ratio.

Introduction

The Use of compression for storing text files has become inherent part of personal as well as commercial computing. The various compression applications available to perform compression and decompression functions. The text document is first compressed and then the entire document is decompressed when required. This has some implications such as the unnecessary use of disk space for storing the compressed document as well as uncompressed document at the same time. Another implication is that even though an end user may require only a part of the document, the entire document as a whole is decompressed.

The algorithm (and application) described in this paper addresses both of the above-mentioned problems associated with compression applications and readers. The algorithm performs compression of the text file and displays only the section of the text file required by the user in decompressed format [11]. Therefore it provides a better and efficient way of storing and reading the text files, saving unnecessary wastage of disk space.

Related Work

In [1] it has been shown that text file compression can be done by assigning 2 character and 3 character ASCII codes. It has also been shown that about 75% reduction in size is achieved by using it with gzip and bzip2. Also, the number of possible codes is 73680, which is lesser than the number of words in the scheme of compression highlighted in this paper. However, only the compression scheme has been highlighted. [1] does not specify a decompression scheme for dynamically decompressing data.

[9] Presented a variable length word-coding scheme, which allows the direct search of the compressed file without decompression of the entire file using a variant of the Boyer Moore algorithm. The compression technique used in [9] provides for efficient decoding of arbitrary portion of text as well as smaller vocabulary representation. In comparison, the compression technique used in this paper uses a much simpler algorithm for compression and searching operations are directly done on the compressed file without decompressing the file either. However, in addition to the features in [9], the compression algorithm used by us allows line wise decoding of the compressed file, which allows dynamic decompression of the file as required by the user.

Compression Principle

The Compression algorithm applied assigns a unique 3-character code to every word in the source file. The position of the codes in the compressed file corresponds to the position of the respective words in the source file, i.e. their order of occurrence is not changed. Codes assigned consist of combinations of lowercase and uppercase ASCII alphabets [5].

Numbers are not compressed and are retained in the compressed file in their original form.

The compression algorithm uses page markers by using a ~ at the start of each page. Each page is determined by assigning a specific number of lines and characters for each line. The page markers play an important role in the current dynamic decompression algorithm as the decompression function uses a page marker count to determine the page number, and then decompresses the required page.

[10] A sequence of spaces is coded by counting the number of spaces and prefixing the two-digit count with #. Newline characters are encoded as the character y. When either a # or y is encountered in the source file, they are prefixed with the character z to ensure that there is no ambiguity while decompressing. One letter and two letter words are coded by prefixing Y and X respectively to them. Words greater than 20 letters in length are not coded either and are prefixed with the character t and suffixed with #. For words with length >2 and <20, every code is in the form of $\alpha\beta\gamma$ where β and γ take values as from a to z and then from A to Z in sequence. The α value is determined from the Table I.

Table I: Alpha Values for Assigning Codes

Word Length	α value	Number of Codes
3	a, A	5408
4	b, B	5408
5	c, C	5408
6	d, D	5408
7	e, E	5408
8	f, F	5408
9	g, G	5408
10	h, H	5408
11	i, I	5408
12	j, J	5408
13	k, K	5408
14	l, L	5408
15	m, M	5408
16	n, N	5408
17	o, O	5408
18	p, P	5408
19	q, Q	5408
20	r, R	5408

An example sequence of codes assigned for 3 letter words would be αaa – αaz – αaA - αaZ - αba – αbz αza – αzz – αzA – αzZ – αAa – αAz – αAA – αAZ – αBa – αZa – αZz – αZA – αZZ initially with α = a and after the completion of the sequence with α=A.

The compressed file structure consists of two parts, the codes and the words. The coded form of the source file is stored in the compressed file first. This is followed by a ~ which denotes that the codes have all been stored. For the decompression algorithm to relate each word with its corresponding code, the words are first classified based on their length and then words of a each length are ordered based on the order of their first occurrence in the source file. The words are then stored at the end of the compressed file. Since codes are assigned based on the order of the occurrence, this is a reliable technique for the decompression algorithm to associate each code with its corresponding word.

Dynamic Decompression

The principle of Dynamic Data Decompression refers to the decompression of a compressed file when it is being viewed. The entire file is not decompressed, but only the page that is currently being viewed is decompressed. The rest of the file remains in the compressed state. [6] Hence better storage efficiency is obtained as the space required on the secondary storage media is reduced by a factor almost equal to the Compression Ratio of the algorithm used for compression. Given the source file size as S bytes with N pages and the compression ratio as C, the compressed file size = S/C bytes. The average page size in the source file = S/N bytes. The compressed file is then subject to dynamic data decompression. Therefore, average size occupied on disk is

(S / C) + (S / N) bytes (1)

This is a considerable decrease in size from the original S bytes. The dynamic data decompression functionality in its current implementation is provided by a User Interface front end and a decompression function that operates as the back end. When a compressed file is first opened, the first page of it is decompressed and displayed. Command buttons are provided on the user interface to view the next and previous pages. Whenever one of these buttons is clicked on, the User Interface control sends a page number as a parameter to the decompression function. The page specified by the page number is then decompressed by the decompression function, which identifies pages by maintaining a count of page markers encountered. Hence, the page requested only is decompressed. Therefore, at any point of time, the space occupied on disk is the sum of the compressed file size and the decompressed size of the page that is currently being viewed.

Decompression Function

The Decompression function is used to decompress the compressed text file. It takes a parameter, the page number of the page to be decompressed and decompresses only this page. It uses a combination of page markers and line markers to distinguish between pages. For every code in the compressed file, the decompression function refers Table 1 to determine the length of the corresponding word for the code. Once the length is determined, it also calculates the displacement from the initial code for words of that length. Based on this displacement value, it determines the corresponding word for the code. This is done by checking the words of the determined length at the end of the compressed file. Since the codes are assigned and words are stored based on the order of occurrence, the displacement value leads the decompression function to the required word.

An added advantage of storing codes is faster searching operations as each word is reduced to a three-character code. Hence, once the code for the word that has to be searched is determined by comparison at the end of the file, the searching operation speed is increased. For Example, even a 15 character word is reduced to a 3 character code and by just comparing the first character of each code, we would be able to determine if it is represents a 15 character word. In any other case, a comparison that extends till the 15th character is required. But in this case a maximum of 3 comparisons are required.

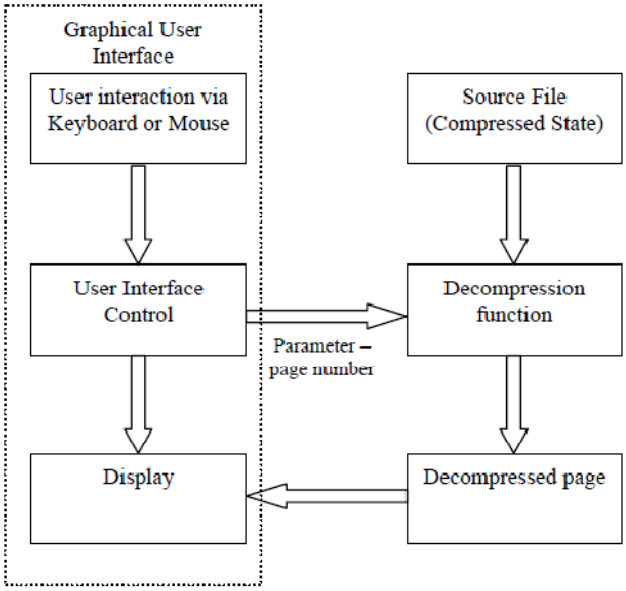


Fig. 1: A Block Diagram showing the interaction between various components of the dynamic decompression system.

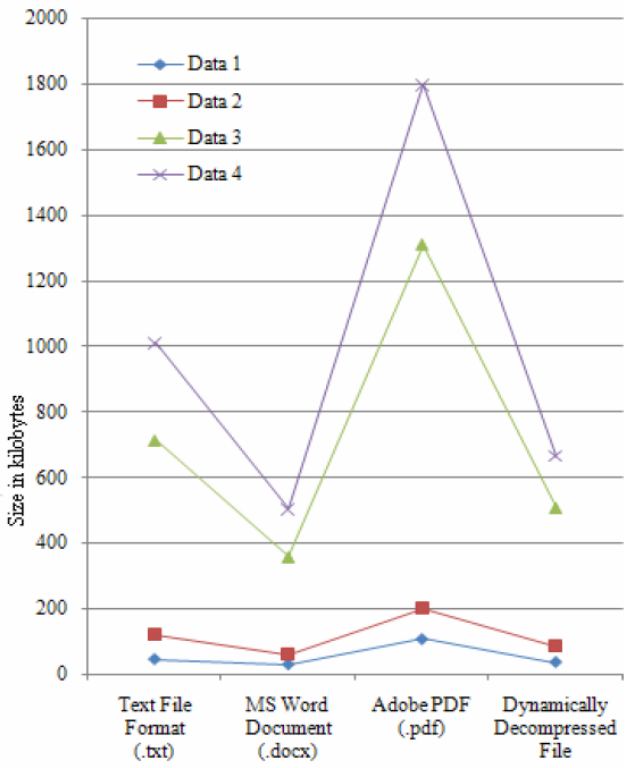


Fig. 2 Graph indicating size of data on disk stored in various files Formats

Experimental Results

For the purpose of comparison with other popular file formats for text file, experiments were conducted. The other file formats used were .docx (Microsoft Word 2007 document format), .pdf (Adobe Portable Document format), and .txt (Text file format). Every source data was converted into these formats and the size was observed.

Additionally, the source data was compressed using the compression principle that is discussed in this paper. Then, the compressed file was subjected to dynamic decompression and an average of the decompressed page size for a certain range of pages was computed. At any point of time, the size occupied on disk was taken to be the sum of the average decompressed page size and the compressed file size from Equation (1). This value was computed and documented.

Source data of four different sizes were taken and each of them were converted into the different file formats above and also compressed using the principle discussed in this paper and the value for each as per equation (1) were calculated and tabulated in Table II.

Table II: Size Comparison for Data in Different File Formats

All value in KB	Text file (.txt)	MS word (.doc)	Adobe pdf (.pdf)	Compressed File
Data1	48	32	109	39
Data2	120	60	202	85
Data3	714	359	1310	509
Data4	1012	505	1800	668

Conclusions

Dynamic decompression reduces the disk space needed to store a text file drastically. For compression applications currently available, there is a point in time wherein both the text file decompressed by the application and the compressed file exist simultaneously on disk. This leads to inefficient disk utilization. The dynamic decompression scheme addresses this anomaly by having only the coded file and only a section of file required by the user decompressed on the disk at any point in time. Therefore, the space occupied on the disk can be reduced significantly. This feature alone makes the developed reader very suitable for low-end consumers as well as in commercial industry where the disk space is of great value and in comparison to other text file formats currently available, offers much more compression ratio.

Since the compression scheme, even though providing compression more than the most of the available text file formats, doesn't quite compress the text files as effectively as other stand alone compression applications available, there is still scope for improvement in the encoding function that we have used. One way in which it can be achieved is by encoding the words based on their frequency of occurrence in the text file and also doing the encoding at the bit level.

The compression time taken by the application can be reduced by using better data structures available such as a BTree, and also by using binary search when searching for a keyword in a list. Since the Graphical User Interface that we have developed currently uses page markers and line markers to decompress pages according to need, we can solve the problem of Graphical User Interface window resizing, where

in the number of characters in a line increase or decrease when resizing a window, by having a character count for a particular window size instead.

A further improvement can be made to the application, by providing encryption standards along with the features of dynamic decompression. An editor can also be integrated to the developed reader where in a user can edit the page being viewed and it gets correctly encoded again after he navigates to another page.

References

[1] Md. Ziaul Karim Zia, Dewan Md. Fayzur Rahman, and Chowdhury Mofizur Rahman. "Two-Level Dictionary-Based Text Compression Scheme". Proceedings of 11th International Conference on Computer and Information Technology.

[2] Behrouz A. Forouzan and Richard F. Gilberg, Computer Science A Structured Programming Approach Using C, Thomson, 2003

[3] Data Structures using C, Aaron M. Tenenbaum, Yedidyah Langsam and Moshe J. Augenstein, Pearson Education, 2006.

[4] Michael J. Folk, Bill Zoellick, Greg Ricardi. File Structures-An Object Oriented Approach with C++, Addison-Wesley, 1998

[5] B.S. Shajeemohan and V.K.Govindan, Intelligent Compression Scheme for Faster and Secure Transmission of Text and Image Data over Internet, International Conference on Human Machine Interface ICHMI 2004

[6] Marc L. Corliss , E. Christopher Lewis , Amir Roth, The implementation and evaluation of dynamic code decompression using DISE, ACM Transactions on Embedded Computing Systems (TECS), v.4 n.1, p.38-72, February 2005.

[7] R. Franceschini, H. Kruse, N. Zhang, R. Iqbal, and A. Mukherjee, "Lossless, Reversible Transformations that Improve Text Compression Ratio," Project paper, University of Central Florida, USA. 2000.

[8] U. Manber, "A Text compression scheme that allows fast searching directly in compressed file," ACM Transactions on Information Systems, Vol.52, N0.1, pp.124-136, 1997.

[9] "A Scheme That Facilitates Searching and Partial Decompression of Textual Documents. Ashutosh Gupta . Intl. Journal of Advanced Computer Engineering, Volume 1, No 2, pages 99 -109, 2008.

[10] F. Awan, N. Zhang, N. Motgi, R. Iqbal, and A. Mukherjee, "LIPT: A Reversible Lossless Text Transform to Improve Compression Performance," Proceedings IEEE Data Compression Conference, pp. 481-210, 2001.

[11] D. A. Huffman, "A method for the construction of minimum redundancy codes," In Proc. IRE 40, volume 10, pages 1098–1101, September 1952.

[12] Terry A. Welch, "A Technique for High Performance Data Compression," IEEE Computer, Vol. 17, pp. 8-19, June 1984.

Evolutionary Algorithm based Optimal Location of Facts Devices

[1]S. Mohammad Rafee and [2]Dr A. Srinivasula Reddy

[1]Electrical Engg. Dept. Samskruti College of Engg & Tech, Hyderabad, India
E-mail: mdrafee1980@gmail.com
[2]Principal for Samskruti College of Engg. & Tech, Hyderabad, India
E-mail: svas_a@ rediffmail.com

Abstract
With the increasing size of power system, there is a thrust on finding the solution to minimize the utilization of existing system and to provide adequate voltage support. Flexible AC transmission system(FACTS) if placed optimally can be effective in providing voltage support, controlling power flow and in turn resulting in to lower losses. The algorithm to find the optimal location of TCSC and STATCOM based on GA has been devoleped. The effect of these devices on line flows and bus voltage profile has benn studied by placing at random with optimal ratings dictated by GA. This has been implemented on 5 bus and 30 bus system.

Index Terms: facts, genetic algorithm, tcsc, statcom, optimal location.

Introduction
A Flexible Alternating Current Transmission System is a system comprised of static equipment used for the AC transmission of electrical energy. It is meant to enhance controlalability and increase power transfer capability of the network. The conventional solutions such as capacitor, reactor, and phase shifting transformers are manually less expensive than FACTS devices but limited in their dynamic behavior and are less optimal. The concept of FACTS and FACTS controllers was first defined by Hingorani, 1988 in [2,3]. FACTS can provide versatile benefits to transmission utilities such as control of power flow, increasing capabilities of lines to their thermal liomits, reducing loop flows, providing greater flexibility [9-10]. The value of FACTS application lies mainly in the ability of the transmission system to efficiently transmit power or to transfer under contingency conditions [6]. Thyristor controlled series compensator is a variable impedance series compensator, which controls the effective line reactance by connecting a variable reactance in series with the line [10-11]. STATCOM is a second generation FACTS device used for shunt reactive compensation. It is applied to improve voltage security, provide interface with the real power source, higher response to system changes and mitigation of harmonics [12-13].

Radman and Raje discussed power flow calculation of power system with multiple flexible AC transmission system by modifying and adding new entries in Jacobian equation because of major FACTS controllers i.e. STATCOM, SSSC and UPFC [14].

Problem formulation
The main objective is to determine the optimal location of the FACTS devices (TCSC and STATCOM) in the power network to minimize the losses, the following performance index is selected.

Min $F_l = \sum_{k=1}^{n+1} P_{l+l}$

F_l is the objective function to minimize active power losses

P_{ltl} is the active power loss in tl^{th} line

N_{tl} is the number of transmission lines in the system.

This objective function subjected to constraints

$Pi^{min} \le Pi \le Pi^{max}$ i = 1,2,……ng

$Q_i^{min} \le Q_i \le Qi^{max}$ i= 1, 2,……ng

$Vi^{min} \le Vi \le Vi^{max}$ i = 1,2……..n

$\delta i^{min} \le \delta i \le \delta i^{max}$

For TCSC

$-0.7\ X_l \le X_{TCSC} \le 0.2\ X_l$

For STATCOM

$Q_{S\,min} \le Q_S \le Q_{S\,max}$

Where Q_S is the reactive power injected by the STATCOM in to the system.

$Q_{S\,min}$ is the minimum limit of rective power injected.

$Q_{S\,max}$ is the maximum limit of reactive power injected

ng = number of buses

X_{TCSC} = reactance of TCSC

X_l = reactance of line

Proposed methodology
Genetic Algorithms operate with a set of strings instead of a single string. This set of strings is known as a population and put through the process of evolution to produce new individual strings. The population size depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions. Initial population is generated on the basis of population size and string length.

For 5 bus system , initialization is done as follows

Step 1 : First generate random values for the TCSC reactance in the working range ($-0.7\ X_l < TCSC < 0.2\ X_l$)

Step 2 : Then generate a set of random number equal to number of transmission lines consists of 0's and 1's.

Step 3 : Multiply the values of TCSC reactance or STATCOM reactance generated with the set of random numbers generated in step 2.

Fitness Function Calculation

After initialization, the fitness is evaluated for each individual of the population. The fitness function for loss minimization problem can be expressed as

$$F(X) = \frac{1}{1+Fl}$$

As it is easy to find the maximum value of objective function using GA, so inverse function is selected to convert the objective function in to a maximum one.

Reproduction

This is a process where the individual is selected to move to a new generation according to its fitness. In this problem formulation Roulette-Wheel selection criterion is used and detailed as follows.

Each individual in the population has its interval. The size of each interval corresponds to the fitness of the individual and can be defined as

$P_i = F_i / >. F_i$

Where P_i = probability of selecting i^{th} string

F_i = The fitness of i^{th} individual

Crossover

Crossover is applied after reproduction. Purpose of crossover is to exchange information among the strings of mating pool to create new strings. Crossover rate P_c is the probability of cross over and varies from 0 to 1. Typically cross over rates vary from 0 to 1. Typically crossover rate s vary from 0.7 to 1 for a population of 30 to 200.

Mutation

Mutation is used to introduce some sort of diversification in the population in the population to avoid premature convergence to local optimum. Mutation involves just flipping of the bit for binary structure i.e. 1 to 0 or 0 to 1 and alteration of number in case of real value structure. Mutation rate is the probability of mutation. Typically mutation rate varies from 0.001 to 0.5 for a population size of 30 to 200.

Flow chart

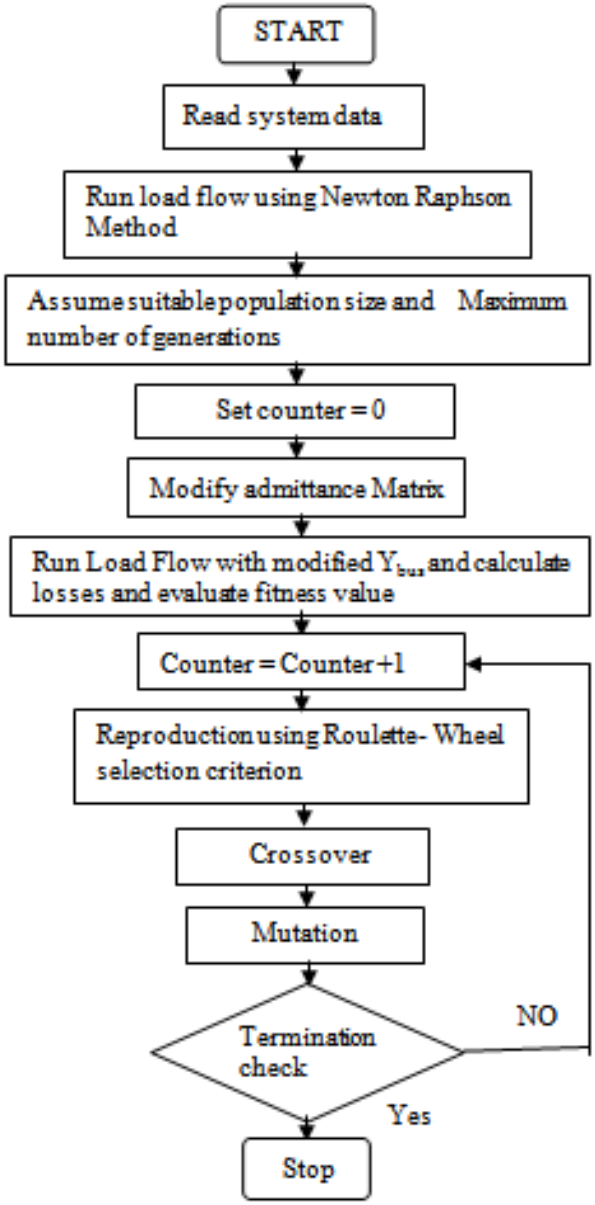


Fig: Flow chart for optimal placement of TCSC or STATCOM

Results and discussions

Genetic algorithm parameters:

Population size = 100, maximum no of generations = 50,
Crossover probability = 0.9, mutation probability = 0.1

Case 1.

Optimal location of FACTS devices by using GA has been implemented on 5 bus system.

TCSC placement: Here TCSC is placed in line 1-3 and the results are shown below.

Result analysis of 5 bus system with and without TCSC

Bus code	X_{TCSC}	Active power without TCSC (pu)	Active power with TCSC (pu)
1-2	0.0732	0.8933	0.8865
2-3	-0.0699	0.2447	0.2459
2-5	-0.0748	0.5466	0.5531

Line No	X_{TCSC}	Active power without TCSC(pu)	Active power with TCSC (pu)
2-5	-0.0536	0.8241	0.9155
9-11	-0.1429	0	0.0000
15-18	-0.0298	0.0653	0.0667

Voltage and angle with TCSC using GA

Bus No.	Voltage magnitude (pu) without (TCSC)	Angle (degree) without TCSC	Voltage magnitude (pu) with TCSC	Angle (degree) with TCSC
1	1.06	0.000	1.06	0.00
2	1.00	-2.0612	1.00	-2.2056
3	0.9872	-4.6367	1.00	-4.9572
4	0.9841	-4.9570	0.9944	-5.2434
5	0.9717	-5.7649	0.9752	-5.9580

Result analysis of 5 bus and 30 bus system for real power loss

S. No.	Total Real Power loss without FACTS devices (pu)	Total Real Power loss with TCSC	Total Real Power loss with STATCOM
For 5 bus system	0.0612	0.0599	0.0606
For 30 bus system	0.1776	0.1771	0.1770

STATCOM placement:
STATCOM is placed at bus 4 . STATCOM works in voltage control mode and it improves the voltage profile of the bus. With the placement of STATCOM in the system, there is improvement in the real power losses of the system, there is reduction of 0.5% in the losses.

Voltage and angle with STATCOM using GA

Bus No.	Voltage magnitude (pu) without (STATCOM)	Angle (degree) without STATCOM	Voltage magnitude (pu) with STATCOM	Angle (degree) with STATCOM
1	1.06	0.000	1.06	0.00
2	1.00	-2.0612	1.00	-2.0543
3	0.9872	-4.6367	1.00	-4.8403
4	0.9841	-4.9570	0.9944	-5.1094
5	0.9717	-5.7649	0.9752	-5.798

STATCOM data

STATCOM bus	E_p	P(pu)	Q_{sh} (pu)
3	1.0205	-4.9577	-0.2049

Case 2:
Optimal location of FACTS devices by using GA has been implemented to 30 bus system.

Result analysis of 30 bus system with and without TCSC using GA
For the placement of TCSC in 30 bus system using GA, the maximum no of TCSC used are four and out of four three are optimally plced by GA

Conclusions
The optimal placement of FACTS controllers has been attempted using GA. The study is carried out on 30 bus and 5 bus system. From the study following conclusions are drawn.
The developed algorithm is effective in deciding the placement of FACTS dev ices
TCSC helps in diverting flow from heavily loaded line and results in reduction in active power losses.

Future Scope
The allocation can be carried out by accounting the cost of FACTS devices and other economic considerations
The speed can be enhanced by reducing search space by incorporating some sensitivity index.

References

[1] Y.H. Song and A.T. Johns, Flexible AC Transmission system, IEE Press London, 1999.
[2] N.G. Hingorani and L.Gyugi, " Understanding FACTS- concepts and technology of Flexible AC Transmission systems", Standerd Publishers distributors, IEEE press, New York, 2001.
[3] A.A. Edris, R.Aapa, M.H. Baker, L. Bohman, K. Clark, "proposed terms and definitions for Flexible AC Transmission Systems", IEEE Transctions on power Delivery, vol.12, no.4, pp. 1848-1853, 1997.
[4] E. Alba and M.Tomassini, " Parallelism and Evolutionary Algorithm", IEEE Transctions on Evolutionary Computations, vol.6,no.5, pp.443-462, 2002.
[5] D.E. Goldberg, " Genetic Algorithms in search Optimization and Machine Learning", Addison-Wesley publishers, 1999.
[6] D.Gothm and G.T. Heydt, " Power flow control and

Power flow studies for systemwith FACT devices", IEEE transction on power systems, vol.13, no.1, pp.60-65, 1998.

[7] R. Rajarama, F. Alvardo, R. Camfield and S. Jalali, " Determination of location and amount of series compensation to increase power transfer capability", IEEE transactions on power systems, vol. 13, no.2, pp. 294-299, 1998.

[8] M. Noorzian and G. Anderson ," Power flow control by use of controllable series components", IEEE Transctions on power systems, vol.8, no.3, pp. 1420-1429, 1993.

[9] J.Mutale and G.Strbac, ' Tranmission network reinforcement versus FACTS an assessment", IEEE Transctions on power system, vol.15, no.3, pp.961-967, 2000.

[10] M. Noorzian, L. Angquis and G. Anderson, " Improving power system dynamics by series-connected FACTS devices", IEEE Transctions on power delivery, vol.1, no.4, pp. 1635-1641, 1997.

[11] R.Rajaraman, F.Alvarado, " Determination of location and amount of series compensation to increase the power transfer capability", IEEE Transctions on power systems, vol.13, no.2, pp.294-299, 1998.

[12] L. Gyugyi, C.D. Shuder and K. K. Sen , "Static synchronous series compensator a solid state approach to the series compensation of transmission line", IEEE Transctions on power delivery, vol.12, no.3, 1997.

[13] M.O. Hassan, S. J. Cheng and Z. A. Zkaria, " Steady – State modeling of static synchronous compensator and thyristor controlled series compensator for power flow analysis", information Technology journal, vol.8, issue 3, pp. 347-353, 2009.

[14] G. Radam and R.S. Raje, " Power flow model calculation for power system with multiple FACTS controllers", Electric power system research, vol.77, issue 12, pp. 1521-1531, 2007.

Authors Biography

S. Mohammad Rafee: He is persuing his PhD from JNTU Hyderabad . He has done his M.Tech from JNTU Anantapur , Andhra Pradesh India. Presently he is working in Samskruti College of Engineering and Technology as Associate professor in EEE Department. His area of interests are reactive power compensation, power quality. He also published papers in various international and national conferences.
E-mail : mdrafee1980@gmail.com

Dr. Srinivasula Reddy: He has done his PhD from JNTU Anantapur, Andhra Pradesh, India. Presently he is working as principal Samskruti College of Engineering and Technology. He published papers in various international journals, international conferences, national conferences to his credit. His area of interest is power systems, drives, FACTS devices.
Email: svas_a@rediffmail.com

Faster Algorithms for Real Time Data Base Updations using Deferrable Scheduling

[1]Rajesh Babu. Movva, [2]A.P.N.G. Krishna and [3]Bomma Manikanta

[1]Assitant Professor, [2]Student, [3]Student
E-mail: mrb.csebec@gmail.com, phalgunaanagani@gmail.com, manikantabomma1@gmail.com

Abstract
The deferrable scheduling algorithms are very impressive for minimizing real-time update transaction workload but suffer from its on-line scheduling overhead. In this paper, we propose two enlarged versions of deferrable scheduling fixed priority algorithms to reduce the on-line scheduling overhead. These algorithms produce a hyper period which can use endless times by satisfying the temporal constraints. The first one is Deferrable Scheduling with hyper period by Schedule Construction. It searches the deferrable scheduling fixed priority schedule for a hyper period. The second one is deferrable Scheduling with hyper period by Schedule Adjustment. It adjusts the deferrable scheduling fixed priority schedule in an interval to form a hyper period. Both deferrable Scheduling with hyper period by Schedule Construction and deferrable Scheduling with hyper period by Schedule Adjustment can overcome the drawbacks in deferrable scheduling fixed priority algorithms, and deferrable Scheduling with hyper period by Schedule Adjustment works better than deferrable Scheduling with hyper period by Schedule Construction by performing more number of update transactions in the system.

Keywords: Real-Time databases, Temporal validity constraint, Fixed priority scheduling, Deferrable scheduling.

Introduction

Real-time embedded systems are important components of many time-critical applications that require timely processing of massive amount of real-time data. The correct functioning of a real-time embedded system depends not only on meeting real-time constraints of application jobs, but also on the accuracy of real-time data values sampled from real-world entities. Examples of real-time data include sensor data in sensor networks, positions of aircrafts in air traffic control systems, and vehicle velocity in adaptive cruise control applications. To provide better management of sampled real-time data and to support effective processing of application jobs, real-time data are typically managed by a real-time database system (RTDBS). passage of time since the status of the corresponding entity in the real-world may change continuously. Sensor update transactions should constantly sample real-world data values and install them into the RTDBS.

One efficient way to determine the correctness of real-time data in a RTDBS is to define a validity constraint or age constraint, which determines a validity interval length for each real-time data object. A real-time data value is only valid within its validity interval. For reliability reasons, real-time embedded systems require continuous generation of update transactions to refresh real-time data objects regardless of how much the status of the corresponding real-world entities have been changed. To meet this constraint, it is important to produce a schedule for all update transactions such that for any consecutive updates of a real-time data object, the next update is completed before the previous validity interval expires. Thus one of the crucial issues in the design of real time embedded systems is to schedule the update transactions efficiently to maintain the validity of real-time data while minimizing the total update transaction workload.

Existing Methods

Most of the previous work in update transaction scheduling assumed a periodic transaction model. The update problem for periodic update transactions consists of two parts.

1. the determination of the sampling periods and deadlines of update transactions; and
2. the scheduling of up-date transactions.

One of the proposed approaches is the Half-Half (HH) scheme in which an update transaction has a fixed period that is half of the validity interval. If the set of transactions is schedulable in HH, the validity constraints of the corresponding real-time data objects can also be guaranteed. Similarly to HH, More-Less (ML) is another periodic approach in which up-date transactions are scheduled based on the deadline monotonic algorithm Compared to HH, ML can guarantee the validity of real-time data objects with less update transaction workload. Recently, the DS-FP algorithm was proposed to further reduce the total update transaction work-load. The main idea of DS-FP is to adopt sporadic task model instead of the periodic model.

Compared to ML, DS-FP increases the separation of two consecutive trans-action jobs by releasing an update transaction job as late as possible based on the sampling time of its previous job. Both theoretical analysis and experimental results have demonstrated that DS-FP outperforms HH and ML significantly in reducing the update transaction workload while still maintaining the real-time data validity. One major problem of DS-FP is its time complexity for on-line schedule computation The variation in its run-time overhead leads to unpredictability of the system performance.

A real-time data object (X_i) at time t is temporally valid (or absolutely consistent) if, for its j th update finished latest before t , the sampling time ($r_{i,j}$) plus the validity interval (V_i) of the data object is not less than t , i.e., $r_{i,j} + V_i \geq t$.

A data value for real-time data object X_i sampled at any time t will be valid up to ($t + V_i$). The actual length of the temporal validity interval of a real-time data object is application dependent. One of the important design goals of RTDBS is to guarantee that real-time data remain fresh, i.e., they are always valid. We assume that the network delay for a sensor update transaction job to be sent from a sensor to the RTDBS (i.e., jitter between sampling time at the sensor and release time at the RTDBS) is zero for convenience of presentation.

More Les

ML adopts the periodic task model for sensor update transactions whose derived deadlines are not larger than their periods. Consider synchronous transactions whose first jobs all start at time 0. A time instant after which a transaction job has the worst-case response time is called a critical instant, e.g., time 0 is a critical instant for all the transactions with deadlines no larger than their periods if those transactions are synchronous. Note that we only consider synchronous transactions. In ML, there are three constraints to follow for transactions τ_i ($\forall$ i, $1 \leq i \leq m$) .

- Validity constraint: the sum of the period and relative deadline of transaction τi is always less than or equal to V_i , i.e., $P_i + D_i \leq V_i$.
- Deadline constraint: the period of an update transaction is assigned to be more than or equal to half of the validity length of its updated object, while its corresponding relative deadline is less than or equal to half of the validity length of the same object. For τi to be schedulable, D_i must be greater than or equal to Ci, the worst case execution time of τ_i , i.e., $C_i \leq D_i < P_i$.
- Schedulability constraint: for a given set of update transactions, the Deadline Monotonic scheduling algorithm is used to schedule the transactions.

DS-FP

DS-FP ML is pessimistic on the deadline and period assignment. This is because it uses a periodic task model that has a fixed period and relative deadline for each transaction, and the relative deadline D_i is equal to the worst-case response time of the transaction. According to the validity constraint in ML, the larger the deadline D_i , the smaller the period P_i . In order to increase the separation of two consecutive jobs (and thus reduce the sensor update workload), DS-FP adaptively derives the relative deadline and separation of one job from its previous job and preemptions from higher priority transactions. Given release time ri,j of job Ji,j and deadline $d_{i,j+1}$ of job $J_{i,j+1}$ ($j \geq 0$),

$$d_{i,j+1} = r_{i,j} + V_i \quad (1)$$

guarantees that the validity constraint can be satisfied, as depicted in Fig. 1. Correspondingly, the following equation follows directly from (1):

$$(r_{ij+1} - r_{ij}) + (d_{ij+1} - r_{ij+1}) = V_i \quad (2)$$

If $r_{i,j+1}$ can be shifted onward to $r'_{i,j+1}$ along the time line in Fig. 1, it does not violate (2). After the shift, temporal validity can still be guaranteed as long as $J_{i,j+1}$ is completed by its deadline $d_{i,j+1}$. The idea of DS-FP is to defer the sampling time (i.e., release time), $r_{i,j+1}$, of $J_{i,j}$'s next job as late as possible while still guaranteeing the validity constraint. According to the fixed priority scheduling theory, $r_{i,j+1}$ in DS-FP can be derived backwards from its deadline $d_{i,j+1}$ as follows:

$$(r_{ij+1}) = (d_{ij+1} - R_{ij+1})(r_{i,j+1}, d_{i,j+1}) \quad (3)$$

$$(R_{ij+1})(r_{i,j+1}, d_{i,j+1}) = \theta i(r_{i,j+1}, d_{i,j+1}) + C_i \quad (4)$$

where $\Theta_i(a, b)$ denotes the total cumulative processor demands made by all jobs of higher-priority transaction τ_k ($\forall k$, $1 \leq k \leq i - 1$) during time interval [a, b), and $R_{i,j+1}(r_{i,j+1}, d_{i,j+1})$ (or $R_{i,j+1}$ for simplicity in DS-FP) the response time of $J_{i,j+1}$ deriving backwards from its deadline $d_{i,j+1}$. Similarly to ML, DS-FP also assigns priorities to transactions according to SVF. Readers are referred to the Appendix for the details of the DS-FP algorithm. Now we summarize the algorithm as follows. First we set $r_{i,0} = 0, \forall i$, $1 \leq i \leq m$. The highest priority job among the outstanding jobs is always scheduled first. It is only preempted when a new job with higher priority is ready.

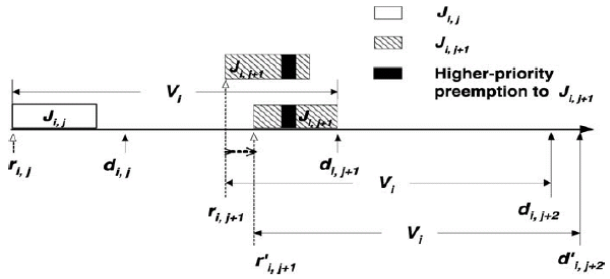


Fig 1. illustration of DS-FP scheduling.

As soon as a job $J_{i,j}$ is completed, we derive the $r_{i,j+1}$ of its next job according to above calculations. The algorithm fails when a job misses its deadline. Otherwise it keeps running. It is proved that any task set that is scheduled by ML is also scheduled by DS-FP. The EDL algorithm proposed in Chetto and Chetto processes tasks as late as possible based on the Earliest Deadline algorithm. EDL assumes that all deadlines of tasks are given whereas DS-FP and DESH algorithms derive deadlines dynamically. The validity constrained scheduling, e.g., ML, DS-FP, and DESH algorithms, are different from the distance constrained scheduling, which guarantees an upper bound to the finishing times of two consecutive instances of a task.

Next, we present two deferrable Scheduling with hyper period (DESH) algorithms for constructing periodic schedules off-line from the DS-FP algorithm so that the on-line scheduling time complexity can be reduced. Note that this will undoubtedly increase the space overhead for keeping the DESH schedules. But the space overhead can be kept reasonably low. Our DESH algorithms satisfy the following properties:

- Property 1 A schedule satisfies the validity constraint.
- Property 2 The on-line scheduling time complexity is O(1).

Proposed Methods

We propose two Deferrable Scheduling with Hyper period (DESH) algorithms, which construct the hyper period schedule off-line and reduce the on-line scheduling time complexity to O(1). The key contributions of our work include the followings.

1. To reduce on-line scheduling overhead, our first algorithm, named Deferrable Scheduling with Hyper period Construction (DESH-SC), searches the hyper period of the set of update transactions so that the transactions can be scheduled by repeating the hyper period schedule. However, the hyper periods found by DESH-SC are exponentially long, which could incur significant space overhead for maintaining the hyper period information for on-line scheduling.
2. Our second algorithm, named Deferrable Scheduling with Schedule Adjustment (DESH-SA), adjusts the DS-FP schedule in an interval such that the adjusted schedule can be repeated infinitely. DESH-SA improves DESH-SC on producing much shorter hyper periods and accommodating significantly more update trans-actions.
3. Our experimental results demonstrate that both DESH-SC and DESH-SA can reduce scheduling overhead of DS-FP, and DESH-SA outperforms DESH-SC by accommodating significantly more update transactions in the system.

Deferrable Scheduling with Hyper period: Schedule Construction (DESH-SC)

In this subsection, we present Deferrable Scheduling with Hyper period based on Schedule Construction (DESH-SC), a DS-FP based algorithm that can reduce the online scheduling overhead. The basic idea of DESH-SC is to search for an interval of DS-FP schedule, the hyper period, that could be repeated infinitely without violating the validity constraint. Note that DESH-SC could return without finding a hyper period. The DESH-SC algorithm consists of two parts: an algorithm for finding the hyper period off-line and an algorithm for scheduling transactions on-line. The latter is trivial once a hyper period is found because it only needs to repeat the hyper period schedule.

For a DS-FP schedule and a time period $[t_s,t_e]$we say $[t_s, t_e]$ is a hyper period for the transaction set if for all transactions τ_i $(1 \le i \le m)$, the following schedule satisfies τ_i 's validity constraint: it is the same as the DS-FP schedule from time 0 to te. From te onward, it repeats the DS-FP schedule in $[t_s, t_e]$ to infinity. Please note that t_s and t_e do not need to be idle time points in DESH-SA, which is different from the requirements of ts and te in DESH-SC.

Theorem 1. $[t_s, t_e]$ is a hyper period in DESH-SC if for all τi $(1 \le i \le m)$ the following conditions hold.

1. t_s and t_e are CPU idle time points.
2. $t_s > V_i$.
3. τ_i is scheduled at least once in $[t_s, t_e]$.
4. $I(t_s, \tau_i) \ge I(t_e, \tau_i)$, where function $I(t, \tau_i)$ is defined as the time distance between t and τ_i 's latest release time before t .

Proof Given any τ_i $(1 \le i \le m)$, we first prove in the hyper period schedule that the distance of the finish time of the first τ_i job after t_e and the release time of its latest job before t_e satisfies the validity constraint. Note that the first job after t_e repeats the first job in $[t_s, t_e]$. Because $t_s > V_i$, the first job of τ_i in $[t_s, t_e]$ must finish by $(t_s - I(t_s, \tau_i)) + V_i$, in other words, by $V_i - I(t_s, \tau_i)$ after t_s. Since $[t_s, t_e]$ is repeated after t_e, the first job of τ_i after t_e also finishes by $V_i - I(t_s, \tau_i)$ after t_e. The distance between the finish time of its first job in the second hyper period $[t_e, 2t_e - t_s]$ and the release time of its latest job in the first hyper period $[t_s, t_e]$ is no longer than:

$$I(t_e, \tau_i) + \big(v_i - I(t_s, \tau_i)\big) = v_i + \big(I(t_e, \tau_i) - I(t_s, \tau_i)\big) \le v_i \quad (5)$$

Similarly, we can prove that the distance between the finish time of its first job in the $(k + 1)^{th}$ $(k = 1, 2, \ldots)$ hyper period and the release time of its latest job in the k^{th} hyper period is no longer than V_i . Thus, the jobs of τ_i satisfy the validity constraint. To better satisfy the fourth condition, we need to assign ts to be the end of an idle period and te to be the beginning of another idle period. By idle period $[t_1, t_2]$ we mean that CPU is busy right before t_1, it idles between t_1 and t_2, and it is busy again after t_2. Once the hyper period is found, we could increase t_e as long as the conditions in theorem 1 satisfied.

The idea is presented in Algorithm 1. In the algorithm, we continuously push t_2 of idle periods into a queue Q as possible candidates for t_s of a hyper period. For each subsequent idle period, we check its t_1 against each t_2 saved in Q to see if they form a hyper period. If the hyper period is found, we could then further increase te as long as $[t_s, t_e]$ still satisfies the conditions in thorem 1. The increase cannot exceed $I(t_s, \tau_i) - I(t_e, \tau_i)$ for any transaction τ_i , $1 \le i \le m$. We define w to be the minimum of $I(t_s, \tau_i) - I(t_e, \tau_i)$ in the algorithm.

Algorithm 1. Search Hyper period:

Input: A DS-FP schedule, a utilization limit Umax, and a time limit tmax.
Output: The hyper period with utilization ≤ Umax.
U ←1.001; // Initialization of hyper period utilization.
t_2←max{V_i | $1 \le i \le m$};
$[t_1, t_2]$← first CPU idle period after t_2;
Append t_2 to Q; //Q is a FIFO queue of t_2.
while (U > Umax) do
$[t_1, t_2]$← next CPU idle period after t_2;
// tmax is the maximum time to search.
if ($t_1 > t_{max}$) then return failure; endif
t_e←t_1;
for t_s = first in Q to last in Q do
if ($[t_s, t_e]$ satisfies Condition 4)
then
Signal that a hyper period exists;

w←min{I (t_s, τ_i)−I (t_e, τ_i) | 1 ≤ i ≤ m};
t_e←t_e +min(w, (t_2 − t_1)); // Fine-tune te.
U′← utilization in [ts, te];
if (U > U′) then
U ←U′;
if (U ≤ U_{max}) then goto RTN;
endif
endif
endif
end
if Q is full then Dequeue the oldest; endif
Append t_2 to Q;
end

RTN: return [t_s, t_e] as the hyper period;

Deferrable Scheduling With Hyper Period: Schedule Adjustment (DESH-SA)

In this subsection, we present Deferrable Scheduling with Hyper period based on Schedule Adjustment (DESH-SA), a DS-FP based algorithm that can reduce the online scheduling overhead while achieving processor utilization close to that of DS-FP. By schedule adjustment, we mean changing release times and deadlines of jobs. The basic idea of DESH-SA is to construct a hyper period schedule SH off-line for T , a set of validity constrained transactions. Suppose the first hyper period of the SH schedule has length $\| S_H \|$. If the first hyper period of SH can be constructed by adjusting the DS-FP schedule in the time interval $[0, \| S_H \|]$,the complete S_H schedule can be constructed by repeating the first hyper period of S_H infinitely every $\| S_H \|$ time units.

Thus, similarly to DESH-SC, the DESH-SA algorithm consists of two parts: an algorithm for constructing the hyper period off-line and an algorithm for scheduling transactions on-line. We next describe how the first hyper period schedule of S_H in the interval $[0, \| S_H \|]$ is derived from the schedule of DS-FP. Given time $t_e > 0$, note that a DS-FP schedule in the interval [0, t_e] can be constructed off-line. Assume that jobs $J_{i,ki-1}$ and $J_{i,ki}$ of τ_i ($k_i \geq 1$ & $1 \leq i \leq m$) satisfy the following condition for t_e:

$$g(n_1 t_e, t) = g(n_2 t_e, t) \quad (7)$$

where g(nt_e, t) is a function returning a pair of integers ‹i, k›, which indicates that the k^{th} job of τ_i in the n^{th} hyper period is active at time t (i.e., at time nte +t). Equation (7) implies that any two hyper periods have the exactly same schedule.

$$g(nt_e, t) = \begin{cases} \langle i, j - n(k_i + 1)\rangle, & the\ CPU\ is\ allocated\ to\ J_{i,j}\ at\ time\ nt_e + t; \\ \langle 0,0\rangle, & the\ CPU\ is\ idle\ at\ time\ nt_e + t. \end{cases} \quad (8)$$

Note that n, j are integers, and n ≥0 & j ≥ 0 hold for (8). Equation (7) ensures that all transactions are released synchronously at time 0, t_e, $2t_e$, . . . , etc. If the processor is allocated to job $J_{i,j}$ at time nt_e + t , then it is the (j – n (k_i + 1))th job of τ_i from time nt_e (Note that there are (k_i +1) τ_i jobs during the interval [0, t_e]). Equations (7) and (8) ensure that the complete SH schedule is constructed periodically by repeating the schedule of the interval [0, t_e] every t_e units.

Theorem 2. Given a DS-FP schedule for a validity constrained transaction set T, suppose t_{idle} is an idle time in the schedule and the schedule before t_{idle} is feasible. Let r_i,k_i-1 ($k_i \geq 1$) be the latest release time of jobs of τ_i ($1 \leq i \leq m$) before t_{idle}. If ∀i ($1 \leq i \leq m$), holds, then the interval [0, t_{idle}] can be used as the first hyper period of the DESH-SA schedule without any adjustment. Proof Note that once a job is released under DS-FP, the processor cannot be idle until the job completes. Thus, if all jobs J_i,k_i are released at time t_{idle}, i.e., $r_i,k_i = t_{idle}$, then the schedule of the interval [t_{idle}, 2 * t_{idle}] is the same as that of the interval [0, t_{idle}]. Moreover, if (9) holds,

$$\begin{aligned} &\left(d_{i,k_i} - r_{ij}\right) = \left(d_{i,k_i} - t_{idle} + t_{idle} - r_{ik_i} - 1\right) \\ &= d_{i,0} - 0 + t_{idle} - r_{ik_i} - 1 \\ &\leq V_i \end{aligned} \quad (9)$$

That is, two consecutive jobs $J_{i,ki-1}$, J_i,k_i (∀i, $1 \leq i \leq m$) across two neighboring hyper periods satisfy the validity constraint. Thus a feasible DESH-SA schedule can be constructed by having the schedule of the interval [0, t_{idle}] as the first hyper period schedule.

Note that if te is set to be tidle, then it is not necessary to adjust the schedule of any transactions in the interval [0, t_{idle}] for making the first hyper period of DESH-SA.However, it is not always possible to find such a time tidle for all transactions satisfying (9), in which case the DS-FP schedule in the interval [0, t_e] corresponding to a subset of the transactions needs to be adjusted. Specifically, if transaction τ_h ($1 \leq h \leq m$) is the highest priority transaction whose schedule needs to be adjusted due to violation of (9), then the schedule of all lower-priority transactions τ_i ($h < i \leq m$) also needs to be adjusted due to the impact of release time and deadline adjustment of τ_i 's higher-priority transactions in the interval [0, t_e]. This is described in Algorithm 2.

Algorithm 2 AdjustScheduleForHyper period(T , t_e):

Input: Transaction set T and time $t_e > 0$.
Output: Adjusted schedule SH in [0, t_e] satisfying (7), and for all i, k_i .
Construct DS-FP schedule SH in [0, t_e] for all $\tau_i \in T$;
//$J_{i,j}$ has $r_{i,j}$, $d_{i,j}$ computed in SH ($j \leq k_i$ by (6)).
h←mini {i | τ_i belongs to T & τ_i violates (9)}.
//$J_{i,ki}$ is the latest τi job in the interval [0, t_e].
for all(i < h), k_i ←k_i −1;
// No adjustment for i < h
s ←t_e; // Schedule in [t_s, t_e] will be adjusted.
for i = h to m do
// Adjust SH in [0, t_e].
$d'_{i,ki}$ ←t_e;
j←k_i ;
// $d'_{i,j}$ is adjusted from $d_{i,j}$;
while (j > 0) do
if ($d'_{i,j}$ −$r_{i,j}$ <Θi($r'_{i,j}$, $d'_{i,j}$)+Ci) then
// $J_{i,j}$'s response time> $d'_{i,j}$ −$r_{i,j}$.
// $r'_{i,j}$ is adjusted from $r_{i,j}$.
$r'_{i,j}$ ←$d'_{i,j}$ − Θi ($r'_{i,j}$, $d'_{i,j}$)−C_i ;
if (((j < k_i) Λ ($d'_{i,j}$+1 − $r'_{i,j}$ > V_i))V
($r'_{i,j}$ < 0)) then report failure;

endif
if ($r'_{i,j} < d_{i,j}-1$) // Ripple impact.
then $d'_{i,j}-1 \leftarrow r'_{i,j}$; // Change $d_{i,j}-1$.
else $d'_{i,j}-1 \leftarrow d_{i,j}-1$; // No change.
endif
$j \leftarrow j - 1$;
else // No adjustment for this job.
if ($t_s \geq d'_{i,j}$) // No more adjustment for τ_i .
then
$t_s = d'_{i,j}$;
break; // Jump out of while loop
else
// Examine the previous job of τi .
$d'_{i,j}-1 \leftarrow d_{i,j}-1$;
// No change.
$j \leftarrow j - 1$;
endif
endif
end
if $((j = 0) \wedge ((d'_{i,j} < \Theta_i (0, d'_{i,j}) + C_i)))$
then report failure;
else $t_s \leftarrow 0$;
endif
end

RTN adjusted SH in $[0, t_e]$ and $\forall i$, k_i ;

Examples

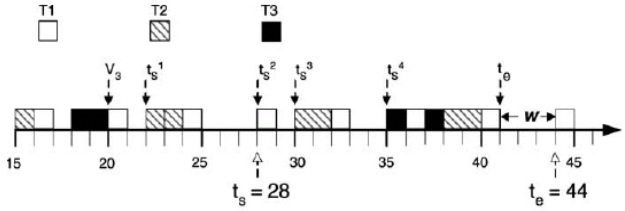


Fig 2. Desh-Schedule Construction

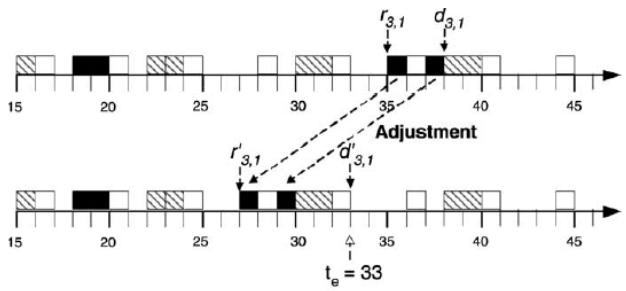


Fig 3. DESH-Schedule Adjustment

Performance Evaluation

This section presents important results from our simulation studies of the DESH algorithms. Our goal is to find out whether the DESH algorithms are effective for reducing the DS-FP overhead. The primary performance metrics used in our experimental studies are the CPU workload and the number of transactions supported in the system. In the experiments, we investigate whether DESH-SA and DESH-SC can find a hyper period, and if so, how much excess CPU workloads they may incur compared to DS-FP. We also compare the hyper period length of DESH-SA and DESH-SC to study the space efficiency of the approaches, and demonstrate the percentage of transactions to be adjusted when we calculate the hyper period for DESH-SA. For simplicity, only one version of a real-time data object is maintained. Upon refreshing a real-time data object, the older version is discarded.

We ignore the on-line scheduling overhead in our experiments, and consider it to be O(1) for all algorithms (which is true for DESH algorithms). This is in favor of DS-FP as its scheduling overhead is ignored for the CPU workload in our experiments. We define Nadjust to be the average number of jobs whose release times or deadlines are adjusted in $[0, t_e]$ under DESH-SA.

Conclusion

This paper presents new approaches originate from DS-FP, named Deferrable Scheduling with Hyper period by Schedule Construction and Deferrable Scheduling with Hyper period by Schedule Adjustment, that decreases the on-line scheduling overhead to O(1). Deferrable Scheduling with Hyper period by Schedule Construction searches the deferrable scheduling fixed priority schedule for a hyper period. Where Deferrable Scheduling with Hyper period by Schedule Adjustment adjusts the deferrable scheduling fixed priority schedule in an interval to form a hyper period. Deferrable Scheduling with Hyper period by Schedule Adjustment works better than Deferrable Scheduling with Hyper period by Schedule Construction by performing more number of update transactions in the system and it will also assures the age constraint. It is both space and time efficient. But there are some un answered questions like what is the necessary and sufficient condition is for the Deferrable Scheduling with Hyper period by Schedule Adjustment to produce a hyper period. We are going to address these problems in our future work.

References

[1] Burns A, Davis R (1996) Choosing task periods to minimise system utilisation in time triggered systems Inf Process Lett 58:223-229.
[2] Chen D, Mok AK (2004) Scheduling similarity-constrained real-time tasks. In: ESA/VLSI, pp 215–221.
[3] Chetto H, Chetto M(1989) Some results of the earliest deadline scheduling algorithm. IEEE Trans Softw Eng 15(10):1261-1269.
[4] Gerber R, Hong S, Saksena M (1994) Guaranteeing end-to-end timing constraints by calibrating intermediate processes. In:IEEE real-time systems symposium, December 1994.
[5] Gustafsson T, Hansson J (2004a) Datamanagement in real-time systems: a case of on-demand updates in vehicle control systems. In: IEEE real-time and Embedded technoly and applications symposium, pp

182-191.

[6] Gustafsson T, Hansson J (2004b) Dynamic on-demand updating of data in real-time databasesystems. In: ACM SAC.

[7] Han CC, Lin KJ, Liu JW-S (1995) Scheduling Jobs with temporal Distemce Constrints. SIAMJ Comput 24(5): 1104-1121.

[8] Kang KD, Son S, Stankovic JA, Abdelzaher T (2002) A QOS-Censitive approach for timelyness and freshness guarantees in Real-Time Databases. In: Euro Micro Real-Time Systems Conference, June 2002.

[9] Kuo T, Mok AK (1992) Real-Time Data Symantics and Similarity-Based Conqurency Control. In: IEEE Real-Time Systems Symphosium, Dec 1992.

[10] Kuo T, Mok AK (1993) SSP: A Symantics-Based Protocol for Real-Time Data Access. In: IEEE Real-Time Systems Symphosium, Dec 1993.

[11] Ho S, Kuo T, Mok AK (1997) Similaty-Based Load Adjustment for Real Time Data Intesive Applications. IN: IEEE Real-Time Systems Symphosium.

[12] Lam KY, Xiong M, Liang B, Guo Y (2004) Statistical Quality of Service Guarantee for Temporal Consistancy of Real-Time Data Objects. In: IEEE Real-Time Systems Symphosium.

[13] Leung J, Whitehead J (1982) On the Complexity of Fixed_Priority Scheduling of Periodic Real-Time Tasks. Perform Eval 2 : 237-250.

[14] Liu CL, Layland J (1973) Scheduling algorithms for multiprogramming in a hard real-time environment. J ACM 20(1).

[15] Locke D (1997) Real-Time Databases: Real-World Requirements. In: Bestavros A, Lin K-J, Son SH (eds) Real-Time Database Systems-Issues and Applications. Kluwer Acadamic, Dordrecht, pp 83-91.

[16] Ramamritham K (1996) Where Do Time Constrints come from and Where Do They Go? Int J Database Manag 7(2): 4-10.

[17] Song X, Liu JWS (1995) Maintaining Temporal Constistancy: Pessimistic vs. Optimistic Concurrency Control. IEEE Trans. Knowl Data Eng. 7(5): 786-796.

[18] Xiong M, Ramamritham K (2004) Deriving Deadlines and Periods for Real-Time Update Transactions. IEEE Trans. Comput 53(5): 567-583.

[19] Xiong M, Ramamritham K, Stankovic JA, Towsley D, Sivasankaran RM (2002) Scheduling Transactions with Temporal Constriants: exploiting data symantics. IEEE Trans. Knowl Data Eng. 14(5): 1155-1166.

[20] Xiong M, Han S, Lam KY (2005) A deferrable scheduling algorithm for real-time transactions maintaining data fressness. In: IEEE Real-Time Systems Symphosium.

[21] Xiong M, Han S, Chen D (2006) Deferrable scheduling for Temporal Consistancy: Schedulability analysis and Overhead Reduction. In: IEEE International Conference on Embedded and Real-Time Computing Systems and Applications, Aug 2006.

[22] Xiong M, Han S, Lam KY, Chen D (2008) Deferrable scheduling for maintaining Real-Time Data Freshness: algorithm, analysis and results. IEEE Trans Comput 57(7):952-964.

[23] Xiong M, Han S, Chen D, Lam KY, Shan Feng (2009) DESH: overhead reduction algorithms for deferrable. In Real-Time Systems.

Proposed Algorithm of System Log Process for Application Software in Linux

Dhirender Kumar [1], Ajay Kumar [2] and Garima Verma[3]

[1,2]Dehradun Institute of Technology, Dehradun, Uttarakhand, India
E-mail: [1]dhiruphy@gmail.com, [2]kumarajay7th@gmail.com
[3]Noida Institute of Engineering and Technology, Greater Noida
E-mail: garimaverma76@gmail.com

Abstract
Operating system is an interface between the user and hardware; it is resposible for the management and coordination of activities and the sharing of resources of the computer. This relieves application program from having to manag these details and make it easier to write applications . Log files are the files that conatiln messages about the system including the kernel, services and the application running on it. There are different log files for the diffrent information.

This paper represent the proposed algorithm of the system log process which contains the specific contents of the application software running on the linux operating system attached to a server. This log file can be used for many purpose like to find out the number of process to do a particular task, to fid out the intruders on the network etc. This userdefined log file can implement for many purposes depending upon the need of the user and the history of process or system call of the application.

Keywords: operating system, process_id, system calls, system log process, message queue, message buffering.

Introduction

Operting system is responsible for the management and coordination of the activities and sharing of resources of the computer. It act as a host for the computing applications running over the machine. As a host one of the ourposes of an operating system is to handle the details of the operations of the hardware and software. Log files are pressent in the system to find out the history of the application software running on it. There are diffrent log files for the diffrent information. These log files are generally used by the operating system[3].

We can use the log files for the specific purposes if it is made in such a way that it contains the messages according to the need of the user.

The operating system observability require the interpersonal communication with the system.IPC describe the different way of message passing between the different processes running on some operating system. The histories of these message passing is used for many purposes in the form of log files.

An operating system have the two modes user mode and kernel mode. In the user mode we have the application program which use the library function of the the software package and generate the system calls[3].

Kernel of an operating system interact with the machines hardware and the shell interact with the user. The user interact with the systems hardware by using the services of kernel through a set of function called system calls. System call is the interface of the user and hardware. Every process generate the system call to get the resource of hardware needed to execute that process. It contains the information of process_id, user(terminal) and command for which it was generated.[7]

About System log process

Program is the statement of any specific programing language or any executable file residing in a memory disk. A program is read into the memory and executed by the kernel. The executing instance of a program is called the process. A process consist of an executing (running) program, its current value, the state information and the resource used by the operating system to manage the process[3]. Every process running on the client node has a unique identification number called process_id.

Log files are the files that contain the messages about the system including the kernel, services and applicaions running on it. There are the diffrent log files for diffrent information[7].

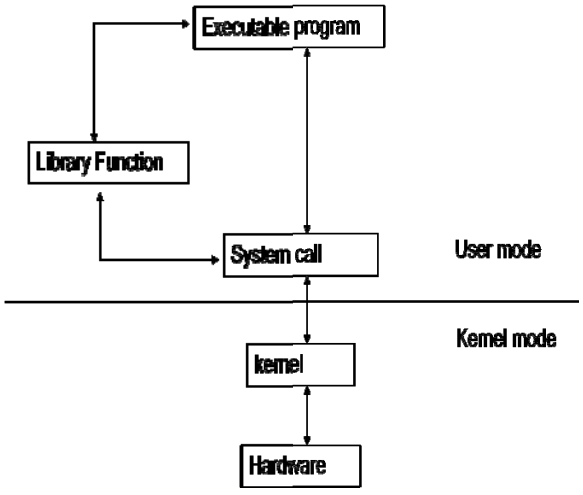


Figure 1. Hardware and software layers of an operating sysem

User defined system log process is the method through which we make the log files on the server for all the processes running on the client nodes. According to this method we have many terminals(called node) which are connected to the server. Every node generate the system calls which will be sent to the server. These system calls will be collected in the message queue through the sendQ(), it is a method through which the system calls(messages) of the diffrent nodes on the network are sending to the message queue[5]. Message queue is initialized through getQ() method. After this the massages are segregated on the basis of their process_id and terminal numbers. Afer this these messages are stored in the local buffers. If the buffer get full then the data of the buffer will be sent to the permanent file called the user defined log file, and the buffer being empty to store the new messages in it[2].

Proposed algorithm for system log process

Message queue operation:

Step 1. Every system calls come to the server in the form of a message.

Step 2. Needed information of these system calls extracted to send through function sendQ() in the message queue.

Step 3. Through function getQ() a message queue is initialized and all the messages are collected in it on the server

Step 4. After this we receive the messages one by one from the message queue through function rcv().

Step 5. Receive the messages from message queue and segregate them on the basis of their process_ID and their terminal numbers.

Process log:

Step 6. On receiving the message from message queue it is stored in its buffer(called log file) if the buffer is not full.

Step 7. If the buffer is full then the content of the buffer is sent to the permanent log files and the buffer being empty to store the other messages in it.

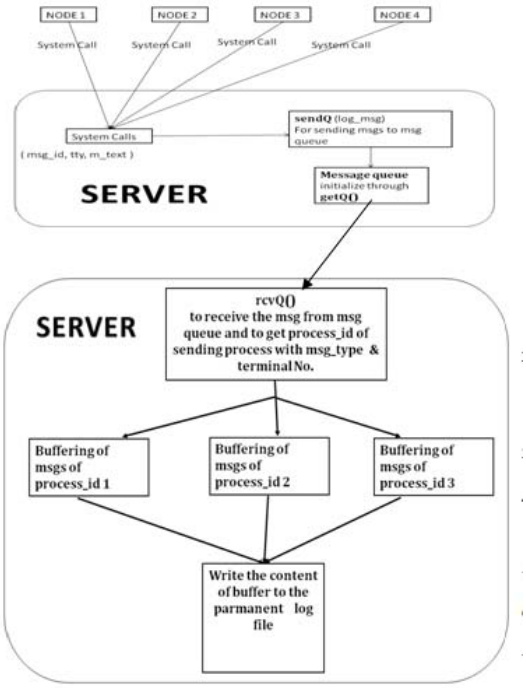


Figure 2. model of system log process

Applications of user defined system log process:

1. Every event of the software process can be stored in the log files and by using the log file analyzer we can analyze the log file to find out any error or any kind of failure of the software to perform any task[1].
2. It can be used for the test result checking in the process of software testing[4].
3. Log files of a particular software can be used by log file analyzer to check the efficiency of software to perform any particular task under the controlled way.
4. The path for any process running on any terminal in the network can be tracked by the log files[6].
5. Log files are useful to trouble shoot any problem of the system[6].

Conclusion

System log process is an approach through which we can make the log record of all the processes running on any particular node in any software to check its efficiency and check its error. System log process is a part of software testing which can be used to test the error and efficiency of software to perform a particular task. By making the use of log file analyzer we can recognize that which process is responsible for the error in the software so that we can remove that easily.

References

[1] A. James H, Z. Yingjun: General Test Result Checking with Log File Analysis, IEEE Trans. On Software Engineering, vol 29, No. 7, pp. 634-648 (2003)

[2] A. Goel, S.C. Gupta, S.K. Wasan: Probe Mechanism for Object-Oriented Software Testing. In Mauro Pezze,

editor, In Proceedings of Fundamental Approaches to Software Engineering (FASE 2003), Lecture Notes in Computer Science, LNCS 2621, pp. 310-324, Warsaw, Springer, Poland, (2003)

[3] A.Silberschatz, P.B. Galvin: Operating System Concept, Fifth Edition,John Wiley & Sons (2000).

[4] J.H. Andrews, "Testing Using Log File Analysis: Tools, Methods and Issues", Proc. Int'l Conf. Automated Software Eng. (ASE '98), pp. 157-166, Oct. 1998.

[5] J.S. Gray, "Interprocess Communications in Linux", Prentice Hall PTR, (2003).

[6] R.J. Moore: A universal dynamic trace for Linux and other operating systems. In Proceedings of FREENIX Track (2001)

[7] W.R. Stevens, "Advanced Programming in the UNIX Environment", Addison -Wesley Longman, Singapore Pte. Ltd., (2001).

Path Tracking Algorithm for a Robot Manipulator

Neha Kapoor[1], Jyoti Ohri[2] and Gopal Krishan[3]

[1&3] Technological Institute of Textile and Sciences, Bhiwani, India
[2]National Institute of Technology, Kurukshetra, India
E-mail: ernehakapoor@rediffmail.com

Abstract
A number of algorithms for mobile robot path have been described in the robotics control literature. This paper presents a comparative experimental study of the classical controllers for the path tracking 1-link robot. Dynamic model of the robot has been taken here and two types of controls i.e. PD and PID have been compared. Experiments have been done by using MATLAB. Gains for both the controllers have been determined by TAE (Trial And Error) method. PID controller's efficiency over PD controller has been proved. Both the controllers are composed of mainly four components: a pre-defined path, a predictive model, an offline controlling algorithm and a feedback tuning model.

Keywords: PD, PID, 1-link robot.

Introduction

A robot is a virtual or mechanical artificial agent. Robots have replaced slaves in the assistance of performing those repetitive and dangerous tasks which humans prefer not to do or unable to do due to size limitations or even those such as in outer space or at the bottom of the sea where humans could not survive the extreme environments. The word robotics, used to describe this field of study, was coined by the science fiction writer Isaac Asimov. Robotics is the engineering science and technology of robots, and their design, manufacture, application, and structural disposition. The control of a robot involves three distinct phases - perception, processing, and action (robotic paradigms). Sensors give information about the environment or the robot itself (e.g. the position of its joints or its end effectors). This information is then processed to calculate the appropriate signals to the actuators (motors) which move the mechanical. For a suitable control strategy the robot model must capture the flexible

Literature Review

The dynamics of a robot arm is highly non-linear. The acceleration, velocity and angle of a single joint affect other joints.

Moreover, other external forces including Coriolis force, centrifugal force, gravity, friction etc. are present, and they can influence the states of a robot. The states of a robot have an effect on themselves a well. In order to control the arms, the torque at each joint must be calculated every moment. But with the increase of degree of freedom, calculating time also increases. Moreover the unknown external forces also can be applied to the robot. Noh and Won [1] used a disturbance observer to find the effect of unknown forces like friction and damping effects. So, by taking the known part of the system as linear and unknown part as nonlinear disturbance, a control technique using Proportional Derivative (PD) controller is developed taking linear part into consideration. A closed chain robot has several advantages over an open chain robot. Guo and Zhang [2] taking the fact that closed chain robot system dynamics due to problem associated with unknown link masses and unknown joint frictions used the adaptive control to make robot to follow the right trajectory with minimum possible errors. A PD plus gravity compensation control realized accurate point to point tracking [3], and a PD computed torque control achieved adequate trajectory tracking [4]. To improve the tracking performance, Lin and Chen [5] developed a control system that considered of a model reference adaptive control (MRAC), a modified switching algorithm, a disturbance compensation loop and several feedback loops to control a closed chain linkage.

Seraji in 1986[6] used a simple scheme to control the dynamics of robot using PD controller. Nirav A. Patel et al [7] used a microcontroller to control and to make a robot walk. They had simulated results of a moving robot using stepper motors and microcontroller and realization of the physical robot is under progress. Dexterous and skilled motions in robot manipulators require reliable and robust joint controllers for achieving accurate joint motion tracking despite uncertainties in the robot dynamics, external disturbances, friction and unknown pay load. Jyoti Ohri et al [8] in 2007 used a decentralized robust H∞ PID controller applied to tracking problem to guarantee arbitrary disturbance attenuation. Morris and Madani [9] used a method of approach to develop a single link model and then to expand this into a two link model, taking proper account of the coupling between the two links. Computed torque and quadratic optimal controllers based on this model have been developed. Pole placement control and adaptive control schemes are designed by Ru Lai and Fujio Ohkawa [10] by discretizing the robot model by using the trapezoidal rule and eliminating the non-linear force terms and external force term from the robot equation.

Each method of control is different in terms of accuracy, scope, time horizon and cost. To facilitate an adequate level of control, the developer has to be responsive to the characteristics of different methods, and determine if a particular method is appropriate for the undertaken situation

before embarking its usage in real application. As a result, the choice of a control method is one of the important factors that will influence the control accuracy. Control techniques can broadly be divided into two categories i.e. intelligent and Non-Intelligent techniques. One of the major limitations of non-intelligent methods is that it requires a very accurate model, without that there are many errors in the system. So, to reduce the modeling errors and to make the controls which can work without models, Artificial Intelligent (AI) techniques came into existence. James M. Adams and Kuldeep S. Rattan [11] used a multistage Fuzzy Logic Controller (FLC) for a 2-link, direct drive robot and results have been compared with the classical PID controller without any change in the plant parameters. Also, in 2003, Partricia Melin and Oscar Castillo [12] used type-2 FLC in spite of type-1 FLC. The main advantage of using AI control technique is due to the greater ability of this theory in modeling uncertainties in the control of non-linear plants.

Robot Dynamics

There are several methods to develop a trajectory that a robot has to follow. Robot has to pass from the initial point and then has to follow the path as close as possible. There are many control algorithms are present to control the robot and to make it move on the desired trajectory. A set of pre-decided points are taken to develop a path, robot now start from the initial point of path and then move. Feedback from the robot's position is taken at every point and is given to the controller to decide the input torque to the robot. This type of control is called as the position control in the robot dynamics. Here, for these experimental results the path taken is given in equation (1) and its pictorial view is given in Fig.1.

$$q_d = 30 + 6t^2 - 1.2t^3 \quad (1)$$

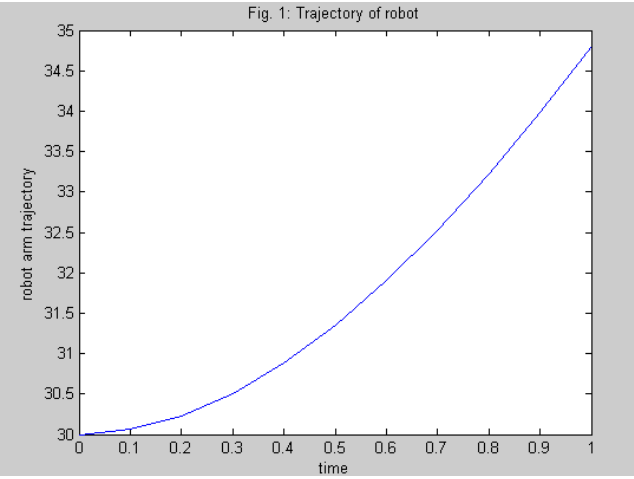


Fig. 1: Trajectory of robot.

It is always necessary to analyze the dynamic characteristics of a robot in order to control it accurately and to evaluate its performance. A robot is most often an open loop link mechanism, which may not be a good structure from the view point of dynamics. This structure, however allow us to derive a set of simple, easily understandable equations of motion. Also, recently, as the need for more rapid and accurate operation of manipulators has increased, the need for real time computation of the dynamics equations has been felt more strongly [13]. 1-link robot taken in this research has a model equation as given in equation (2).

$$M(q)\ddot{q} + V\dot{q} + G = \tau \quad (2)$$

Where M (q) = (10+6sinq) is a inertia matrix;

V= (6sinq) is a factor representing coriolis and centrifugal forces;

G= (2cosq) represents the gravitational forces.

Two Control Approaches

For a suitable control strategy the robot model must capture the flexible dynamics. In this section, two approaches for control of a 1-link robot, whose dynamics are mentioned has been given investigated.

PD controller: The majority of existing industrial manipulators are controlled using proportional derivative (PD) controllers [14]. PD control is a conventional feedback control approach which has been extensively used. This type of controller is very simple and easy to design, which still makes this controller to be used in almost all the industrial robots. General equation for PD controller is given in equation (3).

$$\tau = K_p e(t) + K_d \dot{e}(t) \quad (3)$$

In most current robotic applications, PD controllers are functional and sufficient due to the high reduction ratio of the transmissions used.

Here, PD feedback controller is used to take the feedback from the output of the robot dynamics. Error is being calculated by comparing the feedback quantities with the desired one. These errors are further used to set the values of the PD controller. Values of the controller constants i.e. K_d and K_p are being decided by TAE (Trial And Error) method. Error in the path covered by robot and the actual path is shown in Fig. 2. Average error coming out is 1.2801.

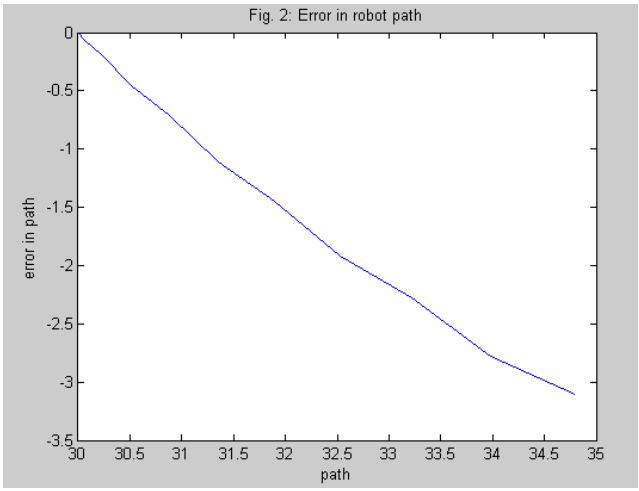


Fig. 2: Error in robot path with PD controller.

PID controller: PID stands for Proportional, Integral and Derivative. Controllers are designed to eliminate the need for continuous operator attention. A proportional–integral–derivative controller (PID controller) is a generic control loop feedback mechanism (controller) widely used in industrial control systems – a PID is the most commonly used feedback controller. A PID controller calculates an "error" value as the difference between a measured process variable and a desired set point. The controller attempts to minimize the error by adjusting the process control inputs. General equation for PD controller is given in equation (4).

$$\tau = K_p e(t) + K_d \dot{e}(t) + K_i \int e(t) \_\_\_\_\_ \quad (4)$$

In most current robotic applications, PID controllers are functional and sufficient due to the high reduction ratio of the transmissions used.

Here, in this block diagram it has been shown that PID feedback controller is using the feedback from the output of the robot dynamics. Error is being calculated by comparing the feedback quantities with the desired one. These errors are further used to set the values of the PID controller. Values of the controller constants i.e. K_d , K_p and K_i are being decided by TAE (Trial And Error) method. Error in the path covered by robot and the actual path is shown in Fig. 3. Average error coming out is 0.6867.

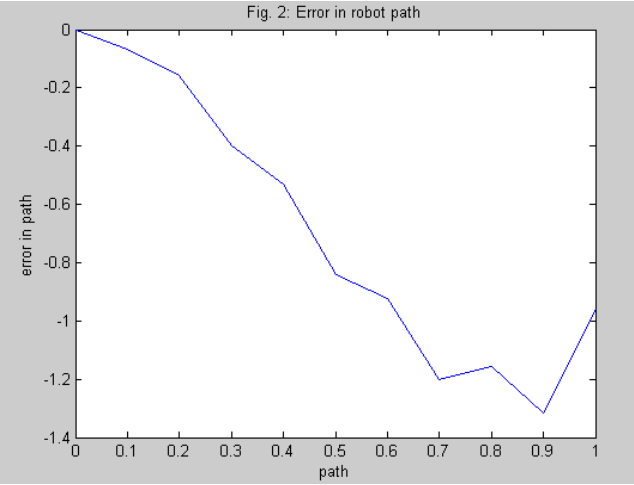


Fig. 3: Error in robot path with PD controller.

From the above two models it can be seen that the average error of the PID model is lesser than the PD controller. Also the comparison graph for both the controllers is shown in Fig.4.

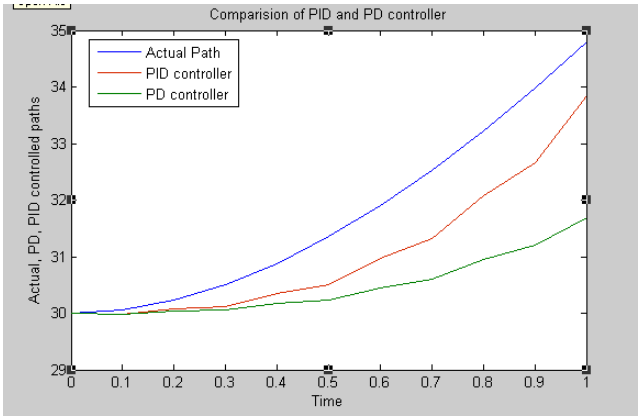


Fig. 4. Comparison of PD and PID controller.

Conclusion

In this research, robot model with dynamic model has been taken into consideration and it has to move on a specific path, which is made by step by step movement of the robot. An open loop control system always gives bad results as compared to closed loop control system. A linear feedback control system consisting of PD and PID controllers are seen in this paper. From Table 1, it can be seen that the average error obtained in a PID controller is lesser than the PD controller, as the integral part is added to the PD controller. Hence, it can be concluded from the above experimental set up that PID controller is a better controller than PD controller. We can also see from the graph that the path covered by PID controller is closer to the actual path than the path covered by the PD controller.

Table1. Average error of PD and PID controller.

Type of Controller	% error
PD	1.2801
PID	0.6867

References

[1] Noh I. and Who S., Control of Two-link Robot attached to a Mass-Spring using Disturbance. Proc. IEEE Int. Conf. on Control, Automation and Systems. 2007 pp. 590-594.

[2] Guo L.S. and Zhang Q., Adaptive Trajectory Control of A Two DOF Closed-Chain Robot. Proc. Of Americian Control Conf., June 2001. pp. 658-663.

[3] Ghorbel, F. and R. Gunawardana, A validation study of PD control of a closed-chain mechanical system. Proc. of the 36th Conference on Decision & Control, San Diego, California, USA. 1997. pp. 1998-2004.

[4] Guo, L.S., Y.F. Li and W.J. Zhang, Trajectory control of two DOF closed-chain mechanical systems, DETC2000/MECH-14158, Proceedings of DETC'OO ASME 2000 Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Baltimore, Maryland. 2000

[5] Lin, M-C and J-S Chen, Experiments toward MRAC design for linkage system, Mechatronics, 6(8).1996, pp. 933-953.
[6] Seraji H., Linear Multivariable Control of Robot Manipulators, IEEE Trans.. 1986, pp. 565-571.
[7] Patel A.N., Pradhan S.N. and Shah K.D., Two Legged Robot Design, Simulation and Realization, Proc. Of 4th Int. Conf. on Autonomous Robots and Agents, Feb. 2009, pp. 426-429.
[8] Ohri J., Dewan L. and Soni M.K., Tracking Control of Robots Using Decentralized Robust Pid Control For Friction and Uncertainty Compensation, Proc. of the World Congress on Engineering and Computer Science, Oct. 2007.
[9] Morris A.S. and Madani A., Computed Torque vs Quadratic Optimal Control for a Two-Flexible-Link Robot, Proc. IEEE Int. Conf. on Control. 1996, pp. 335-340.
[10] Lai R. and Ohkawa F., A Simple Discrete-Time Robot Model and Its Control Applications, pp. 1448-1453.
[11] Adams J.M. and Rattan K.S., Intelligent Control of a Direct-Drive Robot using Multi-Stage Fuzzy Logic, IEEE, 2001, pp. 543-546.
[12] Melin P. and Castillo O., Intelligent Control of Non-Linear Plants using Type-2 Fuzzy Logic and Neural Networks, IEEE, 2003, pp. 1558-1562.

Determination of Spring Constant of Surface Functionalized Micro-machined Micro-Cantilever

A.S. Kurhekar

Department of Electrical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai, India
E-mail: askurhekar@gmail.com

Abstract
One of the bio-sensing mechanisms is mechanical. Than measuring shift in resonance frequency, we adopt to measure the change in spring constant due to adsorption, as one of the fundamental sensing mechanism. This study explain determination of spring constant of a surface functionalized micro machined micro cantilever, which resonates in a trapezoidal cavity-on Silicon <100> wafer, with the resonating frequency of 7000 cycles per second. This thin-flimsy-oxide micro-cantilever has a typical shape, and the tip of the micro-cantilever is dip-coated with chemically and biologically active material. The change in mass, due to adsorption, is detected by measuring the change in spring constant. The Force-Distance spectroscopy is used to detect the change in spring constant. The experimental results, show that the mechanical sensing scheme used by us, permit this surface functionalized micro machined micro cantilever to be used as a molecular mass sensor.

Keywords: Micromachining, Micro-cantilever, Silicon<100>, Bio-Sensor, F-d Spectroscopy.

Introduction

Designing Bio-MEMS/NEMS has always remained a critical task for the simple reason that the structures, on which the biologically or chemically active materials are dip-coated, smeared, spray-painted, and electro-deposited or sintered, should have bio-compatible surfaces. Normal choice of bio-compatible surfaces boils down to Silicon, Silicon Dioxide or Silicon Nitride. We have selected silicon Dioxide as a bio-compatible surface. This bio-compatible surface, upon surface treatments, becomes functionalized for linking to chemically or biologically active species, to use them as either sensors or detectors. This piece of work explains the design and fabrication of MEMS structure [1] [2] [3] – a simple micro cantilever, which is used for designing Bio-MEMS applications.

One of the bio-sensing mechanisms is mechanical. Than measuring shift in resonance frequency, we adopt to measure the change in spring constant due to adsorption, as one of the fundamental sensing mechanism. This study entails determination of spring constant of a surface functionalized micro machined micro cantilever, which resonates in a trapezoidal cavity-on Silicon <100> wafer [4][5][6], with the resonating frequency of 7000 cycles per second. This thin-flimsy-oxide micro-cantilever has a typical shape, and the tip of the micro-cantilever is dip-coated with chemically and biologically active material.

The change in mass, due to adsorption, is detected by measuring the change in spring constant. The Force-Distance spectroscopy is used to detect the change in spring constant.

Micro-fabrication And Surface Functionalization

Micro-fabrication

Silicon<1 0 0 > n-type wafer is cleaned using Piranha cleaning procedure to remove contaminations. The Piranha cleaned wafer was then taken to a pyro-furnace for 700 nano-meters thermally grown oxide deposition, at $1100\,^{0}C$, after calculating deposition parameters. The computed time was 1.10.11 Hrs using Modified Deal-Grove Time Calculations. The observed wafer deposition-colour was bluish-green. The wafer was then annealed for 1.5 Hrs to remove the stresses. After the wafer was drawn out of the oxidation furnace, the oxide thickness was measured to be 698 nanometres using Ellipsometer. The wafer was then given a Dehydration-bake at $120\,^{0}C$ for 45 minutes on the heating iron. The MICROPOSIT® S-1813 positive photo-resist was dispensed at the centre of the wafer, using a dropper, on the wafer-on-the-chuck-of-spinner and spined at 500 rpm for 30 seconds and 3000 rpm for 2 minutes. After Positive Photo-Resist dispensing and spinning, the wafer was then PRE-BAKED at $70\,^{0}C$ for 20 minutes and then taken for Photolithography using Karl-Suss® contact aligner for mask image transfer. The patterned wafer was then developed using a developer and observed under the microscope. The developed wafer was then micro-machined using 3:4 TMAH+WATER at 80 degrees centigrade. Wafer was intermittently observed under the microscope to ensure reliable machining. This micro-machined wafer was the tilted at $45\,^{0}$ and Iso Propyl Alcohol was dispensed on the wafer slowly to ensure release of the micro cantilevers. The chrome-gold sputtered layer improves adhesion of gold to the Silicon dioxide surface. Chrome-Gold was deposited using RF Magnetron sputtering. Substrate heating, while chrome-gold sputtering was not considered to be a necessary step. We have sputtered Chrome for 1 minutes and Gold for 3 minutes.

Surface Functionalization

The sputtered chrome-gold layer has affinity with thio-phenol molecules.[7] Considering this fact, we have dip-coated the piezo-resistive micro cantilevers with 1 micromole thio-

phenol in ethanol solution for 3 Hrs. and then, rinsed with ethanol, for 2 minutes. The surface becomes functionalized for mass-sensing.

Reference Spring Cantilever Calibration With Atomic Force Microscopy

The spring constant of the unknown spring is calibrated by pressing it against a very stiff surface and then against a reference spring of known and lesser compliance.[8] The spring constant of the cantilever under test is then computed using the relation(1),

$$k_{unknown} = k_{std} \cdot [(InvOLS_{std} / InvOLS_{unknown}) - 1] \quad (1)$$

In the above equation, $InvOLS_{unknown}$ is the inverse Optical lever Sensitivity (nm/Volt) for the cantilever under test measured on a very stiff surface and $InvOLS_{std}$ is the same quantity measured on a compliant surface with spring constant k_{std}. To determine the spring constant of the sample cantilever, the spring constant of the AFM cantilever must be known along-with the slope of the force curve. The slope of the force curve gives us the deflection sensitivity in nm/Volt. To compute the spring constant of the micro cantilever fabricated by us, we have used the relation (2),

$$k_{sample} = k_{AFM} [(Deflection\ Sensitivit\ yofHard\ Re\ gion / Deflection\ Sensitivit\ yofSoft\ Re\ gion) - 1] \quad (2)$$

Substituting the measured values of deflection sensitivities and known value of K_{AFM}, we have computed,

$$k_{sample} = 0.58 \cdot [(46.8 / 36.75) - 1] = 0.1622\ Newton / meter$$

Computation of Spring Constant of Thio-Phenol treated Micro-Cantilever

With the above explained method, we have computed the spring constant of a surface functionalized micro cantilever, that reveal the change in the value of resonant frequency, in-turn, change in value of K,

For micro − cantilever1

$$k_{sample} = 0.58 \cdot [(46.8 / 46) - 1] = 0.0100\ Newtons / meter$$

For micro − cantilever2

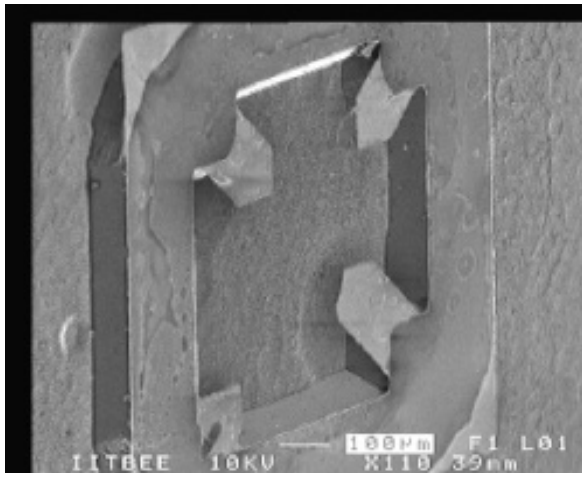
$$k_{sample} = 0.58 \cdot [(46.8 / 32) - 1] = 0.2682\ Newtons / meter$$

For micro − cantilever3

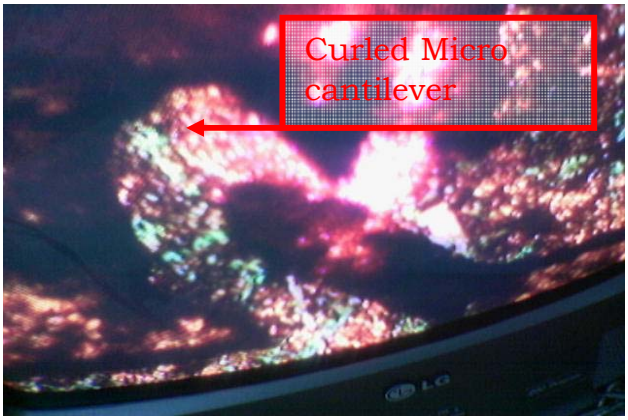
$$k_{sample} = 0.58 \cdot [(46.8 / 41.5) - 1] = 0.0740\ Newtons / meter$$

Results

Figure 1 depicts Scanning Electron Microscopy Micrograph and Atomic Force Microscopy Micrograph of the thio-phenol treated-gold functionalized surface of micro cantilever. Figure 2 depicts Atomic Force Microscopy Micrograph of thio-phenol treated gold surface scan showing clusters of thio-phenol molecule and Figure 3 indicates Surface topography of thio-phenol treated-gold surface. F-d spectroscopy of Micro cantilever before and after surface functionalization is depicted in Figure 4 and Figure 5. The investigation of F-d spectroscopy results reveals that, there is change in slope of the force-distance curve. Table 1 is a measure of change in slope is due to the change in spring constant, due to the adsorption of the thio-phenol molecule to the gold surface.



(a) Scanning Electron Microscopy Micrograph of Micro cantilever on silicon<1 0 0 > surface.



(b) Atomic Force Microscopy Micrograph of Surface Functionalized Micro-cantilever on silicon<1 0 0 > surface

Figure 1. Surface Functionalized Micro cantilever.

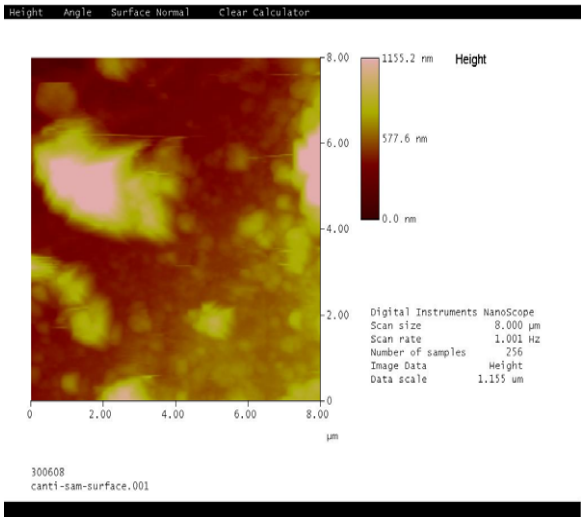


Figure 2: Atomic force microscopy of thio-phenol treated gold surface on silicon <1 0 0 >. Scan showing clusters of adsorbed thio-phenol molecule

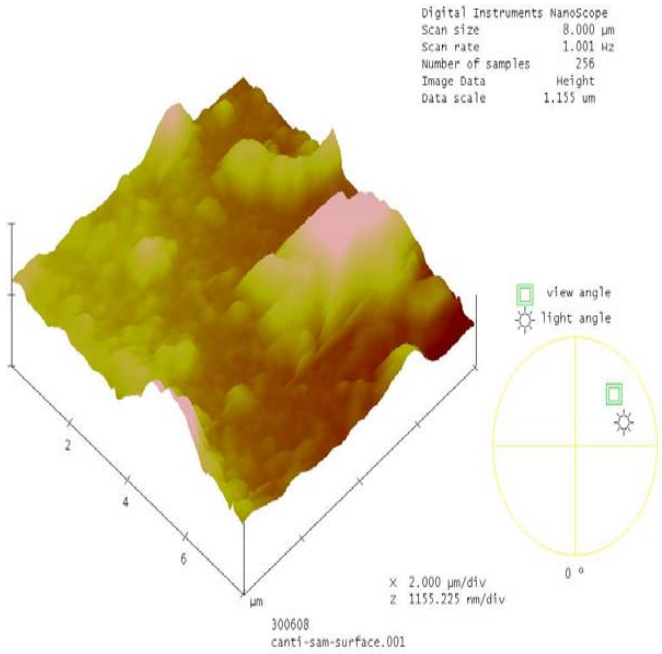


Figure 3: Surface topography of thiophenol treated-gold surface

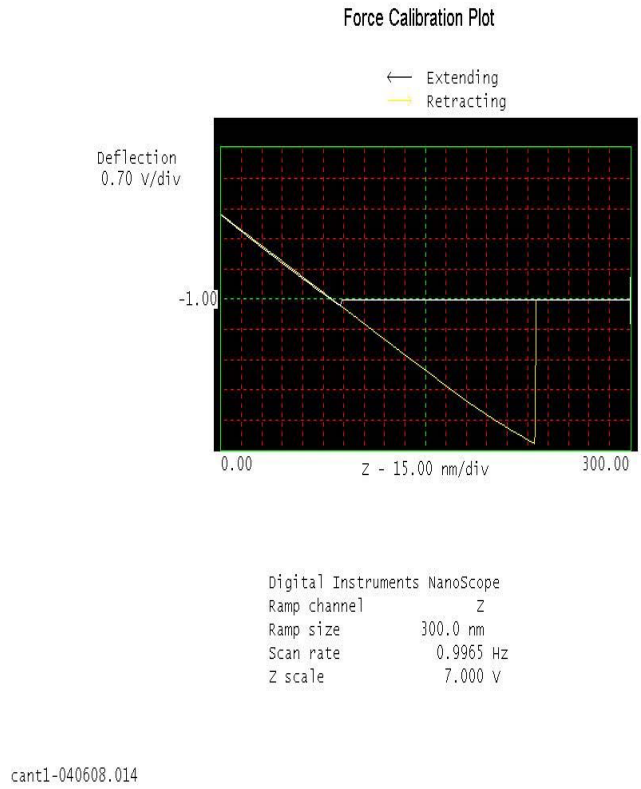


Figure 4. F-d spectroscopy of micro cantilever before the surface functionalization.

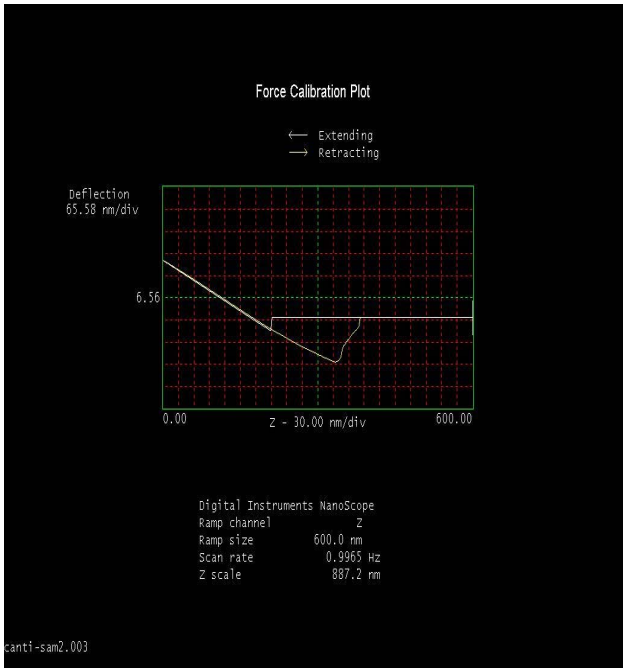


Fig. 5. F-d spectroscopy of micro cantilever after the surface functionalization.

TABLE I Spring Constant: Before and After Surface Functionalization

Sr.No.	Spring Constant of Micro machined Micro cantilever			
	Before Surface Function-alization	After Surface Function-alization	After Surface Function-alization	After Surface Functionalization
1.	*0.16224 Newton/Meter*	*0.0100 Newton/Meter*	*0.26825 Newton/Meter*	*0.0740 Newton/Meter*

Conclusion

The change in mass, due to adsorption, is detected by measuring the change in spring constant. The Force-Distance spectroscopy is used to detect the change in spring constant. The study of F-d spectroscopy curves reveals that the surface functionalized micro cantilever behaves as a molecular mass sensor. The proposed device shall be used as micro-molecular mass sensor.

Acknowledgment

The author acknowledges the Microelectronics Group, Centre of Excellence for Nanoelectronics, Department of Electrical Engineering, Department of Chemistry, Department of Physics, Suman Mashruwala Micro machining Laboratory – Department of Mechanical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai, INDIA

References

[1] Marc J. Madou, Fundamentals of Microfabrication: The Science of Miniaturization, 2nd ed., Boca Raton, Florida: CRC Press, 2002.

[2] *http://www.amazon.com/Fundamentals-Microfabrication-Science-Miniaturization-Second/dp/0849308267*

[3] P. R. Apte, U. D. Vaishnav, S. G. Lokhare, V. R. Palkar, and S. M. Pattalwar, "Micromechanical Components With Novel Properties," Proc. Of S.P.I.E., vol. 3321, pp. 287–297, June. 1996.

[4] *http://bookwebpro.kinokuniya.co.jp/booksea.cgi?ISBN=0819427624*

[5] A. S. Kurhekar, "Out-of-the-plane MEMS Device", Proc. of ICANN 2010, 2010, pp. 199–203.

[6] A. S. Kurhekar, "Etch-Depth Measurement of Anisotropically Etched Trapezoidal Micro-Cavity", Proc. of First IFIP Bioinfo-2010, 2010, pp. 106–108.

[7] http://pdfcast.org/pdf/etch-depth-measurement

[8] A. S. Kurhekar, "In-situ Non-Contact and contact precise measurement of significant parameters of micro-machined micro-cantilever," Proc. Of ICONSAT 2010, vol. 3321 pp. 460-462, January. 2010.

[9] *http://www.iconsat2010.in/news_finallistofacceptedabstracts.php*

[10] A. S. Kurhekar, "Optical Multi-sensing in Micro-machined Micro-cantilever", Proceedings of International conference on Contemporary Trends in Optics and Optoelectronics, 2011, pp. 254–257.

[11] *http://www.iist.ac.in/iist-news/international-conference-on-contemporary-trends-in-optics-and-optoelectronics.html*

[12]

[13] John A. Seelenbinder, Chris W. Brown and Daniel W. Urish," Self-Assembled Monolayers of Thiophenol on Gold as a Novel Substrate for Surface-Enhanced Infrared Absorption", Applied Spectroscopy, Vol. 54, Issue 3, pp. 366-370 (2000)

[14] http://www.opticsinfobase.org/abstract.cfm?uri=as-54-3-366

[15] A. Torii, M. Sasaki, K. Hane, S. Okuma, "A Method for determining the spring constant of cantilevers for Atomic Force Microscopy," J. Meas. Sci., Vol.7,Number 2, pp. 179–184, June. 1996.

[16] *http://iopscience.iop.org/09570233/7/2/010;jsessionid=244391E3F67F22C9E40223000F73DA7B.c2*

A Novel Approach of Combining FFT with Ancient Indian Vedic Mathematics

Nidhi Mittal[1] and Abhijeet Kumar[2]

[1]Electronics & Communication Department, M M University, India
E-mail: mail2bansal@gmail.com
[2]Electronics& Communication Department, M M University, India
E-mail: abhijeet.kumar@mmumullana.org

Abstract
In present scenario every process should be rapid, efficient and simple. Fast Fourier transform (FFT) is an efficient algorithm to compute the N point DFT. It has great applications in communication, signal and image processing and instrumentation. But the Implementation of FFT requires large number of complex multiplications, so to make this process rapid and simple it's necessary for a multiplier to be fast and power efficient. To tackle this problem Urthva Tirvagbhyam in Vedic mathematics is an efficient method of multiplication [4]. Vedic Mathematics is the ancient system of mathematics which has a unique technique of calculations based on 16 Sutras. Employing these techniques in the computation algorithms of the coprocessor will reduce the complexity, execution time, area, power etc. Urdhva Tiryakbhyam one of the sutra of Vedic Mathematics, being a general multiplication formula, is equally applicable to all cases of multiplication. The conventional multiplication method requires more time & area on silicon than Vedic algorithms [8]. More importantly processing speed increases with the bit length. This will help ultimately to speed up the signal processing task. The novelty in this paper is Fast Fourier Transform (FFT) design methodology using Vedic mathematics algorithm. By combining these two approaches proposed design methodology is time-area-power efficient.

***Keywords*:** FFT, Urthva Tirvagbhyam, Vedic Mathematics etc.

Introduction

Direct computation of Discrete Fourier Transform (DFT) requires of the order of N^2 complex multiplication operations where N is the transform size. The FFT algorithm, started a new era in digital signal processing by reducing the order of complexity of DFT from N^2 to $N\log_2 N$, reduces the number of required complex multiplications compared to a normal DFT. Since multipliers are very power hungry elements in VLSI designs they result in significant power consumption [7]. So, the complex multiplication operations are realized using Urthva Tirvagbhyam in Ancient Indian Vedic mathematics is an efficient method of multiplication. It literally means "Vertically and crosswise". This Sutra shows how to handle multiplication of a larger number (N x N, of N bits each) by breaking it into smaller numbers of size (N/2 = n, say) and these smaller numbers can again be broken into smaller. numbers (n/2 each) till we reach multiplicand size of (2 x 2) Thus, simplifying the whole multiplication process [11]. The processing power of this multiplier can easily be increased by increasing the input and output data bus widths since it has a quite regular structure. Due to its regular structure, it can be easily layout in a silicon chip. The Multiplier has the advantage that as the number of bits increases, gate delay and area increases very slowly as compared to other multipliers [2]. In the present scenario high speed digital telecommunication systems such as OFDM and DSL need real-time high-speed computation of the Fast Fourier Transform. Pipeline architecture based on the constant geometry of N point radix-2 FFT algorithm, which uses $N/2\log_2 N$ complex number multipliers and is capable of computing a full N-point FFT in N/2 clock cycles[3], has been proposed. Thus there is a need of innovative algorithms to improve the speed .In this paper, we propose Vedic algorithm for the implementation of multipliers to be used in the FFT. Fast Fourier Transform (FFT) design methodology using Vedic mathematics algorithm provides a fast and a reliable approach to compute the N point DFT**.**

Fast Fourier Transform (FFT)

The computation of the N point DFT by the divide-and conquer approach provides a computationally efficient algorithm. We split the N-point data sequence into two N/2-point data sequences $f1(n)$ and $f2(n)$, corresponding to the even numbered and odd-numbered samples of $x(n)$, respectively, that is,

$$f1(n) = x(2n), \quad (1)$$

$$f2(n) = x(2n+1)\ n = 0,1,\ldots..N/2\text{-}1 \quad (2)$$

Thus $f1(n)$ and $f2(n)$ are obtained by decimating $x(n)$ by a factor of 2, and hence the resulting FFT algorithm is called a decimation-in-time algorithm.

FFT of a sequence x(n) of length N is given by X(K)

$$X(K) = \sum_{n=} x(n) W_N^{nk}, \quad 0 \le K \le N\text{-}1 \quad (3)$$

Where $W_N = e^{-j2\pi/N}$, is a complex valued phase factor.

FFT take the advantage of the periodicity and symmetry of the complex number W_N. Thus reducing the complex number multiplications and additions from N^2 to $N/2\log_2 N$

and $N\log_2 N$ respectively. In FFT, where N is an integer power of 2, i.e. $N=2^L$, the no of stages of computation is L ($=\log_2 N$), then this algorithm is known as radix-2 FFT algorithm. For N=16, which consist of $L = \log_2 16 = 4$, four stages, the first stage computes the eight 2-point DFTs, the second stage computes the four 4-point DFTs, the third stage computes the two 8-point DFTs and finally the fourth stage computes the desired 16 point DFT. The number of complex multiplications are $N/2\log_2 N = 8\log_2 16 = 32$ and the number of complex additions are $N\log_2 N = 16\log_2 16 = 64$. The basic operation of DIT algorithm is the butterfly in which two inputs f(0) and f(1) are combined to give the outputs F(0) and F(1).

$$F(0) = f(0) + f(1)W_{16}^{0} \quad (4)$$

$$F(1) = f(0) + f(1)W_{16}^{8} \quad (5)$$

The corresponding flow graph of a 2-point DFT is shown in fig.1.

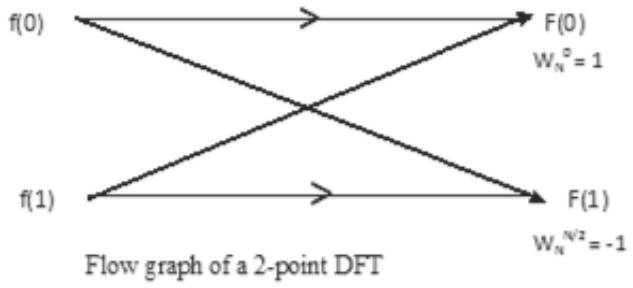


Fig. 1 Flow graph of a 2-point DFT.

Vedic Mathematics

Vedic mathematics is an ancient mathematics concept that provides a fast and a reliable approach to perform arithmetic operation using sixteen sutras [5] which was rediscovered from the Vedas between 1911 and 1918 by Sri Bharati Krishna Tirthaji comprised all this work together and gave its mathematical explanation while discussing it for various applications. Swamiji constructed 16 sutras (formulae) and 16 Upa sutras (sub formulae) after extensive research in Atharva Veda. Vedic mathematics is not only a mathematical wonder but also it is logical. That's why it has such a degree of eminence which cannot be disapproved. Due to these phenomenal characteristics, Vedic maths has already crossed the boundaries of India and has become an interesting topic of research abroad. Vedic math's deals with several basic as well as complex mathematical operations. Especially, methods of basic arithmetic are extremely simple and powerful. The word "Vedic" is derived from the word **"Veda"** which means the store-house of all knowledge. The Vedic mathematics approach is totally different and considered very close to the way a human mind works. A large amount of work has been done in understanding various methodologies. The Sutras apply to cover each and every part of mathematics (including arithmetic, algebra, geometry, trigonometry, astronomy, calculus etc. The beauty of Vedic mathematics lies in the fact that it reduces the otherwise cumbersome-looking calculations in conventional mathematics to a very simple one. This is so because the Vedic formulae are claimed to be based on the natural principles on which the human mind works. This is a very interesting field and presents some effective algorithms which can be applied to various branches of engineering such as computing and digital signal processing.

Description of Urdhva Tiryakbhyam

The multiplier is based on an algorithm Urdhva Tiryakbhyam of ancient Indian Vedic Mathematics. Urdhva Tiryakbhyam Sutra is a general multiplication formula applicable to all cases of multiplication [3]. It literally means "Vertically and crosswise". It is based on a novel concept through which the generation of all partial products can be done and then, concurrent addition of these partial products can be done. Thus parallelism in generation of partial products and their summation is obtained using Urdhava Tiryakbhyam. The Multiplier has the advantage that as the number of bits increases, gate delay and area increases very slowly as compared to other multipliers[9]. Therefore it is time, space and power efficient. This Sutra shows how to handle multiplication of a larger number (N x N, of N bits each) by breaking it into smaller numbers of size (N/2 = n, say) and these smaller numbers can again be broken into smaller numbers (n/2 each) till we reach multiplicand size of (2 x 2).Thus, simplifying the whole multiplication process. The multiplication algorithm is then illustrated to show its computational efficiency by taking an example of reducing a 4x4 bit multiplication to a 2x2-bit multiplication operation [12]. To analyse 4x4 multiplications, say X3X2X1X0 and Y3Y2Y1Y0. Following are the output line for the multiplication result, S7S6S5S4S3S2S1S0. Divide X and Y into two parts, say X3X2 & X1X0 for X and Y3Y2 & Y1Y0 for Y. Using the fundamental of Vedic multiplication, taking two bit at a time and using 2 bit multiplier block, we can have the following structure for multiplication.

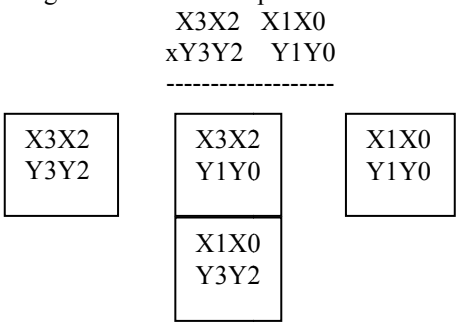


Fig. 2. Block diagram presentation for 4x4 multiplications

Each block as shown above is 2x2 multiplier. First 2x2 multiplier inputs are X1 X0 and Y1 Y0.The last block is 2x2 multiplier with inputs X3 X2 and Y3 Y2. The middle one shows two, 2x2 multiplier with inputs X3X2 & Y1Y0 and X1X0 & Y3Y2. So the final result of multiplication, which is of 8 bit, S7S6S5S4S3S2S1S0, X3X2 and Y3Y2 give multiplication result S33S32S31S30, X3X2 and Y1Y0 give multiplication result S23S22S21S20, X1X0 and Y3Y2 give multiplication result S13S12S11S10, X1X0 and Y1Y0 give multiplication result S03S02S01S00.

For the final result, add the middle product term along with the term shown below.

S33 S32 S31 S30 0 0 S01 S00
S23 S22 S21 S20
S13 S12 S11 S10
0 0 S03 S02

The first two outputs S0 and S1 are same as that of S00 and S01. Result of addition of the middle terms by using two, 4 bit full adders will forms output line from S5S4S3S2. One of the full adder will be used to add (S23 S22 S21 S20) and (S13 S12 S11 S10) and then the second full adder is required to add the result of 1st full adder with (S31 S30 S03 S02). The respective sum bit of the 2nd full adder will be S5S4S3S2. Now the carry generated during 1st full adder operation and that during 2nd full adder operation should be added using half adder so that the final carry and sum to be added with next stage i.e. with S33 S32 to get S7 S6.

Combine Approach of FFT with Vedic Mathematics

As FFT reducing the complex number multiplications and additions from N^2 to $N/2\log_2 N$ and $N\log_2 N$ respectively. For N=16, which consist of four stages, the first stage computes the eight 2-point DFTs, the second stage computes the four 4-point DFTs, the third stage computes the two 8-point DFTs and finally the fourth stage computes the desired 16 point DFT. The number of complex multiplications are $N/2\log_2 N = 8\log_2 16 = 32$ and the number of complex additions are $N\log_2 N = 16\log_2 16 = 64$. The basic operation of DIT algorithm is the butterfly in which two inputs are combined to give the outputs . When the word length to be 16 bits. The single simple multiplier implementation needs 16 rows of partial product generation and each row containing 16 partial product bits. To accumulate these 16 partial product rows large hardware will be needed to get the result in sum and carry form. As implementation of 16 pt radix-2 FFT requires large no of multiplication and these all multiplications are done using Vedic mathematics reduces the time, area and power.

e.g. (a+ib) (c+id) = (ac-bd) +i (ad+bc)

Where $i^2 = -1$ and a, b, c, and d are 4 bit numbers.

a and c are real while b and d are imaginary.

Now results of multiplication of ac, bd, ab and bc are obtained using Vedic algorithm. . In the proposed architecture, the 4x4 bit multiplication operation is fragmented reconfigurable FFT modules. The 4x4 multiplication modules are implemented using small 2x2 bit multipliers. The structure of FFT will be designed, optimized and implemented on SPARTAN-3E FPGA (Field Programmable Gate Array). This FFT have the high speed and small area as compared to the conventional FFT. This particular FFT is to be designed by using Vedic adder, Vedic subtractor, and Vedic multiplier. The delay produced by the Vedic FFT is smaller than the delay produced by the conventional FFT.The application of FFT algorithm include Linear filtering, Correlation, Spectrum Analysis which will further add the field of Communication, signal & image processing and instrumentation. Combine approach of FFT with Vedic Mathematics create the new advancement in various fields of engineering.

Conclusion

In this paper a novel technique of Fast Fourier Transform (FFT) design methodology using Vedic mathematics algorithm is presented. The design is based on Vedic method of multiplication that is quite different from the conventional method of multiplication like add and shift. This also gives chances for modular design where smaller block can be used to design the bigger one. This gives method for hierarchical multiplier design. So the design complexity gets reduced for inputs of large no of bits and modularity gets increased. This will help in designing FFT structure, as its give effective utilization of structural method of modelling. An FFT circuit has been described that provides the high performance with Small area which has great applications in communication, signal and image processing and instrumentation that can also benefit future needs of wireless communications systems.

References

[1] Ashish Raman, Anvesh Kumar, R.K.Sarin, ‘‘High Speed Reconfigurable FFT Design by Vedic Mathematics’’, journal of Computer Science and Engineering, vol.1, pp 59-63 May 2010.

[2] Anvesh Kumar, Ashish Raman, ‘‘Small Area Reconfigurable FFT Design by Vedic Mathematics”, vol 5, IEEE pp 836-838, 2010.

[3] Laxman P. Thakre, Suresh Balpande,Umaeh Akare, Sudhair Lande, ‘‘Performance evaluation and Synthesis of Multiplier Used in FFT Operation Using conventional and Vedic Algorithm”, International Conference on emerging trends in Engineering and Technology, pp 614-619, 2010.

[4] M.E.Paramasivam, Dr.R.S.Sabeenian, ‘‘An Efficient Bit Reduction Binary Multiplication Algorithm using Vedic Methods”, IEEE pp 25-28, 2010.

[5] Anvesh Kumar, Ashish Raman, ‘‘Low Power ALU Design by Ancient Mathematics”, vol 5, IEEE pp 862-865, 2010.

[6] Sumit Vaidya, Deepak Dandekar, ‘‘Delay-Power Performance Comparison of Multipliers in VLSI Circuit Design”, International Journal of Computer Networks & Communications (IJCNC), Vol.2, No.4, pp 47-56 July 2010.

[7] Leonard Gibson Moses S, Thilagar M, ‘‘VLSI Implementation of High Speed DSP algorithms using Vedic Mathematics’’, International Journal of Computer Communication and Information System, Vol.2. pp 119-122 Jul –Dec 2010.

[8] Parth Mehta, Dhanashri Gawelli, “Conventional Versus Vedic Mathematical method for Hardware Implementation of a multiplier”, International Conference on emerging trends in Engineering and Technology, pp 640-642, 2009.

[9] M.Ramalatha, K.Deena Dayalan, S. Deborah Priya, P.Dharani, ‘‘High Speed Energy Efficient ALU Design using Vedic Multiplication Techniques”, ACTEA IEEE Zouk Mosbeh, Lebanon, pp 600-603 July 15-17, 2009.

[10] Ramalatha M, Deena Dayalan, Thanushkodi K, Dharani P,‘‘A Novel Time and Energy Efficient Cubing Circuit using Vedic Mathematics for Finite Field Arithmetic”, International Conference on Advances in Recent Technologies in Communication and Computing , pp 873-875, 2009.

[11] Harpreet Singh Dhillon and Abhijit Mitra, ‘‘A Reduced Bit Multiplication Algorithm for Digital Arithmetic’s”, International Journal of Computational and Mathematical Sciences Spring, 2008.

[12] Anthonyo Brien, Richard Conway, ‘‘Lifting Scheme Discrete Wavelet Transform using Vertical and Crosswise Multipliers”, ISSC, Galway June 18-19 pp 331-336, 2008.

[13] Honey Durga Tiwari, Ganzorig Gankhuyag, Chan Mo Kim, Yong Beom Cho, ‘‘Multiplier design based on ancient Indian Vedic Mathematics,” International SoC Design Conference, pp 65-68, 2008.

[14] Hanumantharaju M.C, Jayalaxmi H, Renuka R. K, Ravishankar M,‘‘A High Speed Block Convolution using Ancient Indian Vedic Mathematics,” International Conference on Computational Intelligence and Multimedia Application , pp 169-173, 2007.

[15] Shamain Akhter, ‘‘VHDL Implementation of Fast NxN Multiplier based on Vedic Mathematics”, Jaypee Institute of Information Technology University, Noida, 201307op, India, IEEE 2007.

[16] Dr. K. S.Gurumurthy, M. S. Prahalad, ‘‘Fast and Power Efficient 16x16 Array of Multiplier Using Vedic Multiplication”, International Conference on Computational Intelligence and Multimedia Application, 2006.

[17] Purushottam D. Chidgupkar, Mangesh T. Karad, “The Implementation of Vedic Algorithms in Digital Signal Processing”, Global J. of Engng. Educ., Vol.8, No.2© UICEE Published in Australia, pp 153-158, 2004.

[18] Abhijeet Kumar, Dilip Kumar Siddhi, “Hardware Implementation of 16*16 bit Multiplier and Square using Vedic Mathematics”, Design Engineer, CDAC, Mohali.

A New Approach to Combined under Voltage and Directional Over Current Protection Scheme

G. Chandra Sekhar[1], P.S. Subramanyam[2] and B.V. Sanker Ram[3]

[1]Vignana Bharathi Institute of Technology, Dept.Of EEE, Aushapur, Ghatkesar(M), Hyderabad, India
E-mail: chandu_vbit@yahoo.com
[2]Vignana Bharathi Institute of Technology, Dept.Of EEE, Aushapur, Ghatkesar M), Hyderabad, India
E-mail: subramanyamps@gmail.com
[3]JN.T.U.College of Engineering, Dept.Of EEE, Kukaptpally, Hyderabad, -500075, A.P, India
E-mail: bvsram4321@yahoo.com

Abstract
To study the protection scheme of Three Phase transmission line the authors have developed Logic Based Under Voltage and Directional Over Current relay for three phase system. A Novel method for the development of a logic Based Combined Under Voltage and Directional Over current Relaying Scheme has been presented here for use in Three Phase Systems using Matlab Simulink tool. When there is any fault the fault voltage will be less and also have a phase difference. The over current detection provides for action when there is any increase in magnitude and phase difference preventing maloperation for legitimate change in power factor or tolerable over load. The Highlight of the scheme is that the present Voltage or Current wave forms are being compared with the previous history of the corresponding wave forms of a few cycles continuously so that when fault occurs the faulted current or voltage wave form is compared with the corresponding previous healthy wave form. It also protects the transmission line from both Under voltage and over load current by using a single relay scheme.

Keywords: Protection, Three phase system, , Directional Over Current relay, Under Voltage relay.

Introduction

As The demand for electric power is increasing day by day and requires additional energy sources and additional transmission lines.

Here the authors have developed Logic Based Under Voltage and Directional Over Current relay for three phase system

Compared to Electro Magnetic Relays and Static Relays, Digital Relays are preferred as they act quickly and can be used in Real Time Control of Power System. Digital relaying requires additional calculations and Algorithms. On the other hand use of Logic Based Protection [5] has the advantage of instant action as in hardware and the use of simulation eliminates development of special Algorithms. The work makes use of traditional Amplitude and Phase Comparators but in Simulation.

In this paper the authors propose a logic based scheme to protect the transmission line from both Under voltage and directional over load current by using a single relay making use of comparison of the faulted wave with the immediately preceding healthy wave over a few cycles just before the acceptance of the faulty condition.

This scheme was simulated using SIMULINK of MATLAB software and is tested for various types of faults for both unsymmetrical faults and symmetrical faults. The results obtained in this simulation are up to the expectations.

This scheme acts if there is any directional over load and also if there is any under voltage magnitude with phase difference. When there is any fault the fault voltage will be less and also have a phase difference. The over current detection provides for action when there is any increase in magnitude and phase difference preventing maloperation for legitimate change in power factor or tolerable over load.

Single Phase Amplitude Comparator for Over Current Protection:

The Scheme for Amplitude Comparison in the case of a Single Phase System making use of different blocks of Matlab Simulink is given in Figure 1 and the corresponding wave forms are shown in Fig.2.

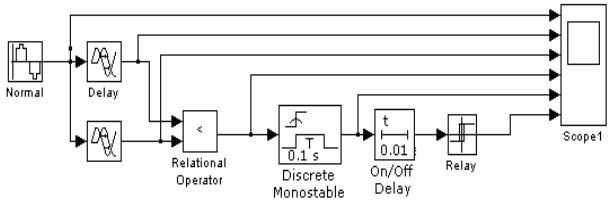


Fig.1. Single phase amplitude comparator for over load protection

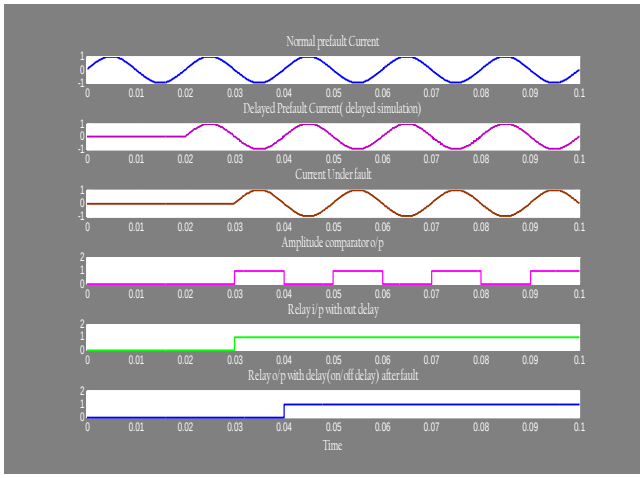


Fig.2 Output wave forms of single phase amplitude comparator with two different delay times

Here only one sine block is used for both prefault and fault currents, but to simulate the fault we use different delays. If we use the same delay times, it belongs to prefault and for fault it is different delay times. For same delay times the relay output is zero. The normal current in a line or bus voltage is taken to be 1.0 p.u for single phase amplitude and also for phase comparators. In the case of bolted faults at any bus with or without fault impedance the prefault bus current is taken as zero.

Normal sinusoidal voltage or current at 50Hz frequency is taken as the reference waveform and represents the healthy condition. The same wave form is extended after the occurrence of disturbance up to the end of simulation period comprising of a few cycles to represent the previous history of healthy condition for comparison with faulty condition. The delayed Sinusoidal Wave Form is considered to represent the Wave Form under Fault current or Under Voltage Conditions which is used for simulation. The prefault bus current will be zero and prefault current in the line will be taken as 1.0 p.u. Under shunt fault with no fault impedance the fault bus voltage will be zero. It should be remembered that there is one and only one Wave under two different conditions of healthy and fault conditions.

The initial delay is given as one cycle arbitrarily. The simulation is taken for minimum of five cycles so as to have four cycles of previous history for comparison. Fault is assumed to occur at the end of the first cycle from the start of simulation only for simulation purpose. It can be after that instant also before the simulation time is over.

The relay has to detect the unhealthy condition in one or less cycle after the fault occurs.

The delay for the second wave form gets automatically decided by the time of occurrence of the fault. (i.e, the instance of occurrence of the fault in the wave cycle)

If there is any difference between the two delays used for simulation purpose it does not matter. This condition is relevant if the time of occurrence of the fault is different from the positive zero crossing of the faulty wave.

Even if there is any indication of relay action before the stipulated delay for the fault wave form in the output, it doesn't matter, because such a condition cannot occur in practice since there is one and only one current or voltage wave form. The second wave form shown in the figures is of the same wave under fault condition shown separately for comparison with the previous history of the first wave form over integral number of cycles.

Even though the phase difference between two waves is zero, if one of the wave forms is shifted, amplitude difference takes place and the relay works. If there is no fault there will be only one wave form, i.e. normal wave and there is no question of amplitude or phase Comparison. The delay in the fault wave form shown here is only meant for simulation and the delay for both wave forms should be same for phase or amplitude comparison and hence comparison takes place.

The relay delay (on/off delay) has no reference for the delay in the waveforms, and this delay is necessarily to be less than one cycle in the case of faults for quick action and it functions as restraining coil torque in the electro Magnetic relay.

By giving sufficient delay mal-operation of the relay for normal and legitimate over loads for a short duration can be prevented in the case of over load protection.

For simulation purpose the fault waveform has to be necessarily shifted only by an integral number of cycles, because there is only one and the same wave form whose nature changes under fault.

In the case of Amplitude Comparison, the amplitude of the normal current wave is compared by the

Relational operator Block with the amplitude of the current wave under fault which will be much larger and gives a digital output 'one' when there is a fault. The Discrete Monostable Block extends the output of the relational block till the end of the simulation time by giving parameter 'pulse duration' equal to simulation time. Under fault, the fault current will be much larger, many times the normal full load current.

Single Phase under Voltage Relay

In the case of under voltage relay, since the normal voltage is compared with fault voltage, the relational block will have the > sign operator since the amplitude of the normal voltage will be greater than the voltage under fault. The bus voltage under fault will be either zero for direct shunt fault with no fault resistance or much less than the normal voltage.

For other under voltage conditions the under voltage will have less amplitude compared to normal voltage. The simulation diagram for Single Phase Under Voltage relay and the corresponding wave forms are shown in Figs.3 & 4 respectively.

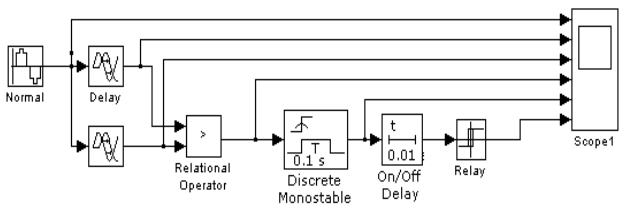


Fig.3 Single Phase Under Voltage Relay

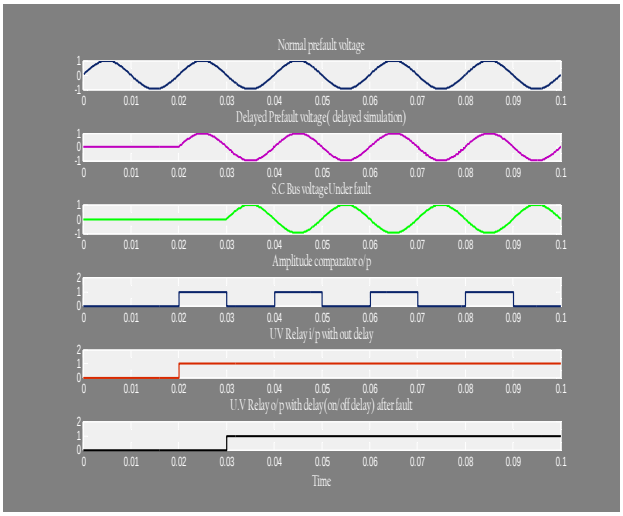


Fig.4 Output wave forms of single phase Under voltage Relay

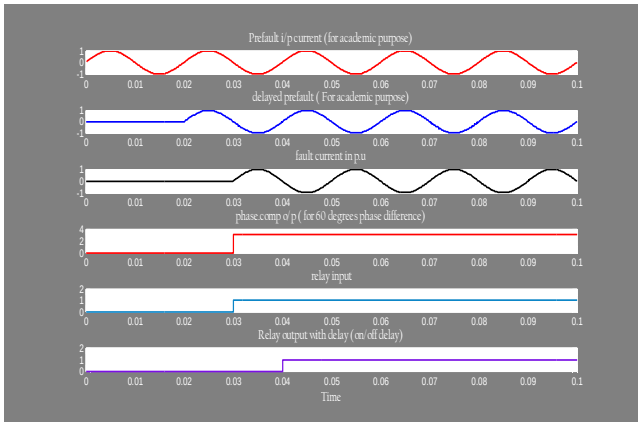


Fig.6 Output wave forms for Phase comparator of Single Phase System

Phase Comparator For Single Phase.

The scheme for phase comparator is similar to the amplitude comparator except that the relational block is replaced by Complex Phase Difference Block which can be obtained from "Communications block set/Utility blocks" tool box. Here also the phase of the first wave form is compared with the phase of the second wave form. If the phase difference is 180^{o} or Π radians this will cause the Relay to act as Reverse Current Relay.

In the Phase Comparison used for Line Protection, the receiving end fault current may have a phase difference with reference to the sending end current.If the phase difference is 180 degrees we have a reversre current relay. The Phase Comparator gives the directional feature to the relay.

Phase comparator need not be used for Under Voltage Relay as the voltage at the faulted bus will be zero for direct shunt fault with no fault resistance. The simulation diagram for phase comparator of Single Phase System and the corresponding wave forms are shown in Fig.5 & 6 resp.

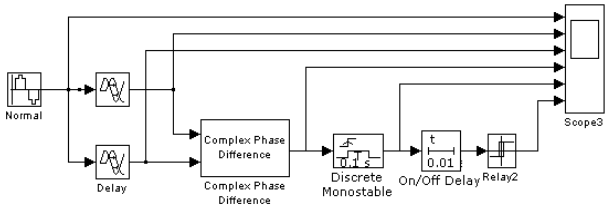


Fig.5 Phase Comparator for Single Phase System

Directional Over Load And Under Voltage Relay For Three Phase System.

The simulation diagram shown in fig.7 gives the protection of three phase system from under voltages and over load currents. Here the authors proposed a unique scheme of logical protection from under voltage and over load currents for a three phase system.

In the case of three phase system for line comparison of any single phase the normal prefault bus current is taken to be 1.0 p.u. In the case of direct shunt fault at bus the prefault bus current I_f =0. But it has been taken as 0.0001p.u for simulation purpose instead of zero.

In the case of three phase system for line comparison of any single phase the normal prefault bus voltage is taken to be 1.0 p.u.In the case of direct shunt fault at bus the prefault bus voltage V_f =0. But it has been taken as 0.0001p.u for simulation purpose instead of zero.

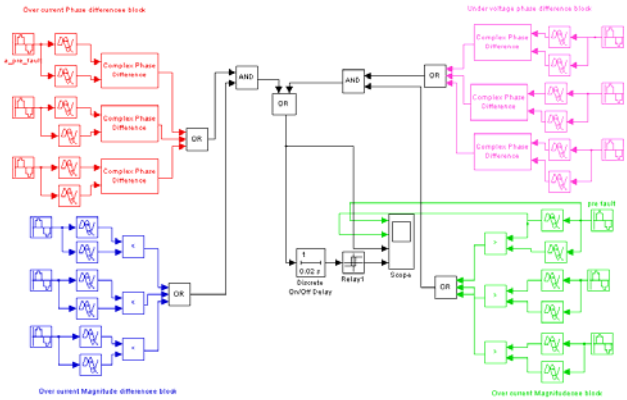


Fig.7. Three phase Under voltage and Over load Protection scheme

The outputs of both magnitude and phase comparators is given to Logic AND gate because the relay will give trip signal when there is amplitude and phase difference preventing maloperation during legitimate change in load power factor. The output of Under voltage scheme and Over

load schme is given to Logic OR gate so that the relay will give trip signal when either or both of the two exists. The above scheme has been studied for Unsymmetrical (L-G) fault [7] and the results obtained are on expected lines. The wave forms after simulation are shown in Fig.9. The proposed circuit is also counter checked by giving fault values as prefault values for getting no signal to the relay.

Example

The single line diagram for the sample problem [7] is shown in Fig.8 and the line and generator data are given in Table No.1 and Table No.2.(All parameters are expressed in p.u).
The Protection scheme for the same problem has been checked and the respective wave forms are presents here. here for unsymmetrical L-G fault at bus 3.

Table.1

Bus code	Self Impedance			Mutual Impedance	
	Positive	Negative	Zero	Zero Sequence	Coupling Element
1-2(1)	0.05	0.05	0.1	0.05	1-2(2)
1-2(2)	0.05	0.05	0.12	0.05	1-2(1)
2-3	0.06	0.06	0.12	-	-
1-3	0.1	0.1	0.15	-	-

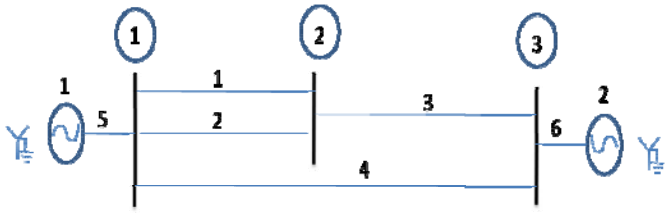


Fig.8 Single line diagram for sample problem

Table.2

Generator Number	General Reactances Impedance		
	Positive	Negative	Zero
1	0.25	0.15	0.04
2	0.2	0.12	0.02

After doing the short circuit analysis the fault currents at bus 3 for Unsymmetrical L-G fault are found [7] to be as

$$I_F^{a,b,c} = \begin{pmatrix} 14.265\angle 0^0 \\ 0 \\ 0 \end{pmatrix}$$

The phase components of bus voltage at bus3 during single L-G fault is

$$V_{3(F)}^{a,b,c} = \begin{pmatrix} 0 \\ 0.6942\angle -100^0 \\ 0.6942\angle 100^0 \end{pmatrix}$$

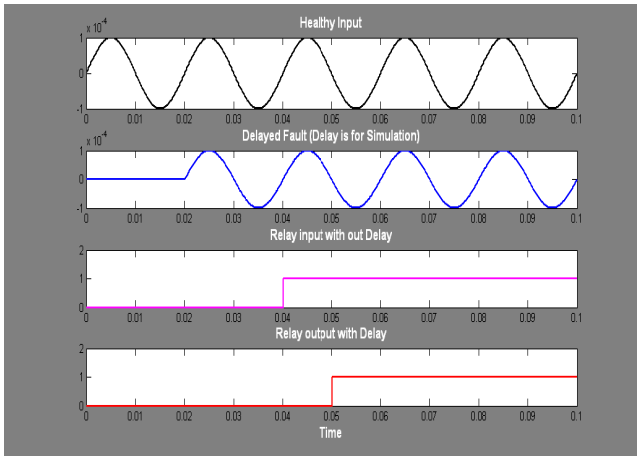


Fig.9 Output wave forms for Three Phase directional over current Protection scheme (for L-G fault at bus 3)

Conclusion:

The proposed scheme consists of Comprehensive Logic Based Protection Schemes for thee Phase System which will be comparable or better to the existing Schemes.

Traditional Amplitude and Phase Comparison had been adapted to apply for Logic Based Protection Schemes.

Logic Based Protection has the advantage of instant action as in hardware and the use of simulation eliminates development of special Algorithms.

The Highlight of the scheme is that the present Current or Voltage wave forms are being compared with the corresponding wave forms of a few cycles continuously so that when fault occurs the faulted current or voltage wave form is compared with the corresponding previous history of the healthy wave form.

This scheme was simulated using MATLAB and is tested for various types of faults for both unsymmetrical faults and symmetrical faults. First the Scheme had been developed for Single Phase System and then extended to Three Phase System. The results obtained in this simulation are up to the expectations.

Instead of using two different relays for under voltage and over load current, this scheme gives better results to protect the transmission line from both Under voltage and over load current by using a single relay.

This Scheme Can be extended to work for Logic Based Protection of Six Phase System.

References

[1] H.C. Barnes, L.O. Barthold, " High phase order power transmission", Presented by *Cigre Sc. Electra No.24, 1973, pp. 39-153.*

[2] P.S.Subramanyam, "Contributions to the analysis of six phase system" *Ph.D. Thesis, IIT, Madras, March 1983.*

[3] P.S.Subramanyam, A. Chandra Sekharan, S.Elangovan, "Dual three phase transformation for comprehensive fault analysis of as six phase system", *Electric power systems research, 1997, Paper No. EPSR 1113.*

[4] J.R. Stewert, D.D. Willems, " High phase order transmission- A feasibility analysis part-I steady state considerations, Part-II- Over voltages and insulation requirements, " *IEEE Trans.On PAS, Vol.91No.6, Nov/Dec.1978, pp.2300-2317.*

[5] G.Chandra Sekhar, P.S.Subramanyam, B.V.Sanker Ram, "Logic based detection of Negative sequence currents for six phase system *"International Journal of Applied Engineering Research", ISSN 0973-4562, Vol 6, Number 6(2011), pp.1311-1322.*

[6] S.S. Venkata, W.C. Guyker, W.H. Booth, L. Kondragunta, N.K. Saini, E.K. Stanek, " 138kV six phase transmission system-Fault analysis", *IEEE Trans. On PAS, Vol.101, No.5, May 1982, pp.1203-1218.*

[7] "Computer Techniques in Power System Analysis" by M.A.Pai, Tata McGraw-Hill Publishing Company Ltd, First Edition 1980, pp.103-109.

About the Author

G. Chandra Sekhar received his B.E(EEE) in the year 1998 from Andhra University and M Tech in High Voltage Engineering in the year 2001 from JNTU College of Engineering, Kakinada, E.G(Dt), AP, India.He is Pursuing Ph.D from JNTU, Kukatpally, Hyderabad. He has published three research papers in International Journals. His area of interest includes Electrical Power systems, Electrical Machines, Electrical Circuits and Multiphase transmission systems.

Mr. Chandra Sekhar is life member of Indian Society for Technical Education(ISTE) and Member of IEEE.

P S Subrahmanyam received his Bachelor of Engineering in Electrical Engineering from Andhra University & Masters Degree in Electrical Power Systems from Jawaharlal Nehru Technological University. He received his PhD from IIT Madras.

He published a number of papers in National and International Journals and several text books. Basically from Electrical Engineering discipline, he cross migrated to the field of Computer Science and Engineering. His areas of interest are Power Systems including Six Phase Systems, Six Phase Induction Motors and Power electronics.

Dr. Pisupati Sadasiva Subramanyam is a fellow of The Institution of Engineers (India), Fellow of National Federation of Engineers, Senior Member of IEEE, Member of Computer Society of India, and Member of Indian Society for Technical Education

B.V. Sanker Ram received his Bachelor of Engineering in 1982 and Master of Technology(Power systems) in 1984 from Osmania University.He received Ph.D in 2003 from JNTU, Kukatpally, Hyderabad.He published more than 60 papers in National and international Journals. His area of interest is Power electronics, FACTS, Reliability Engineering and Control Systems. Sanker Ram is life member of Indian Society for Technical Education(ISTE).

Structural and Optical Properties of Pure & Aluminium Doped ZnO Thin Films Prepared by Sol-Gel

Neha Aggarwal*, Vijay Kumar Anand*, Kiran Walia* and S.C. Sood*

**Ambala College of Engineering & Applied Research, Devasthli, Ambala 133101, Haryana, India*
E-mail: neha.aggarwal@rediffmail.com, ervijay2222@gmail.com, walia.kiran@gmail.com, soodace@gmail.com

Abstract
Thin film of zinc oxide was prepared by spin coating on pyrex glass using zinc acetate dehydrate, 2-methoxyethanol and monoethanolamine (MEA) as a precursor, solvent and stabilizer respectively. Also, Aluminium-doped thin film of ZnO was prepared by using $AlCl_3$. Deposited thin films were investigated for structural and optical properties using X-ray diffraction (XRD) and UV–VIS-NIR spectrophotometer respectively. Results of XRD prove that thin films have polycrystalline nature and possess typical hexagonal wurtzite structure. It is observed that compared to pure ZnO thin film, the grain size in the Al-doped thin film increases. Optical transmission spectrum illustrate that the thin films were transparent in the visible region and gets absorbed in the UV region.

Keywords: ZnO; AZO; Sol-gel; spin coating; thin films; Annealing Temperature; XRD

Introduction

Zinc oxide (ZnO) is an II-VI compound semiconductor material with a direct-wide band gap of 3.36 eV [1] and exciton binding energy of 60 meV [2] at room temperature. This exciton binding energy is much larger than the room temperature thermal energy (26 meV), suggesting that the electron-hole pairs are stable even at room temperature. It has crystalline structure of the wurtzite type and with lattice constants a = 3.24 Å, c = 5.19 Å [3]. It is one of the most promising materials for the fabrication of the next generation optoelectronic devices in the UV region. It has potential uses in photo detectors [4], solar cells [5], light emitting diodes (LEDs) [6], gas sensors [7], surface acoustic devices [8], transparent electrodes [9] and heat mirrors [10]. Doping is done to control the ZnO physical properties. Usually, n-type doping is obtained by Al, Ga or In. On the other hand, p-type doping is not easily obtained. Various techniques have been used to deposit undoped and doped ZnO thin films on different substrates, including spray pyrolysis [11], metal-organic chemical vapour deposition (MOCVD) [12], pulsed laser deposition [13], magnetron sputtering [14] and sol-gel process [15]. Among these, the sol-gel technique has several advantages such as deposition of homogeneous, cheaper, large-area thin films at comparatively low temperatures, easy control of chemical composition, ability to produce fine structure and fabrication of thin film at low cost.

Experimental Procedure

Preparation of sol

The precursor solution of zinc acetate $Zn(CH_3CO_2)_2{\cdot}2H_2O$, (0.25 M) was prepared by dissolving in 2-methoxy ethanol ($(CH_3)_2CHOH$). The mixture was vigorously stirred on hot plate with magnetic stirrer keeping the temperature of solution constant at 70°C. The obtained solution was mixed ultrasonically for about half an hour. An equimolar amount of monoethanolamine (MEA) was added to the solution drop by drop which eliminated the obtained precipitates completely. The resultant solution was very clear, transparent and homogenous. The solution was left to age for 24h to obtain optimum viscosity before film deposition.

In a similar manner powdered Aluminium chloride ($AlCl_3$) material as solute was dissolved in above prepared solution of zinc acetate and ethanol to obtain doping solution in order to study the effect of Aluminium concentration on the structural & optical properties of ZnO thin films.

The pyrex glass slices, after being cleaned with trichloroethylene and methanol, were rinsed with deionised water for 5 min and dried in an oven.

Film deposition

The clear solutions were used for spin coating after 24 hours. Thin films of ZnO and AZO were deposited on 2cm × 2cm pyrex glass at room temperature with a spinning speed of 3000 rpm for 30 seconds. After each coating, the samples were heat-treated at 100 °C for 10 minute, then at 300 °C for 15 minute and gradually cooled down to room temperature before applying a new coating. This preheat treatment is necessary for the evaporation of the organic group content present in the thin film. This process was repeated 16 times to deposit films of the desired thickness. Prepared samples were annealed for an hour at 500°C for decomposition and oxidation of the precursors.

Characterization technique

The structural and lattice parameters of the thin films were investigated with an X-ray diffraction. Diffraction patterns of intensity versus 2θ were recorded with a XPERT-PRO diffraction, using a monochromatized X-ray beam having Cu $K_{\alpha 1}$ radiation with λ = 1.54060 Å (40 mA, 45 kV). A continuous scan mode was used to collect 2θ data from 30° to 70°. The average dimension of crystallites was determined by the Scherrer's method. The optical transmittances of the thin films were investigated using Perkin Elmer Scan-Lambda 750

double-beam UV-VIS-NIR spectrophotometer in the wavelength range from 200nm to 1100 nm.

Results and Discussions

Structural properties

The crystallinity, crystallographic orientation and phase evaluation of thin ZnO & AZO films were found by XRD patterns, as shown in Fig. 1.Peaks observed in XRD results exhibits polycrystalline nature of the films. Peak corresponding to (101) plane is dominating in pure ZnO [16], [17] making it suitable for optoelectronic or UV/blue devices in. Table I shows the comparison of XRD results of standard RRUFF (R060027) [17] with experimental values. It has been observed that the difference between experimental and standard value is within the tolerable range. From the table it has been analyzed, there is a right shift in diffraction peaks which exhibits the lattice is under tensile strain [18].

AZO sample exhibits dominating peak corresponding to (002) plane which is a preferential growth orientation [19], [20]. A qualitative idea of the formation mechanism of the preferential oriented thin films could be the minimization of the surface free energy of each crystal plane, and usually films grow so as to minimize the surface free energy [21]. It has been determined that when the intensity of the diffraction peak of (002) is high, the electrical properties such as mobility and resistivity are improved [22].

Table II shows that XRD peak shifts to higher 2θ value for AZO due to the smaller radius of Al^{3+} ions (0.53 Å) compared to Zn^{2+} ions (0.75 Å). Hence Al^{3+} ions only substitute Zn^{2+} ions and are not found at interstitional site [19].

The lattice constant (a, c) and grain size calculated by the Debye Scherrer's formula [23] are listed in Table III. These values are slightly varying than the values given in (JCPDS #36-1451). It is observed that grain size of Aluminium doped thin film is larger than the pure ZnO thin film.

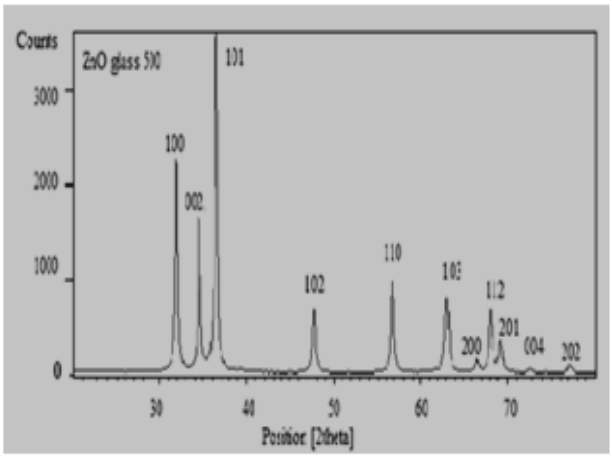


Fig. 1 (a) X-ray diffraction pattern of ZnO thin film on pyrex

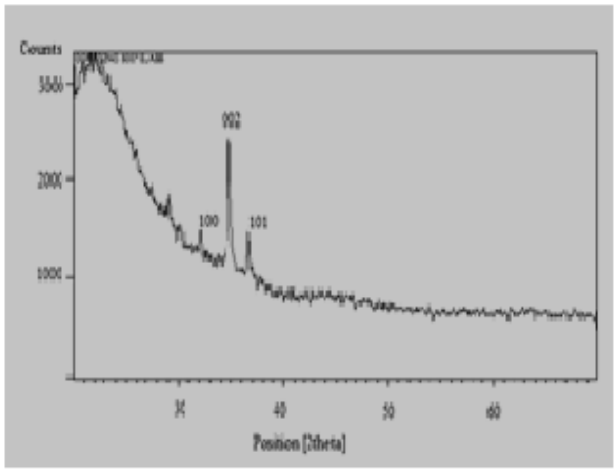


Fig. 1 (b) X-ray diffraction pattern of AZO thin film on pyrex

Table I: The standard rruff data (Zincite_R060027) compared with the ZnO sample XRD pattern

Standard rruff data Zincite_R060027			Experiment			Error	
Plane (hkl)	2θ (deg)	d (Å)	Plane (hkl)	2θ (deg)	d(Å)=nλ/2sinθ n=1, λ=1.54060 Å	d% error	2theta % error
100	31.82	2.8148	100	31.929	2.803	0.3438	0.4214
002	34.474	2.6036	002	34.581	2.5939	0.3112	0.3746
101	36.301	2.4764	101	36.413	2.4675	0.3088	0.3625
102	47.589	1.9113	102	47.698	1.9067	0.2305	0.2431
110	56.642	1.6251	110	56.698	1.6236	0.099	0.0971
103	62.905	1.4774	103	62.99	1.4757	0.1358	0.117
200	66.417	1.4075	200	66.466	1.4067	0.0739	0.0555
112	67.988	1.3787	112	68.051	1.3778	0.0928	0.0694
201	69.128	1.3587	201	69.169	1.3582	0.0589	0.0379
004	72.607	1.3019	004	72.733	1.3002	0.1737	0.1316
202	77.005	1.2381	202	77.083	1.2363	0.1012	0.1448

Table II: Comparison of diffraction peak position

Plane	ZnO	AZO
002	34.581	34.8301
101	36.413	36.6466

Optical properties

Fig. 2 shows UV–VIS-NIR absorption spectra of ZnO & AZO thin films. The optical transmittance is about 96% for ZnO thin film. However transmittance decreases to 72% for Aluminium doped film in the visible range (400-800 nm). The

ZnO thin film shows sharp absorption edge at 365.5 nm while AZO thin film at 358.6 nm. The threshold of optical absorption shifts toward lower wavelengths for Aluminium-doped ZnO thin films, suggesting an increase in the band gap due to Aluminium doping. This shift of the absorption edge of Al-doped crystalline films is attributed to the poor crystallinity of the films and also to increase in disorder with Al doping [24]. There have been reports of this kind of shift according to which increase of the carrier concentration due to Al doping results in a shift of the fermi level and block some of the lowest states, thereby causing widening of the bandgap. Using the relation

optical band gap of the ZnO and AZO films were found to be 3.392eV and 3.458 eV respectively.

Table III: lattice parameters for ZnO and AZO thin film

Substrate	Film	a	c	D
Glass	ZnO	2.8 A	4.93 A	35.7 nm
Glass	ZnO: Al	2.97 A	5.14A	55nm

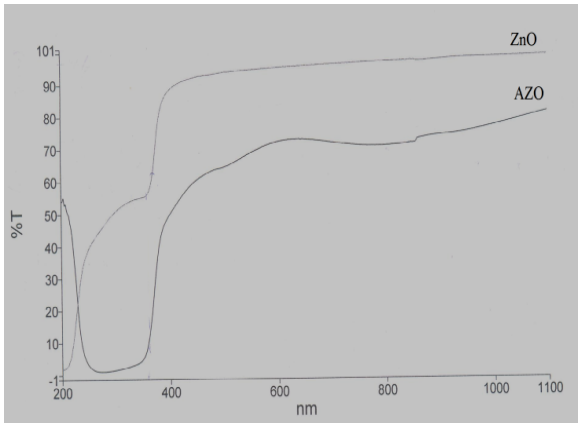


Fig. 2 Transmittance of ZnO and AZO thin films

Conclusions

Structural and optical properties of pure ZnO and AZO thin films deposited on glass substrates using sol-gel spin coating technique were studied. The thin films are found to be polycrystalline in nature. The AZO film exhibit (002) hexagonal c-axis preferred orientation perpendicular to the substrate. In addition, ZnO and AZO has a direct band-gap, (Eg) of about 3.39 eV & 3.45 eV respectively. The optical transmittance was about 72% and 96% in the visible and near IR regions for AZO and ZnO films. Thus making the films suitable for optoelectronic devices, for instance as window layers or as antireflection coating in solar cells. Further we propose to extend the study to intermediate temperature range [25] so that optimum growth conditions can be found which can be used in its device applications such as in solar cells.

Acknowledgement

The authors would like to thank Dr. J.K.Sharma, Director, ACE, Sh. Nalini Kant; mentor of the college, and Dr Jaidev, Chairman, for their kind support & motivation, without their support & encouragement, work was not possible.

References

[1] Bixia Lin, Zhuxi Fu, Yunbo Jia, *Applied Physics Letter*, 79 (2001), 943
[2] Srikant V., Clarke D.R., *J. Appl. Phys.*, 81 (1997), 6357
[3] Minami T., Nato H., Takata S., *Thin Solid Films*, 124 (1985), 43
[4] Chopra K. L., Major S., Panday D.K., *Thin Solid Films*, 102 (1983)
[5] Hupkes J., Rech B., Kluth O., Repmann T., Zwaygardt B., Muller J., Drese R., Wuttig M., *Sol. Energ. Mat.Solar Cells*, 90 (2006) 3054
[6] Jeong W.J., Kim S.K., Park G.C., *Thin Solid Films*, 506-507 (2006), 180
[7] Suchea M., Christoulakis S., Moschovis K., Katsarakis N., Kiriakidis G., *Thin Solid Films*, 515 (2006), 551
[8] Water W., Chu S.-Y., Juang Y.-D., Wu S.-J., *Mater. Lett*., 57 (2002), 998
[9] Michelotti F., Belardini A., Rousseau A., Ratsimihety A., Schoer G., Mueller J., *Non-Cryst. Solids*, 352 (2006), 2339
[10] Kubo M, Oumi Y, Takaba H, Chatterjee A, Miyamoto A, Kawasaki M, Yoshimoto M and Koinuma *Phys. Rev.* B61 16187, H 2000
[11] Nunes P., Fortunadeo E., Martins R., *Thin Solid Films*, 383 (2001), 277
[12] Roth A.P., Williams D.F., *J. Appl. Phys*., 52 (1981), 6685.
[13] Lu Y.F., Ni H.Q., Mai Z.H., Ren Z.M., *J. Appl. Phys*., 88 (2000), 498
[14] Jiang X., Wong F.L., Fung M.K., Lee S.T., *Appl. Phys. Lett.*, 83 (2003), 1875
[15] Jimenez-Gonzalez A.E., Urueta J.A.S., Suarez-Parra R., *J. Crystal Growth*, 192 (1998), 430
[16] Ghaida S. Muhammed, "Efficiency enhancement of crystalline silicon solar cell by the deposition of undoped ZnO thin film", *Indian Journal of Science and technology* Vol. 4 No. 6 (June 2011)
[17] Kihara K, Donnay G, The Canadian Mineralogist http://rruff.info/RRUFF ID: R060027 23 (1985) 647-654. 401.
[18] Jan News Letter 2009.pdf
[19] Hyunchul Oh, JohannesKrantz, IvanLitzov, TobiasStubhan, LuigiPinna, ChristophJ.Brabec, "Comparison of various sol–gel derived metal oxide layers for inverted organic solar cells," *Solar Energy Materials & Solar Cells* 95 (2011) 2194–2199
[20] P. Sagar, M. Kumar, R.M. Mehra, "Electrical And

Optical Properties Of Sol-Gel Derived Zno:Al Thin Films," *Materials Science-Poland*, Vol. 23, No. 3, 2005

[21] Sumetha Suwanboon, "The Properties of Nanostructured ZnO Thin Film via Sol-Gel Coating," *Naresuan University Journal* 2008; 16(2):173-180

[22] Young Baek Kim1, Bum Ho Choi, Jong Ho Lee, and Jin Hyeok Kim, "Morphological and Electrical Properties of Self-Textured Aluminum-Doped Zinc Oxide Films Prepared by Direct Current Magnetron Sputtering for Application to Amorphous Silicon Solar Cells," *Japanese Journal of Applied Physics* 50 (2011) 06GG09

[23] A.R. West, "Effect of crystal size on the powder pattern-particle size measurement," *Solid State Chemistry and Its Applications*, John Wiley & Sons, New York, 1984, pp 173–175 (Chapter 5.6.5)

[24] S Tewari and A Bhattacharjee, "Structural, electrical and optical studies on spray-deposited Aluminium-doped ZnO thin films," *Pramana journal of physics* Vol. 76, No. 1, January 2011 pp. 153–163

[25] Hiroyo Segawa, Hideaki Sakurai, Reiko Izumi, Toshiharu Hayashi, Tetsuji Yano, Shuichi Shibata, "Low-temperature crystallization of oriented ZnO film using seed layers prepared by sol–gel method," *J Mater Sci* (2011)

Eigen Frequency Analysis of High – G MEMS Accelerometer with and without Packaging

T. Sampath and G. Dharani Bai

[1]*Asst. Prof., Lovely Professional University, Punjab, India*
E-mail: sampath.t78@gmail.com
[2]*Associate Prof., VIT University, Vellore, India*
E-mail: gdharanibai@vit.ac.in

Abstract
In this work the modeling and simulation of MEMS High g accelerometer for with and without packaging has been done. The micromechanical structures such as accelerometers have moving parts and narrow gaps and the response of the moving parts are affected by gas in the gaps, so called squeeze film effect. The Eigen frequency analysis of high g MEMS accelerometer was simulated using finite element method based on packaging and without packaging model. In order to understand the effect of adhesive material, which is filled in the gap between the sensor chip and the package bulk, on the output of frequencies, a simplified packaged structure has been adopted in this analysis. The results from the simulations show that Young's moduli of seal adhesive have important influences on the Eigen frequency of packaged accelerometer.

Keywords: High g MEMS accelerometer, squeeze film damping, Packaging.

Introduction

High-g accelerometers fabricated by advanced silicon micro-machining technology have been widely used in many harsh occasions including collision, explosion or impact by the advantage of its small-volume, low-cost, high-precision, good reliability and being prone to mass production, in which the measured shock acceleration usually can be so high as ten thousand gravities[1]. High g MEMS accelerometers are desirable for many commercial, military, and space applications. The high g accelerometers will endure high impact loads up to 100-200kG (1G=9.81m/s^2). MEMS high amplitude shock accelerometers should be capable of measuring long duration transient motion, as well as respond to and survive extremely fast rise times, typical of high g shock event.

High g MEMS Accelerometer design is intended to full fill the most demanding aerospace, industrial, and commercial application requirements in Smart fuzes, Penetrator tests, Weapons data recorders/launch characteristics, Explosives environments (pyroshock), metal to metal impact/armor piercing, Blast loading of structures/Nuclear blast survivability. MEMS technology has found its principal application in sensor technology, but the majority of military and aerospace applications are still too unique to be satisfied by the very low cost MEMS automobile air bag type of accelerometers. At present, a business area is developing that will fill in the gap between the low-volume, higher cost, test-and-evaluation market and the high-volume,very low cost MEMS commercial market.

Packaging is another path toward cost reduction. An accelerometer designed for test and evaluation applications is placed in a housing that allows it to be attached to the surface of a component or system. This housing serves to provide isolation from electrical and mechanical interference from the operating environment. High amplitude shock accelerometers are available in both single axis and triaxial configurations.

Design Methodology

Cantilever Beamstructure Design

With the development of micromachining and package technologies, silicon micromachining accelerometers are increasingly being applied in the fields of sensing of automobile crashes, control of vehicle dynamics, and high-g-force impact assessment. The accelerometers are required to have many features such as wide frequency response, high sensitivity, and low cost. It is important to protect the sensor from resonance, since when the resonance of such a device is excited, the output signal is suspected. Piezoresistive accelerometers are characterized with low mass, extremely small size, and unique construction of the element, which blends an exceptionally high resonance frequency. During the design, fabrication and packaging of high- g accelerometers the care should be made on, Effective overload protection and Stress free mounting of the sensing element.

These can be achieved by designing the special structures. In this work the model of high g MEMS accelerometer designed and fabricated by Dongxiang et al [2] is used to study the effects of squeeze film damping and the packaging on the shock response performance. The 3D schematic of the sensing element of this accelerometer is shown in the figure 2.1.

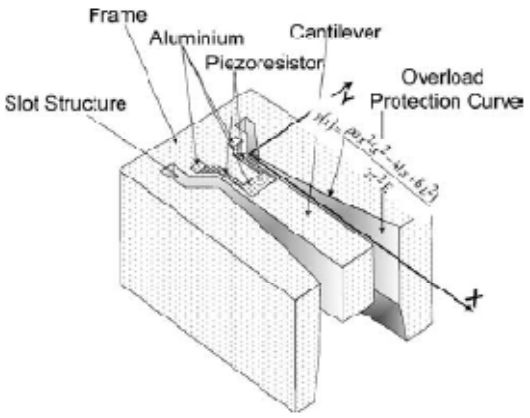


Figure.2.1. Scheamaticstructureof lever of accelerometer

In High g environments, pyroshock refers to short-duration, high-amplitude, high-frequency, transient structural responses in aerospace vehicles, on rocket or missile systems is attributable to explosive bolts and nuts, pin pullers, separation of spent rocket booster stages, linear cutting of the structure, and other actions that produce a near-instantaneous release of strain energy. To support structural analysis of typical military and aerospace systems, a 20,000 Hz frequency response is always more than adequate. To meet this requirement the accelerometer sensing component is designed as per the parameters listed in Table 2.1

Table 2.1 Design Parameters of the sensing element

Parts of sensor	Value
length of cantilever(μm)	515
thickness of cantilever(μm)	16
Density of silicon($\frac{Kg}{m^3}$)	2330
Elastic modulus of silicon(GPa)	170
Piezoresistive coefficient(Pa^{-1})	60 X 10^{-11}

According to Rayleigh–Ritz method [3], the fundamental frequency of cantilever at the sensitive direction can be expressed as

$$f_0 = 1.019\frac{h}{2\pi L^2}\sqrt{\frac{E}{\rho}} \tag{2.1}$$

The fundamental frequency of the cantilever beam is predicted from equation.2.1 is 47.007 KHz. Indicating that the accelerometer sensing component would meet the requirement of high g application

The 3D structure of the cantilever beam is designed with the curved protection of sensing element capable of enduring high g shock pulse. In this work a shock pulse of amplitude $1x10^5$ g and 50μsec duration is used to study the transient response of the accelerometer. Based on the uniform load mechanical model [3], according to the coordinate system as shown in Figure(2.1), the displacement of cantilever along the *x*-axis can be express as

$$Y(x)=\frac{\rho a x^2 - 4Lx + 6L^2}{2T^2E} \tag{2.2}$$

Displacement of the tip of the cantilever beam predicted by the equation (2.2) is 17.5μm for an applied acceleration of $1x10^5$ g. Distance of the cantilever beam surface along X-direction to the curved stop is formulated as

$$H(x) = \frac{\rho a x^2 - 4Lx + 6L^2}{2T^2E} + p \tag{2.3}$$

The curved stop of the structure is designed in such a way that its distance to the cantilever surface at the root is equal to PSD (0.5 μm) and at tip is 18 μm. To study the effect of damping gap the PSD is varied in steps of 0.5μm from 0.5μm to 2.0μm.

Design of Accelerometer with package

In this work finite element simulation has been applied in frequency-domain and time-domain analyses for a packaged accelerometer used in high-G environments. In order to understand the effect of the adhesive material, which is filled in the gap between the sensor chip and the package bulk, on the output signal of accelerometer, a simplified packaged structure has been adopted in this analysis[5]. In order to obtain the conceptive conclusion, the sensor assembly is simplified to a simple configuration, that includes accelerometer chip, alumina ceramic bulk, and two pieces of seal adhesive fixing the chip in the bulk, in the direction of shock loading as shown in the figure(2.2)

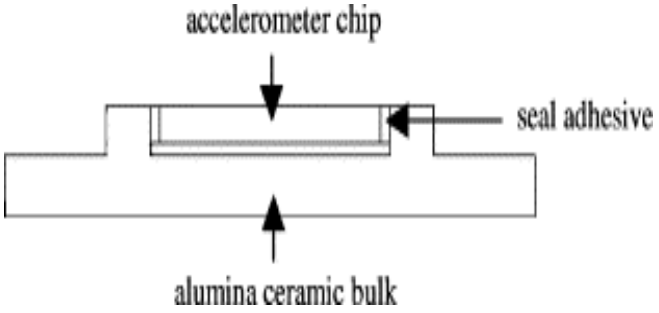


Figure.2.2. Simplifiedconfigurationfor accelerometer with packaging.

The material parameters adopted in the present simulations for each component of the packaging construction are shown in the Table2.2.

Table 2.2 Material parameters for package structure

Material	Young's modulus (Mpa)	Poisson's ratio	Density (g/cm^3)
Ceramic	296000	0.28	3.97
Silicon	131000	0.28	2.33
ADHESIVE 1	0.01	0.30	1.8
ADHESIVE 2	100	0.30	1.8

ADHESIVE 3	1000	0.30	1.8
ADHESIVE 4	9000	0.30	1.8
ADHESIVE 5	16000	0.30	1.8
ADHESIVE 6	296,00	0.30	1.8

The encapsulation resins with relatively high Young's moduli (E ~ 8-12 GPa) correspond to epoxy encapsulation resins with high filled SiO2 particles while the relatively soft encapsulation resins (E< 4 Gpa) correspond to epoxy encapsulation resins with low-filled Si02 particles, or even without SiO2 particles, or silicone. Because the compositions of the encapsulation, resins were, very complex, the densities were designated as 1.8g/cm^3 to study the influence of Young's moduli of the encapsulation.

The ADHESIVE 1 and 2 with Young's modulus *E*=0.01 MPa and 100 MPa are soft seal material, and its mechanical properties are similar to rubber. The Young's moduli of ADHESIVE 3 and ADHESIVE 4 were *E*=1000 Mpa and *E*=9000 Mpa, corresponding to two kinds of plastic molding compounds (low SiO_2 powder filled epoxy and highSiO_2 powder filled epoxy) used normally in IC packaging, respectively. The ADHESIVE 5 (*E*=16,000 Mpa) corresponds to epoxy-based FR4, which was usually used as a material for printed circuit boards (PCBs). The Young's modulus *E*=296,000 Mpa of ADHESIVE 6 was selected to coincide with the superfine alumina ceramics, in order to simulate the situation in which seal adhesive material was the same as ceramic bulk material.

Structure Simulation

The designed accelerometer sensor with and without package is simulated in COMSOL to study the effects of damping gap and packaging material.

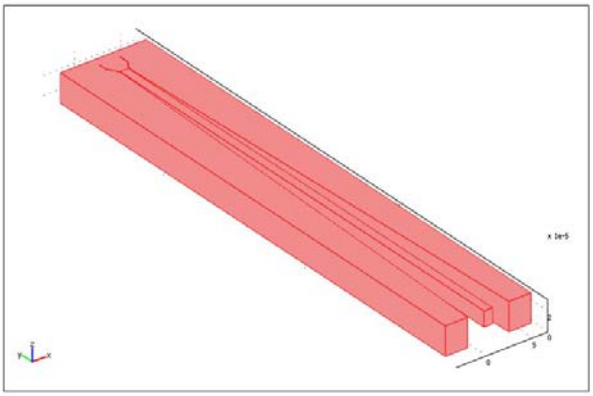


Figure. 2.3 Comsol structure without package

Effects of Packaging

The 3D geometry of the simple packaged structure of the single cantilever beam accelerometer as shown in the figure (2.4) has been built and then finite element simulation was conducted with COMSOL Multiphysics. The boundary condition of this model was that the bottom surface of the package was fixed. Eigen frequency and transient analysis has been conducted on this packaged structure with different sealing adhesive materials and the results are discussed in the chapter 4.

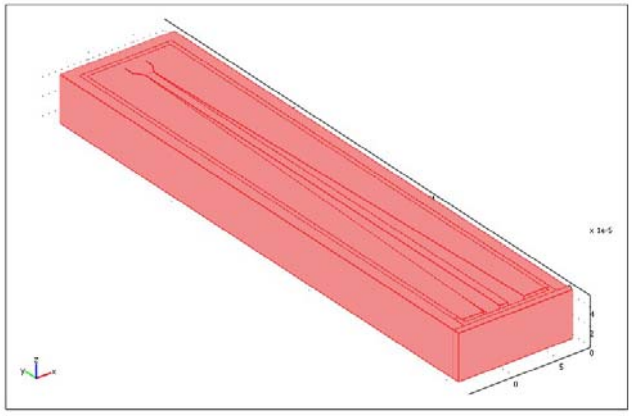


Figure.2.4 Comsol structure with package

Results and Discussion

Natural Frequencies Of Accelerometer Without Packaging

Eigen frequency analysis of the single cantilever beam structure without packaging was simulated and the first 10 natural frequencies are listed in the table.

Table. 3.1 First 10 Natural frequencies of the accelerometer chip without packaging

Modes	Frequency
1	46,171.363052
2	81,331.73589
3	2.892275e^5
4	5.079942e^5
5	8.096055e^5
6	1.4144804e^6
7	1.586071e^6
8	2.194482e^6
9	2.6213e^6
10	2.751176e^6

From the table it is seen that the first mode of frequency correspond to the natural frequency of the sensitive component in the accelerometer. The main vibration frequency of sensitive compound obtained from this simulation, 46.171KHz is very close to the analytical result of Equation.(3.1) The current trend for the design of high-G accelerometer is that the frequency of the lowest natural frequency of main vibration mode of sensitive compound should be raised over 15,000 Hz [6]. And in some applications, the lowest natural frequency should be even higher. The simulation results of frequency-domain analysis

for accelerometer chip without packaging in this work showed that the lowest natural frequency of sensitive cantilever in accelerometer chip was 46.171K Hz, indicating that the accelerometer chip developed would meet the requirement of the high-G application.

Natural Frequencies Of Accelerometer With Packaging

Eigen frequency analysis of the single cantilever beam structure with packaging considering different adhesive materials has been simulated and the first 10 natural frequencies are listed in the table3.2

Adhesives materials have an important influence on the vibration behavior of packaged accelerometer. At the same ordering number of vibration mode, the natural frequency of packaged accelerometer will become higher as the Young's modulus of adhesive materials used in the assembly increases. The frequencies of main vibration mode of sensitive cantilever in accelerometer chip have no obvious variations for different seal adhesive materials and nearly same that of the accelerometer chip without packaging

However as the young's modulus of the sealing material decreases, the ordering number of vibration mode is delayed. This leads to distortion in the output signal of the accelerometer.

Table. 3.2 First 10 natural frequencies of the accelerometer chip with packaging.

Mode	1e4	100	10000	9000	16000	29600
1	22518.85032	46788.80611	46918.27145	46979.60429	46992.15480	47003.74549
2	28572.25874	81007.92603	81306.71737	81688.28331	81760.58250	81819.33071
3	29741.843581	2.999098e5	3.011478e5	3.017234e5	3.018364e5	3.019401e5
4	31152.68503	5.086911e5	5.109304e5	5.139701e5	5.145587e5	5.150416e5
5	33578.76215	9.296526e5	9.362681e5	9.389808e5	9.394862e5	9.399451e5
6	36527.75607	9.572173e5	1.117645e6	1.233134e6	1.249527e6	1.263827e6
7	50455.15010	1.389676e6	1.471606e6	1.483949e6	1.486565e6	1.488779e6
8	84479.29725	1.462776e6	1.772348e6	2.026363e6	2.027872e6	2.029187e6
9	95819.03772	1.932333e6	2.017959e6	2.209924e6	2.211474e6	2.211685e6
10	1.874843e 5	1.984314e6	2.20987e6	2.211814e6	2.351795e6	2.500281e6

Conclusion

Frequency domain analysis showed that seal materials with high Young's moduli effected a less change in Eigen frequencies. In conclusion FR4-like epoxy based adhesives could be considered to be adopted as seal material for the high g sensors packaging.

Future scope

In this study only frequency analysis was studied. For the commercially available accelerometers, an exact model can be developed that includes the protective cap, potting material and header.

References

[1] Zunxian Yanga, Xinxin Lib," Simulation and optimization on the squeeze-film dampingof a novel high-g accelerometer" Microelectronics Journal 37 (2006) 383–387.

[2] Davies, C. Barron, S. Montague, et.al. High g MEMS integrated accelerometer. Proc SPIE 3046 (1997) 52

[3] Dong Jian, Li Xinxin, Wang Yuelin, Lu Deren, Ahat Shawkret. "Silicon micromachined high-shock accelerometers with a curved-surface application structure for over-range stop protection and free-mode-resonance depression". J Micromech Microeng 2002;12:742–6.

[4] Bao Minhang. Micro mechanical transducers — "pressure sensors, accelerators and gyroscopes"[M]. Elsevier: Amsterdam Publisher; 2000.

[5] Weidong Huang, Xia Cai,Bulu Xu,Le Luo,Xinxin Li, Zhaonian Cheng,"Packaging effects on the performances of MEMS for high-G accelerometer with double-cantilevers.Sensors and Actuators A 102(2003)268-278.

[6] P.L. Walter , Trends in "Accelerometer design for military and aerospace application". *Sensors* **16** (1999), pp. 21–25.

Effect of Pressure on Thermal Expansivity of Ionic Solids and Nanomaterials

[1]Nidhi Verma and [2]Dr. Sanjeev Srivastava

[1]Department of Applied Sciences, RPIIT, Karnal, India
E-mail: nidhidawer.1981@gmail.com
[2]Department of Applied Sciences, GIMT, KKR, India
E-mail: sanjeevsrivastava1980@rediffmail.com

Abstract
A simple theory is proposed to predict the effect of pressure on thermal expansivity of ionic solids and nanomaterials which is supported by the theory of thermal expansivity for various substances as formulated by different observers. The results obtained are found to present a good agreement with the experimental data. Here we have considered a number of ionic solids and nanomaterials like Lithium-alumino-silicate, Rb_3C_{60}, nanocrystalline nickel (20 nm, Ni), NaCl, KCl, MgO, CaO, Nanocrystalline iron, α-Fe_2O_3, rutile TiO_2, $Zr_{0.1}Ti_{0.9}O_2$, γ-Si_3N_4, Ni-filled and Fe-filled multiwalled carbon nanotubes etc. A good agreement is observed when results obtained are compared with experimental data which demonstrates the validity of present approach.

Keywords: Nano-materials, Thermal Expansivity, Bulk Modulus, EOS.

Introduction

Nanomaterial is a field that takes a material science based approach to nanotechnology. This term nanomaterials' is sometimes also used for materials smaller than one micrometer. Hence nanomaterials differ from bulk materials by virtue of their small size. Certain numbers of physical properties vary with the change in macroscopic systems. The physical properties of materials depend strongly on the structure and interatomic distances. The materials reduced to the nanoscale can suddenly show very different properties compared to what they exhibit on macroscale, enabling unique applications. The study of nanomaterials with varying pressure can help us to study a wide range of solid state materials. High temperature – pressure applications have the potential to constitute a unique way for the elaboration of new materials in a controlled manner [1].

Several experiments have been performed to study and explain the behavior of nonmaterial by Sharma and kumar [2]. A simple model was developed to study the effect of pressure and temperature on C_{60} solid. Pressure and temperature dependence of volume $V/V0$, bulk modulus B, and coefficient of volume thermal expansion were investigated by varying the temperature from room temperature to 2000 K, and pressure from room pressure to 200 kbar.

Due to high pressure on nanomaterials many effects happen such as pressure ionization, modification in electronic properties etc. For this pressure versus volume relation of condensed matter known as EOS (Equation Of State) is a vital input. As many EOS exist in literature, but still there is a need to explain the effect of pressure at constant temperature for nanomaterials. Lithium-alumino-silicate ceramics have gained considerable commercial attention because of very low thermal expansion, transparency, high chemical durability and strength [3].

The compression behavior of Rb_3C_{60} was first measured by X-ray diffraction under hydrostatic pressures upto 2.8 GPa at 300K by Zhou et al. [4]. The compressibility of Rb_3C_{60} was later re-measured by Ludwig et al. [5] up to 6GPa using a similar technique. To understand the size effect on the bulk modulus and to look for possible new high pressure phases, nanocrystalline nickel (20 nm, Ni) was studied under high pressure by Chen et al. [6]. High pressure compression behaviour of carbon nanotube has been studied experimentally by Tang et al. [7]. Nanocrystalline iron has been the subject of many experimental and theoretical studies [8]. The EOS of nanocrystalline CuO (24 nm) has been studied by Wang et al. [9] up to 16 GPa using high energy synchrotron radiation and Raman spectroscopic techniques . High pressure compression behaviour of α-Fe_2O_3 has been studied experimentally by Clark et al. [10]. Nanocrystalline rutile TiO_2 has been studied using X-ray diffraction at ambient temperature up to 47.4 GPa by Olsen et al. [11]. The study of AlN nanocrystal under hydrostatic condition has been performed by Wang et al. [12] up to the pressure of 36.9 GPa. The high pressure behaviour of γ-Si_3N_4 has been investigated experimentally by Kiefer et al. [13].

To investigate the effect of particle size on the compressibility of MgO, Rekhi et al. [14] performed X-ray diffraction study on MgO with particle size 100 nm. The compression behaviour of Zr-doped nano anatase $Zr_{0.1}Ti_{0.9}O_2$ synthesized by thesol–gelmethod was studied by Holbig et al. [15] using a DAC upto 13 GPa. No phase transition was seen in Zr-doped nano anatase upto a pressure of 13 GPa. The high pressure behaviour of Ni-filled and Fe-filled multiwalled carbon nanotubes has been investigated up to 27 and 19 GPa, respectively, with the help of synchrotron based angle dispersive X-ray diffraction by Poswal et al. [16]. These nanotubes do not show any structural transformation up to the highest pressures studied. Singh and Chauhan [17] analyzed the temperature dependence of thermal expansivity and isothermal bulk modulus in terms of the Anderson–Gruneisen

parameter and the thermal pressure. The analysis was done for ionic solids viz. NaCl, KCl, MgO and CaO in the temperature range starting from room temperature up to the temperatures close to their melting temperatures. An isothermal equation of state was developed by M Kumar and Upadhyay [18] to study the compression of lithium and sodium halides at pressures ranging from zero to their structural transition pressures. The calculations were performed by developing an ionic model based on Harrison's quantum mechanical treatment of overlap repulsive potential. The compression data thus obtained were used to investigate the pressure dependence of the volume thermal expansion coefficient by using the formula due to Anderson derived on the basis of thermodynamic analysis by making some crude approximation. The expansivity, constant-pressure heat capacity, and isothermal bulk modulus of sodium chloride (NaCl) have been obtained by using molecular dynamics method by Q F Chen et. al. [19]. Here, the thermodynamic properties of NaCl were summarized in the pressure range 0–500 kbar and the temperature up to 1000 K. A new relation for predicting volume thermal expansion of alkali halides at high temperatures was derived by Z H Fang [20] based on the assumption that the two different diffusional driving force models presented were equivalent both at room temperature and zero pressure. The thermal pressure for the MgO was evaluated by Srivastava and sharma [21] from room temperature to 3000 K at atmospheric pressure with the help of data suggested by Jacobs and Oonk. Various relationships between thermal pressure and volume expansion ratio were examined. Thermoelastic properties viz. C_{11}, C_{44}, C_S, and K_S were calculated at high temperatures for the solids with the help of Murnaghan and Tallon models. A new relationship was developed by Sinha and Srivastava [22] to investigate temperature dependence of elastic constants and thermal pressure for ionic solids NaCl, KCl, MgO and CaO by using a formulation which is valid up to extreme compression limit.

Although, several experimental studies have been performed to understand the high pressure behaviour of ionic solids and nanomaterials. More- over, the theoretical efforts are lacking. Therefore, it may be useful to present a simple theoretical model to study the behaviour of nanomaterials under high pressure.In the present paper, we present a simple theoretical analysis to study the behaviour of nanosystems and compare it with some earlier formulation and we include the effect of temperature also.

Method of analysis

Sharma & Kumar [1, 2] predict the effect of pressure on various nano-materials with the help of following expression:

$$P = a_1(1-V/V_0) + a_2(1-V/V_0)^2 \quad \ldots\ldots\ldots\ldots \quad (1)$$

When V -> 0 then

$$P = a_1 + a_2$$

$$\text{Where } a1 = K_0 \quad \ldots\ldots\ldots\ldots \quad (2)$$

$$\text{And } a_2 = K_0 (K'_0+1)/2 \quad \ldots\ldots\ldots\ldots \quad (3)$$

Here, K_0 and K'_0 are the bulk modulus and its pressure derivative at ambient conditions. It is clear from equation (1) that the density of the material approaches the infinite value (when v -> 0) at a finite pressure which is given by equation

$$P = K_0 + K_0 (K'_0+1)/2 = \text{finite value} \quad \ldots\ldots\ldots\ldots \quad (4)$$

This pressure is not far above the terrestrial range, it means that the material collapses to infinite density at finite pressure. The equation (1) also represents that the bulk modulus decreases and the first pressure derivative of bulk modulus (K') becomes negative with further pressure yielding $K\infty = 0$ and $K'\infty = 1$ [3]. Here subscript ∞ referred to the value of the concerned parameter at infinite pressure. Various investigators suggested that K' decreases with increasing pressure and approaches to a positive finite value at infinite pressure. Now using the basic definition of the K' as given below:

$$K' = -V/K\ (dK/dV) \quad \ldots\ldots\ldots \quad (5)$$

With the help of equation (5), we can examine the behavior of bulk modulus K at infinite pressure (P -> ∞ or V -> 0) which gives:

$$\int_{V0}^{0} K'dV/V = -\int_{K0}^{K\infty} dK/K = \ln(K_0/K_\infty) \quad \ldots\ldots\ldots\ldots \quad (6)$$

Equation 6 reveals that if K' remains finite at V -> 0 then $K\infty$ -> ∞. This conclusion contradicts the result obtained by equation (1) as it gives $K_\infty = 0$.

It needs not to be explained that infinite pressure extrapolation is necessary to judge the suitability of an EOS (Equation of State). According to Stacey and Devis "Infinite pressure properties are simply EOS parameters, not observable in any conventional solid were to approach infinite compression it would undergo dramatic phase transition to exotic forms and EOS do not carry through phase transitions. However, parameters such as K'_∞ are just as legitimate as physical entities as are zero pressure properties, V_0 and K_0, for high pressure materials that do not survive decompression to P = 0". Since equation 6 does not follow the infinite pressure constraints therefore we disregarded equation 6 in our study and adopted an appropriate EOS. An EOS which follows the necessary boundary conditions imposed by extreme pressure thermodynamics is known as reciprocal K-prime EOS. The reciprocal K-prime EOS is based on the linear relationship between 1/K' and P/K i.e.

$$1/K' = a + b\ (P/K) \quad \ldots\ldots\ldots\ldots \quad (7)$$

Where the coefficient $a = 1/K'_0$

$$\text{And } b = 1 - (K'_\infty/K'_0) \quad \ldots\ldots\ldots\ldots \quad (8)$$

is determined using the condition $K' = K'_0$ at P=0 and the infinite pressure extrapolation condition given by following equation:

$$(K/P)_{P->\infty} = K'_\infty \quad \ldots\ldots\ldots \quad (9)$$

This gives a guarantee that equation (7) is asymptotically valid. Equation (7) has been integrated analytically to find [3].

$$K/K_0 = (1 - K'_\infty P/K)^{-K'_0/K'_\infty} \quad \ldots\ldots \quad (10)$$

And

$$\ln(V/V_0) = K'_0/K'^2_\infty \ln(1-K'_\infty P/K) + (K'_0/K'_\infty - 1)P/K \quad (11)$$

The value of $K'\infty$ can be estimated empirically with the help of next relationship.

$$K'_\infty = 3/5K'_0 \quad \ldots\ldots \quad (12)$$

It is clear that to predict the pressure volume relationship for the material with the help of reciprocal K-prime EOS should know the values K_0 and K'_0. Fortunately, the values of K_0 and K'_0 for nano-materials considered in the study, are available in the literature [5-21].

Results and Discussions

Using the reciprocal K-Prime EOS we have estimated the variation of volume with the effect of pressure. To solve the reciprocal K-Prime equations, we started with the fitting of ratio P/K for which the value of bulk modulus (K) and V/ V_0 are computed with the help of equation (10) and (11). The obtained values of K are used to estimate the value of pressure with the help of non-values of P/K. The required input parameters for various nano-materials considered in the study are given in the Table1. These values are extracted from the literature. The computed values of PV relationship are compared with available experimental data as well as those obtained by equation (1). The comparison is shown in figures (1, 2, 3…) for various nano-materials. Figures (1, 2, 3…) show an agreement between experimental and theoretical values. Although both theoretical approaches considered in the study present good agreement with the experimental but the reciprocal K-Prime EOS should be preferred over the other theoretical approach as discussed above.

Table 1

Sr. No.	Materials	Ko(GPa)	Ko'
1	Fe-filled MWCNT	167	4
2	Ni-filled MWCNT	190.4	4
3	γ-Si_3N_4	339	4
4	AlN (hexagonal)	321	4
5	$Zr_{0.1}$ $Ti_{0.9}$ O_2	213	17.4
6	TiO_2 (rutile phase	211	8
7	3C-SiC (30 nm)	245	2.9
8	TiO_2 (anatase)	190.4	4
9	α-Fe (filled-nanotube)	89.7	20.9
10	α-Fe_2O_3	336	4
11	γ-Al_2O_3 (67 nm)	238	4
12	CuO	81	4
13	MgO	179	1.5
14	ε-Fe (Hexagonaliron	179	3.6
15	γ-Fe_2O_3	374	4
16	Carbon nanotube (individual)	230	4
17	Ni (20 nm)	185	4
18	Rb_3C_{60}	17.35	3.9
19	CdSe (rocksalt phase)	74	4
20	$LiAlSi_2O_6$	71	4.4

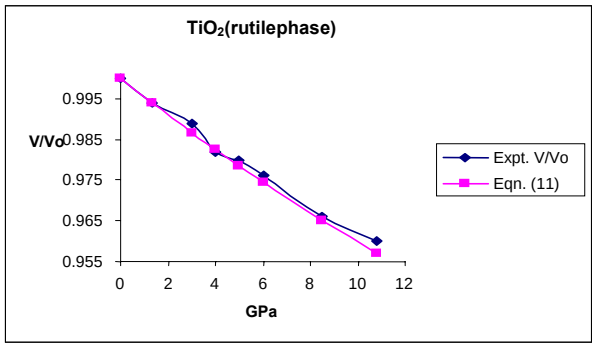


Fig. 1

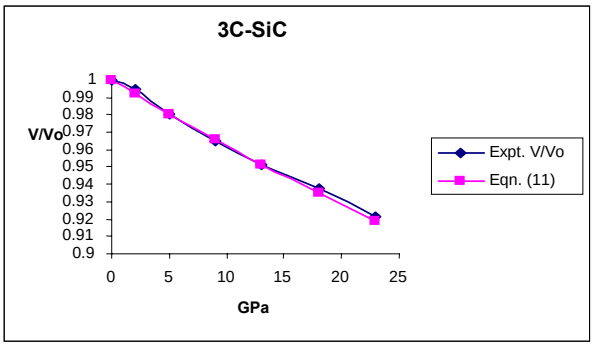


Fig. 2

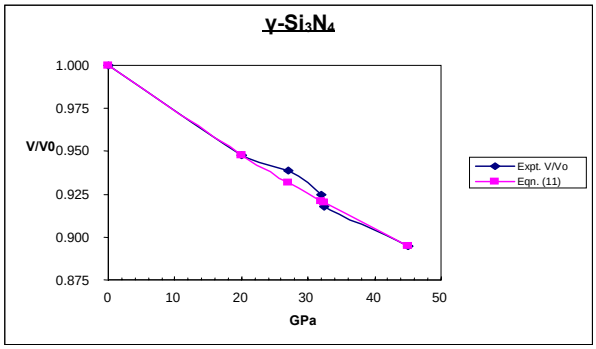


Fig. 3

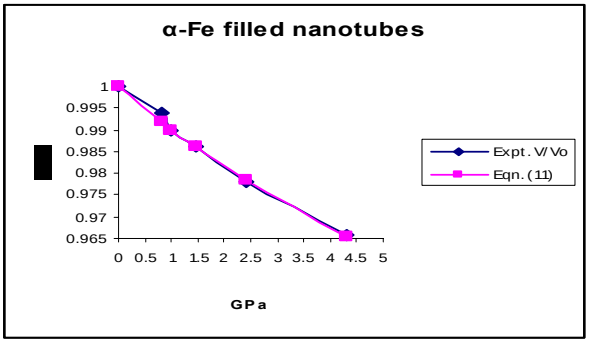


Fig. 4

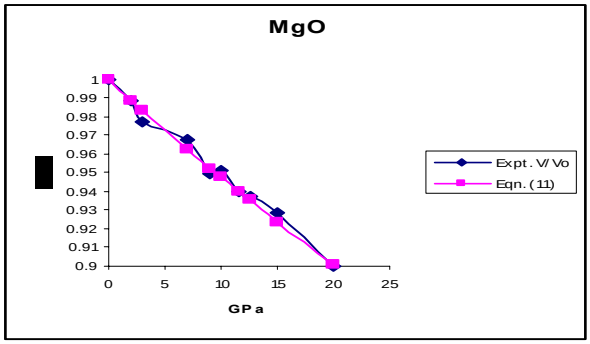


Fig. 5

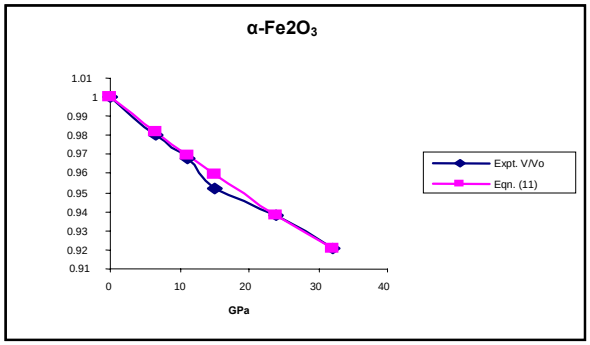


Fig. 6

An important aspect of the present study is to include extreme pressure thermodynamics which is applicable to the interior of the planet is found also applicable to the materials having dimensions in the nano-scales.

At last, it can be concluded, an EOS which is applicable to the extreme compression and to the materials of very high dimensions (The planet earth) is also applicable to the nano-materials. Thus the K-Prime approach can be used from 10^{-6} meter to 10^{6} meters.

References

[1] Uma D Sharma and M Kumar, Physica B, 405, 2820-2826(2010)

[2] Uma D Sharma and M Kumar, IJ PAP, 48, 9, 663-668(2010)

[3] F D Stacey and P M Davis, Phys. Earth Planet. Int., 142, 137-184(2004)

[4] O Zhou, G B M Vaughan, Q Zhu, J E Fischer, P A Heiner, N Coustel, J P J McCauley, A B Smith, Science 255, 833(1992)

[5] H A Ludwig, W H Fietz, F W Hornung, K Grube, B Renker, G J Burkhart, Physica C 234 , 45(1994)

[6] B Chen, D Penwell, M B Kruger, Solid State Commun. 115 , 191(2000)

[7] J Tang, et al., Phys. Rev. Lett. 85 ,1887(2000)

[8] B Chen, D Penwell, M B Kruger, J. Appl. Phys. 89 , 9(2001)

[9] Z Wang, V Pischedda, S K Saxena, P Lazor, Solid State Commun. 121, 275(2002)

[10] S M Clark, S G Prilliman, C K Erdonmez, J Rockenberger, D J Zaziski, J Kwong, Nanotechnology 16, 2813(2005)

[11] J S Olsen, L Gerward, J Z Jiang, High Pressure Res. 22, 385(2002)

[12] Z Wang, K Tait, Y Zhao, D Schiferl, J. Phys. Chem. B 108, 11506(2004)

[13] B Kiefer, S R Shieh, T S Duffy, T. Sekine, Phys. Rev. B 72 , 014102(2005)

[14] S Rekhi, S K Saxena, Z D Atlas, J. Hu, Solid State Commun. 117, 33(2001).

[15] E Holbig, L Dubrovinsky, G S-Neumann, V Prakapenka, V Swamy, Z. Naturforsch 61(b), 1(2006).

[16] H K Poswal, S Karmaker, P K Tyagi, D S Mishra, E Busetto, S M Sharma, A K Sood, Phys. Stat. Solid (b) 244, 3612(2007).

[17] K S Singh and R S Chauhan,Physica B, 3125, 74(2002)

[18] M Kumar and S P Upadhyay, J. Phys. Chem. Sol., 54, 6, 773(1993)

[19] Q F Chen, L C Kai, S Duan, D Q Chen, J. Phys. Chem. Sol., 65, 6,1077-1081(2004)

[20] Z H Fang, Physica B, 357(3-4), 433-438(2005)

[21] S K Srivastava and S K Sharma, Physica B, 373, 2, 258-261(2006)

[22] Pallavi Sinha and S K Srivastava, Physica B, 405, 4, 1197(2010)

[23] A Keane, Aust. J. Phys., 7, 322-333(1954)

[24] L Knopoff, J. Geophys. Research, 68, 2929-2932(1963)

[25] W B Holzapfel, Rep. Prog. Phys., 59, 29-90(1996)

[26] J Hama, K Suito, J. Phys. Condensed Matter, 8, 67-81(1996)

[27] F D Stacey, Geophys. J. Int., 143, 621-628(2000)

[28] F D Stacey, Report Prog. Phys., 68, 341(2005)

Simulation of Indirect Vector Controlled Induction Motor Drive

Kulraj Kaur, SSSR Sarathbabu Duvvuri and Shakti Singh

[1]Electrical and Electronics Engineering Deptt. , LPU, Jallandhar, Punjab, India
[2]Electrical and Inst. Engineering Deptt., Thapar University, Patiala, Punjab, India
[3] Electrical and Inst. Engineering, Thapar University, Patiala, Punjab, India
E-mail: kulraj.15720@lpu.co.in, sarath.duvvuri@thapar.edu, shakti.singh@thapar.edu

Abstract
The paper emphasis a solution for induction motor speed control. In this paper we present the simulation results of vector speed control of an induction motor. In this case, indirect methods were simulated using MATLAB (SIMULINK). A squirrel cage induction motor model was used taking the reference coordinates as the rotor magnetic field. The accuracy of results is given by the precision of motor model used. This paper describes the use of the MATLAB simulation toolbox "SIMULINK" for dynamic modeling of vector controlled motor drive system. The dynamic and transient performance is studied through simulation and the results are presented. The principle of vector control of AC machine enables the dynamic control of AC motors, and induction motors in particular to a performance level comparable to that of a DC machine. The basic equations describing the dynamic behavior of an induction machine in rotating reference frame are designed. Based on these equations the structure of the vector controlled induction motor drive is designed. The procedure is evaluated through extensive computer simulation. The complex nature of the vector controlled scheme places a heavy computational burden on the controller. The power circuit is developed using Insulated Gate Bipolar Transistors (IGBTs). This structure generates the desired reference voltage by acting on both the amplitude and the angle of its components. The motor reaches the reference speed rapidly and without overshoot, load disturbances are rapidly rejected and variations of some of the motor parameters are fairly well dealt with.

Keywords: vector control, induction motor, scalar control, rotor equations, reference frame

Introduction

Modern industrial processes place stringent requirements on industrial drives by way of efficiency, dynamic performance, flexible operating characteristics, ease of diagnostics and communication with a central computer. These coupled with the developments in micro-electronics and power devices have led to a firm trend towards digital control of drives. There is a wide variety of applications such as machine tools, elevators; mill drives etc., where quick control over the torque of the motor is essential. Such applications are dominated by DC drives and cannot be satisfactorily operated by an induction motor drive with constant Volt-Hertz (V/f) scheme. Over the last two decades the principle of vector control of AC machines has evolved, by means of which AC motors and induction motors in particular, can be controlled to give dynamic performance comparable to what is achievable in a separately excited DC drive. In recent decades, many investigations have been done by researchers to control ac motors similar to that of separately-excited dc machines that lead them to vector control theory [1]. Vector control made the ac drives equivalent to dc drives in the independent control of flux and torque. The major disadvantage of the indirect vector control scheme is that it is machine parameter dependant, since the model of the motor is used for flux estimation. The machine parameters are affected by variations in the temperature and the saturation levels of the machine. Any mismatch between the parameters in the motor and that instrumented in the vector controller will result in the deterioration of performance in terms of steady state error and transient oscillations of rotor flux and torque. These types of oscillations are not desired for some exact uses. Regarding the importance of sensitivity of vector control drive to the motor parameters, many investigations have been carried out in this field. In [2] the effects of rotor resistance and mutual inductance variations on output torque and rotor flux have been discussed qualitatively. In the other work the effect of the machine parameters variations on its outputs, referring to simulation results has been investigated and two techniques for rotor resistance estimation have been described [3]. Krishnan in [4] has derived approximate equations for parameter sensitivity of indirect vector control; and finally in some references, motor parameter estimation and compensation techniques and their effect on machine outputs have been described [5]-[9]. In this paper exact equations of parameter sensitivity have been derived. Using derived equations, the effect of parameter variations on the outputs of the machine can be determined. In the next sections, first the basic equations of indirect vector control are presented, and then sensitivity analysis of this type of control carried out and analytical functions for output torque and rotor flux sensitivity are derived.

Indirect vector control rotor field oriented induction machine

The analysis of vector control structures highlights several possibilities to control the instantaneous values of the electromagnetic torque of the induction machine. The position of rotor flux can be estimated directly or indirectly. The

indirect estimation schemes of the rotor flux position ensure a good behaviour in all speed range, and they are the common solution in practice. In the indirect method [4], it is necessary to determine the rotor time constant in order to realize effectively the orientation of the rotor flux space vector. This can be seen from the following equations [7]:

$$\tau_r \frac{di_{mr}}{dt} + [1 - j\omega_m \tau_r] i_{mr} = i_s \quad (1)$$

Where

$\tau_r = \frac{L_r}{R_r}$ is the rotor time constant

L_r, R_r is the rotor inductance and resistance respectively

i_{mr} Rotor magnetization current space

i_s - Stator current space vector

ω_m - Electric rotor speed.

Equation (1) is determined taking into consideration the coordinates of the stator, transforming it to the rotor flux coordinates by multiplying with the unit vector $e^{-j\rho}$ and separating the result into real and imaginary parts, we get

$$\tau_r \frac{di_{mr}}{dt} + i_{mr} = i_{ds} \quad (2)$$

$$\frac{d\rho}{dt} = \omega_{mr} = \omega_m + \frac{i_{qs}}{\tau_r i_{mr}} \quad (3)$$

Figure 1 shows the current phasor diagram in the rotatory reference frame. The synchronous speed of the stator current space vectors is:

$$\omega_1 = \omega_{mr} + \frac{d\delta}{dt} \quad (4)$$

Where δ is the torque angle

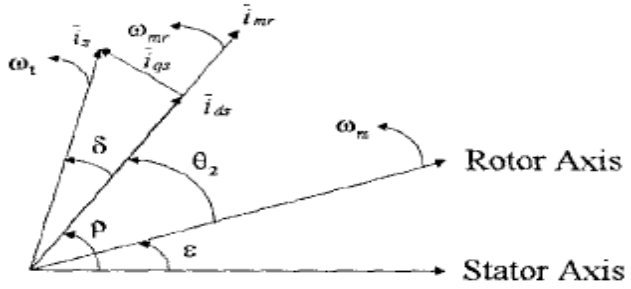


Figure 1 Phasor diagram representation

In steady state, the torque angle is constant i.e. $\omega_1 = \omega_{mr}$ and the vectors i_s and i_{mr}, rotate in synchronism. The components i_{qs} and i_{ds} can be controlled independently thus uncoupling the magnetic rotor flux from the torque, a similar behaviour as a separately excited dc motor. This is the basic principle of field orientation control (FOC). The indirect vector speed control with rotor time constant (z,) estimation is shown in figure 2 [6]. In the direct method [5], a rotor flux observer (rotor flux model) is needed in order to implement closed loop flux control such that a reference flux value is set and the estimated flux is the feedback signal.

The basic equations describing the dynamic behaviour of an indution machine in a rotating reference frame aligned to the rotor flux axis [1, 2] can be given as

$$T_r(di_{mr})/dt + i_{mr} = i_{sd} \quad (5)$$

$$d(\rho)/dt = \omega_1 = \omega_1 + \omega_2 = \omega_r + i_{sq}/T_r i_{mr} \quad (6)$$

$$Jd(\omega_r)/dt = 2/3P[L_m/(1+\sigma) i_{mr} i_{sq} - m_{load}] \quad (7)$$

As evident in the above equations, the electromagnetic torque is directly proportional to i_{sd} component of the stator current, if i_{mr} is kept constant. The ideal method of vector control implementation is to control the stator current by controlling i_{sq} and i_{sd} components separately. While a current controlled inverter using hysteresis controller is easy to realise, it has the $K_{te} = \frac{3}{2}\frac{p}{2}\frac{L_m}{L_r}$ following disadvantages:

1. The switching frequency depends on the nature of the load.
2. The current ripple is high.
3. Performance at higher speeds is unsatisfactory.

These disadvantages can be overcome by using a constant switching frequency Pulse Width Modulated (PWM) inverter to control the stator current by a voltage source inverter. This calls for the translation of stator voltage equations to the rotor flux reference frame.

The possibility to use the squirrel cage motor in high performance control systems in which a dynamic response similar to that given by dc motor opens up new applications in industry for this type of motors.

There exist two vector control methods:

- The direct field orientation method where field sensors or models are used to calculate the magnitude and position of the rotor flux vector subsequently orienting it in a system of rotatory orthogonal coordinates.
- The indirect method in which the slip angular speed is used to obtain the position of the rotor flux vector henceforth orienting it.

In both methods, it is necessary to determine correctly the orientation of the rotor flux vector, lack of which leads to degradation in the speed control of the motor. In the indirect method, the rotor flux obtained from an adaptable reference model was compared to a rotor flux obtained from a fixed reference model thus estimating the rotor time constant. This new value is substituted in the oriented field equations thus

not disorienting the magnetic field framework that rotates at the same speed as the rotor magnetic flux.

The indirect vector controller is derived from the dynamic equations of the induction machine in the synchronously rotating reference frames. The rotor equations of the induction machine are given by:

$$R_r i_{qr}^e + p\lambda_{qr}^e + \omega_{sl}\lambda_{dr}^e = 0 \quad (8)$$

$$R_r i_{dr}^e + p\lambda_{dr}^e - \omega_{sl}\lambda_{qr}^e = 0 \quad (9)$$

Where

$$\omega_{sl} = \omega_s - \omega_r \quad (10)$$

$$\lambda_{qr}^e = L_m i_{qs}^e + L_r i_{qr}^e \quad (11)$$

$$\lambda_{dr}^e = L_m i_{ds}^e + L_r i_{dr}^e \quad (12)$$

In these equations, the various symbols denote the following:

Rr , the referred rotor resistance per phase;

Lm , the mutual inductance per phase;

Lr , the stator referred rotor self inductance per phase; *e*

i_{dr}^e and i_{qr}^e the referred direct and quadrature axes currents, respectively;

p, the differential operator;

ω_{sl} , slip speed in rad/sec,

ω_s and ω_r are synchronous speed and electrical rotor speed both in rad/sec, and

λ_{dr}^e and λ_{qr}^e are rotor direct and quadrature axes flux linkages ,respectively.

The resultant rotor flux is assumed to be on the direct axis, to reduce the number of variables in the equations. Hence, aligning the d axis with rotor flux phasor yields:

$$\lambda_r = \lambda_{dr}^e \quad (13)$$

$$\lambda_{qr}^e = 0 \quad (14)$$

$$p\lambda_{qr}^e = 0 \quad (15)$$

Substituting equations (6) to (8) in (1) and (2) and using equations (4) and (5), the followings are obtained for i_f and ω_{sl} :

$$i_f = \frac{1}{L_m}[1 + pT_r]\lambda_r \quad (16)$$

$$\omega_{sl} = \frac{L_m}{T_r}\frac{i_T}{\lambda_r} \quad (17)$$

Where

$$i_f = i_{ds}^e$$

$$i_T = i_{qs}^e$$

$$T_r = \frac{L_r}{R_r}$$

The q and d axis currents are labeled as torque (i_T) and flux ($i\,f$) producing components of the stator current phasor, respectively. *Tr* denotes the rotor time constant. Also using equations (6) to (8), we can summarize the induction machine torque equation as:

$$T_e = \frac{3}{2}\frac{p}{2}\frac{L_m}{L_r}(\lambda_r i_{qs}) = \frac{3}{2}\frac{p}{2}\frac{L_m}{L_r}\lambda_r i_T = K_{te}\lambda_r i_T \quad (18)$$

Where K_{te} is torque constant and is equal to:

$$K_{te} = \frac{3}{2}\frac{p}{2}\frac{L_m}{L_r} \quad (19)$$

Note that the torque is proportional to the product of the rotor flux linkages and the stator q axis current. This resembles the torque expression of dc motor, which is proportional to the product of the field flux linkages and the armature current. If the rotor flux linkage is kept constant, then the torque is simply proportional to the torque producing component of the stator current (*T i*), as in the case of the separately excited dc machine, where the torque is proportional to the armature current when the field current is constant. The rotor flux linkage and torque equations given in (9) and (14), respectively, complete the transformation of the induction machine into an equivalent separately excited dc machine from a control point of view.

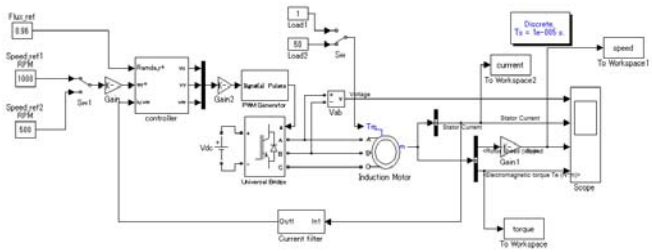


Figure 2 Simulation model of indirect vector controlled induction motor drive system simulation and experimentation

Computer modelling and simulation is widely used to study the behaviour of various complex systems. With proper simulation techniques, a significant amount of experimental cost could be saved in the prototype development. Among several simulation software packages, SIMULINK [7] is one of the most powerful techniques for simulating dynamic systems due to it's graphical interface and hierarchical structure and in addition SIMULINK uses MATLAB as a Tool for mathematical purposes which further enhances the modelling process. This software permits the design of special user blocks, which can be added to the main library. Figure 3 shows the block diagram in the d-q synchronous reference frame of the reluctance synchronous machine without damper windings, by considering a two pole machine. The vector control aspects are studied using synchronously rotating

reference frame. The main part of the dynamic vector controlled reluctance synchronous motor model is the two-axis motor model block which consists of electrical torque, machine voltages and mechanical equations. The (3-2) phase transformation block converts the 3 phase supply to 2-phase but the reverse transformation block converts the 2-phase rotating reference frame into 3-phase stationary equivalent The load torque T is simply represented as a constant value, though it can be represented as a variable quantity. In older to start siniulating the system, the parameters must be known They can be either calculated or measured as in this investigation.The parameters of the motor, used for simulation are as shown in Table.

Table I induction machine parameters.

S.no.	Parameters	Value
1	Voltage(V)	220
2	Frequency (Hz)	60
3	Inertia	0.089

Figure 2 Motor specifications.

Table II: Speed Controller

S.no	Parameters	Value
1	Proportional gain (Kp)	0.5
2	Integral gain (Ki)	30
3	Derivative gain (Kd)	0

Figure 3 Controller parameters.

Results and Discussions

Vector controlled motor (kW) has been simulated and the dynamic responses have been obtained. . The simulation results for speed-time, torque-time and current-time are as shown in figures 3-5 respectively. Figure 3 shows the speed response for the motor with different load torque *(T,)* when the reference speed is 200 (rad sec) while keeping the voltage-frequency ratio (v/f) constant. Figure 6 shows the starting torque produced by the motor when the motor started from rest, and figure 5 shows the speed /torque characteristics of the drive.

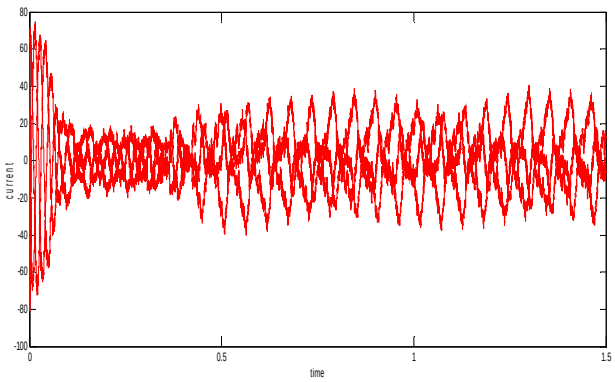


Figure 5 Current Responses with Time Variation

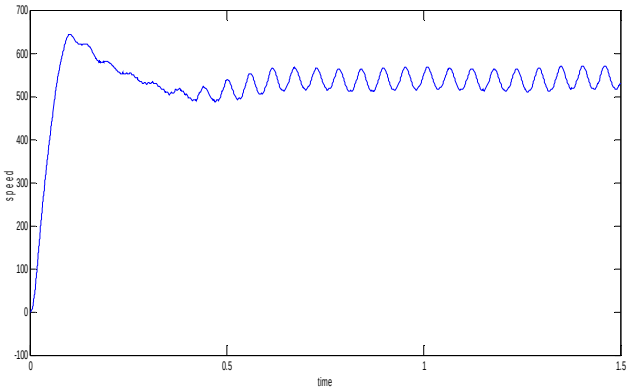


Figure 6 Speed Response with Time Variation

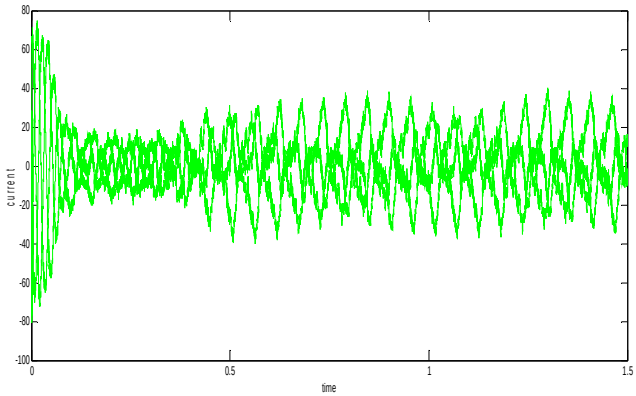


Figure 7 Torque Variation with time.

To validate the simulation results, the prototype of the proposed scheme is worked upon in the laboratory and tests are conducted on it. The motor is run at speed of 1425 rpm in both simulations as well as in experimentation. It can be observed that the simulation results match reasonably with the experimental results. It has been observed that the design of proposed method is flexible for making it suitable for retrofit applications.

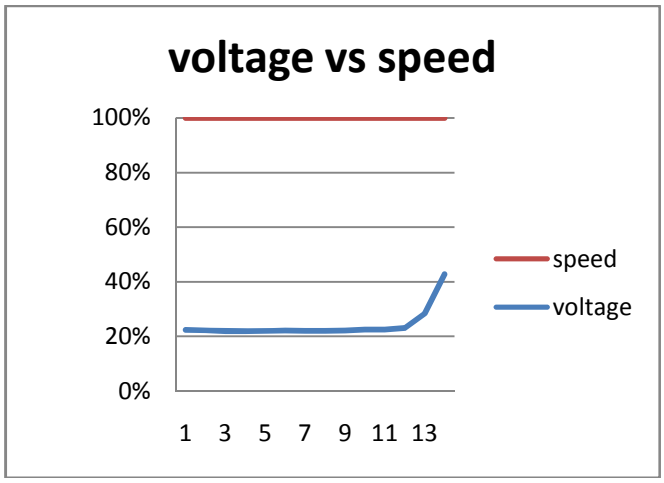


Figure 8 Experimental Speed and voltage response

The experimental set up is shown as under:

Conclusions

A design procedure has been given for various control loops of a vector controlled drive. The design procedure is verified using computer simulation. The simulation results helps to decide the hardware and software structure for the vector controlled induction motor drive. The total system includes an electronic controller for the signal processing involved in vector control, PWM inverter, motor and the load. Although the hardware and software have been dealt with in the context of a vector controlled induction motor drive, they are general purpose in nature. These can be used for control of other machine types such as wound field as well as permanent magnet synchronous motors. It is also possible to use the same hardware and software for the implementation of other control schemes like direct self- control of induction machines.

The operation of 3.75 KW drive was also investigated experimentally. The simulation has been validated using the actual control of induction motor. The simulation results have been shown a reasonably close agreement with experimental results. The results demonstrate that the drive can be operated in a wide speed and load range.

References

[1] Blaschke F., “The Principle of Field Orientation as Applied to the New Transvektor Closed-Loop Control System for Rotating -Field Machines”. Siemens Rev., 39, 5, May 1972, pp. 2 17-220.

[2] Blaschke F., “Das Verfahren der Feldorientierung zur Regelung der Drehfeldmaschine” Techn. Univ. Braunschweig, Dissertation, 1973.

[3] Hasse K., “Zur Dynamik drehzahlgeregelter Antriebe mit stromrichtergespeisten Asynchron Kurzschlusslaufermaschinen”.Techn. Hochsch. Darmstadt, Dissertation, 1969.

[4] Pedro Ponce & Jaime Rodriguez Rivas., “Simulation del Control por Campo Orientado de Motores de Induccion”. IEEE RVP 97.

[5] Bimal K. Bose “Power Electronics and Variable Frequency Drives”, IEEE Press

[6] Pedro Ponce & Jaime Rodriguez Rivas., “Estimacion de la Constante de Tiempo del Rotor del Motor de Induccion Mediante un Modelo de Flujo Adaptable del Sistema”. IEEE RVP 98, pp. 373-379, Tomo I.

[7] Murphy J.M.D. & Turnbull F.G. “Power Electronic Control of A.C. Motors”, Pergamon also estimates the rotor flux satisfactorily. Press, 1988.

[8] Ben-Brahim L., Gastli A., Al-Hamadi M., *Auto-Tuning for Sensorless AC Motor Drive Systems.* IEEE Int. Symp. on Ind. Electronics, Bled, Slovenia, 367-372, 1999.

[9] Onea A, Horga V., *Educational Software Environment for Motion Control.* Proc. of the IASTED 5th Int. Conf. on Comp. and Adv. Techn. in Education, 334-339, 2002.

[10] Bhim Singh, “A Novel Polygon Based 18-pulse AC-DC Converter for Vector Controlled Induction Motor Drive”. IEEE transactions vol.22, no.2, March 2007.

[11] J. C. Basilio, and S. R. Matos, "Design of PI and PID Controllers with Transient Performance Specification," *IEEE. Trans. Education,* vol. 45, No.4, pp. 364-370, November 2002.

[12] M. M. M. Negm, J. M. Bakhashwain, and M. H. Shwehdi, "Speed Control of a Three-Phase Induction Motor Based on Robust Optimal Preview Control Theory," *IEEE Trans. On Energy Conversion,* vol. 21, No. I, pp. 77-84, March 2006.

[13] M. Bounadja, A. Mellakh i, and B. Belmadani, "A High Performance PWM Inverter Voltage-Fed Induction Machines Drive with an Alternative Strategy for Speed Control," *Serbian journal of electrical engineering,* vol. 4, No. I, pp. 35-49, June 2007.

[14] L. A. Zadeh,"Outline of a New Approach to the Analysis of Complex Systems and Decision Processes", *IEEE Trans. Systems* , Man, and Cybernetics, No.3, pp. 28-44.

[15] B. Singh, G. Bhuwaneswari and V. Garg, "Multipulse improved-power-quality AC-DC converters for vector controlled induction motor drives" , IEE Proc. –Electr Power Appl., vol. 153, no. 1, pp. 88-96, Jan. 2006.

[16] T. Iwakane et al., "high performance vector controlled AC motor drives”, vol.IA 23, no.5, September 1987.

[17] J. A. Santisteban, and R. M. Stephan, "Vector control methods for induction machines: an overview," *IEEE Trans. On Education,* vol. 44, No.2, pp. 170-174,2001.

[18] V. Garg, “Power Quality Improvements at ac Mains in Variable Frequency Induction Motor Drives,” Ph.D. dissertation, Indian Inst. Technol. Delhi, New Delshi, 2006.

Magic Number in Neutron-Rich Nuclei using Relativistic Mean Field Formulism

M.S. Mehta[†], Poonam Malik[‡] and K.S. Upadhyaya[†]

[†]Department of Applied Sciences, RBCEBTW, Mohali- 140 104, India
[‡]Kurukshetra Institute of Technology and Management, Bhor Saidan, Kurukshetra-136 119, India

Abstract
The magic numbers associated with closed shells have long been assumed to be valid across the whole nuclear chart. Investigations in recent years of nuclei in the drip-lines have revealed that the magic numbers may change locally in the exotic nuclei leading to the disappearance of shell gaps and the appearance of new magic numbers. We investigate the magic number in neutron drip-line using axially deformed Relativistic Mean field Theory. The neutron numbers N=28, in ^{52}Ca and N=40 in ^{52}Ca become magic or semi magic, while its magicity is broken in ^{60}Ni. A considerable shell gap at N=40 appears in ^{68,78}Ni and ^{60}Cabut almost disappears in ^{86}Ni.

Index Terms: Nuclei, magic number, binding energy, drip-line, single particle energy, quadrupole deformation, separation energy.

Introduction

The magic number is the backbone of nuclear structure physics that explains the structure of nuclei close to stability. The origin of magic number is in the gaps created by the single-particle eigen states of the mean-field. Exploring the formation of shell structure away from the β-stability line is a frontier in the nuclear physics. With advancement in Radioactive Ion Beam Facilities (RIBF) it has started to change the nuclear landscape and have offered the glimpses of exotic new phenomena away from the β-stability line. The new generation of RIBF will attempt give the possible answer to one of the questions that what is the shell configuration of nuclei at the extremes of iso-spin. The answers to the questions have started emerging, for example, exotic nuclei ^{42}Si [1] and ^{78}Ni [2] have been produced at the National Superconducting Cyclotron of Michigan State University. The first evidence for the N = 16 magic number in oxygen came from an evaluation of neutron separation energies on the basis of measured masses [7]. The measurements showed some surprising changes in the nuclear shell structure as a function of proton and neutron number in light nuclei. These observations triggered numerous theoretical investigations, which in turn made new predictions that some magic numbers will disappear and new shell gaps will appear in certain regions of the nuclear chart [3]. The disappearance of the magic numbers N = 8 and 20 in the light nuclei, and/or appearance of the new magic numbers N = 16 and 32 in drip-line nuclei have been reported recently and can be found in Refs. [4-7]. One case of particular interest is the magicity at proton/ neutron number N = Z = 28 and N = 32 and 40 near the neutron drip-line in Ca and Ni-isotopes, which has been a centre of discussion for sometimes now (see Ref.[8], and the references therein). Recently, N=40 is reported to be magic in HF calculations with a number of effective interactions [9]. In the present investigation we calculate the single particle energy levels, quadrupole deformation parameter $\beta_2$and separation energy using axially deformed relativistic mean field model. A brief description of model and the analysis of the result are presented in the following sections.

Lagrangian Density

The RMF model has been proved to be a very powerful tool to explain the properties of finite nuclei and infinite nuclear matter [10], [11], [12] for the last two decades. We start with the relativistic Lagrangian density for a nucleon-meson many-body system,

$$L = \overline{\psi_i}\{i\gamma^\mu\partial_\mu - M\}\psi_i + \frac{1}{2}\partial^\mu\sigma\partial_\mu\sigma - \frac{1}{2}m_\sigma^2\sigma^2 - \frac{1}{3}g_2\sigma^3 - \frac{1}{4}g_3\sigma^4 - g_s\overline{\psi_i}\psi_i\sigma - \frac{1}{4}\Omega^{\mu\nu}\Omega_{\mu\nu} + \frac{1}{2}m_\omega^2V^\mu V_\mu + \frac{1}{4}c_3\left(V_\mu V^\mu\right)^2 - g_\omega\overline{\psi_i}\gamma^\mu\psi_iV_\mu - \frac{1}{4}\vec{B}^{\mu\nu}.\vec{B}_{\mu\nu} + \frac{1}{2}m_\rho^2\vec{R}^\mu.\vec{R}_\mu - g_\rho\overline{\psi_i}\gamma^\mu\vec{\tau}\psi_i\vec{R}^\mu - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - e\overline{\psi_i}\gamma^\mu\frac{(1-\tau_{3i})}{2}\psi_iA_\mu \quad (1)$$

The field for the σ-meson is denoted by σ, that for the ω-meson by V_μ and for the iso-vector ρ-meson by $\vec{R}_\mu$. A_μ denotes the electromagnetic field. The ψ_i are the Dirac spinors for the nucleons whose third component of isospin is denoted by τ_{3i}. Here g_s, g_ω, g_ρ and $\frac{e^2}{4\pi} = \frac{1}{137}$ are the coupling constants for σ, ω, ρ mesons and photon, respectively. g_2, g_3 and c_3 are the parameters for the nonlinear terms of σ- and ω-mesons. M is the mass of the nucleon and m_σ, m_ω and m_ρ are the masses of the σ, ω and ρ-mesons, respectively. $\Omega^{\mu\nu}$, $\vec{B}^{\mu\nu}$ and $F^{\mu\nu}$are the field tensor for the V^μ,$\vec{R}^\mu$and the photon fields, respectively[13].

From the relativistic Lagrangian we get the field equations for the nucleons and mesons. These equations are solved by expanding the upper and lower components of Dirac spinors and the Boson fields in a deformed harmonic oscillator basis with an initial deformation. The set of coupled equations is solved numerically by a self-consistent iteration method. The centre-of-mass motion is estimated by the usual harmonic oscillator formula$E_{c.m.} = \frac{3}{4}\left(41A^{-1/3}\right)$ The quadrupole deformation parameter β_2 is evaluated from the resulting

quadrupole moment [13] using the formula,

$$Q = Q_n + Q_p = \sqrt{\frac{9}{5\pi}} AR^2 \beta_2, \quad (2)$$

Where, R = $1.2A^{1/3}$. The total binding energy of the system is,

$$E_{Total} = E_{Part} + E_\sigma + E_\omega + E_\rho + E_c + E_{pair} + E_{c.m} \quad (3)$$

where E_{part} is the sum of the single-particle energies of the nucleons and E_σ, E_ω, E_ρ, E_c and E_{pair} are the contributions of the mesons fields, the Coulomb field and the pairing energy, respectively. For the open shell nuclei, the effect of pairing interactions is added in the BCS formalism. For pairing strength, the BCS approach provides a reasonably good description of nuclei close to o not too far away from stability line. For the nuclei in the vicinity of drip-lines, coupling to continuum becomes important. However, it has been shown that self-consistent treatment of BCS approximation breaks down when coupling between bound states and the states in continuum takes place [14]. The pairing gaps for proton Δ_p and neutron Δ_n are calculated from the relations [15],

$$\Delta_p = r b_s Z^{-\frac{1}{3}} e^{(sI - tI^2)}$$
$$\Delta_n = r b_s N^{-\frac{1}{3}} e^{-(sI + tI^2)} \quad (4)$$

where r = 5.72MeV , s = 0.118, t = 8.12, b_s = 1 and I = (N −Z)/(N +Z).

Results and Discussion

In the present calculations we use the axially deformed relativistic mean field (RMF) Theory with NL3 parameter set. The pairing interaction is taken care within the BCS scheme, with gap parameters as in Ref. [16].

The results of calculations for the neutron single particle energy levels for ^{60,68,78}Ni and ^{52,60}Ca nuclei are, respectively, shown in Figs. 1 and 2. In figure 1, the shell gap of ~ 4.0 MeV can be seen at N = 28 and N = 40 for all the nuclei considered here. Although, the gaps are not very large but fairly clear at these numbers. The two neutron separation energy is shown in Fig. 3. The magicity corresponding to N= 28 and 40 can easily be found at sudden fall in the separation energies of $^{52-80}$Ni and $^{40-76}$Ca nuclei. In case of Ni, the sharp decrease in separation energy at N = 50 is obtained, which means the magicity at N = 50 is also indicated in RMF calculations. The sharp decrease in the separation energy at N = 28 and 40 is in agreement with the single particle energy levels in Fig. 1, which further gives the insight of the magicity. Considerably large shell gaps (≈ 6 MeV), both at N = 28 and 40 in ^{68}Ni nucleus appears. Note that this nucleus has been suggested to be a doubly magic nucleus experimentally [17]. Similarly, in^{78}Ni nucleus, the large shell gaps at N = 28, 40 and 50are obtained. This shows that the magicity at N = 28 is perhaps not diminished, and the gap at N = 50 is nearly~4.0 MeV. Fig. 5, showing a plot of neutron matter radius for $^{40-76}$Ca isotopes, in which a sharp change in radius at N= 28 and 40. The considerably large difference in the radii at N =28 and 40 gives the indication of the extra stability of nuclei at these numbers. In Fig. 4, we present the variation of the pairing gap (Δ_n) for neutron for the different paring strengths. The pairing gap vanishes at N = 28 and 40 at small pairing strength of G_n = G_p = 0.1 showing the magicity at these numbers.

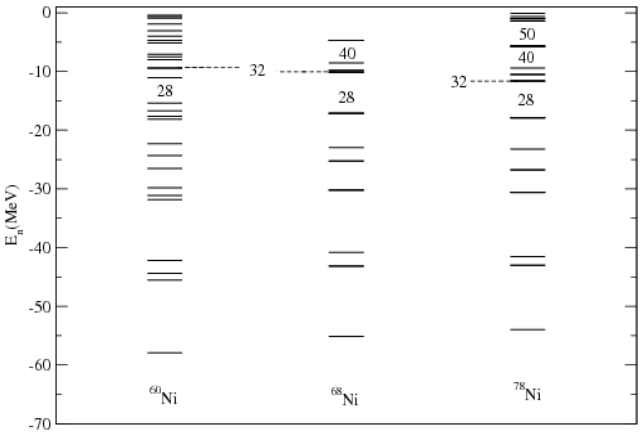


Fig. 1. The neutron single particle energy levels of ^{60,68,78} Ni nuclei with NL3 parameter set.

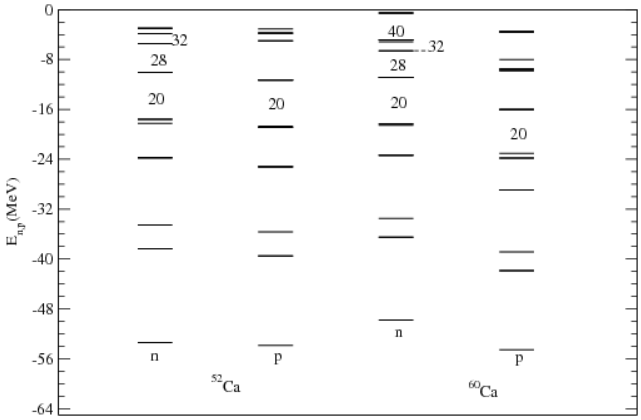


Fig. 2. The neutron and proton single particle energy levels of ^{52,60} Ca nuclei with NL3 parameter set.

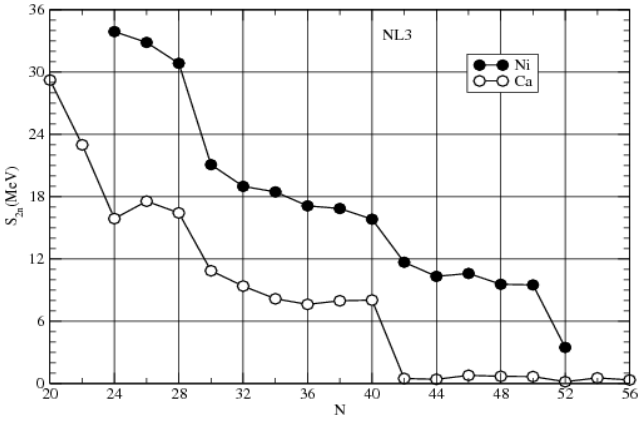


Fig. 3. Two neutron separation energy of Ca and Ni isotopes with NL3 parameter set.

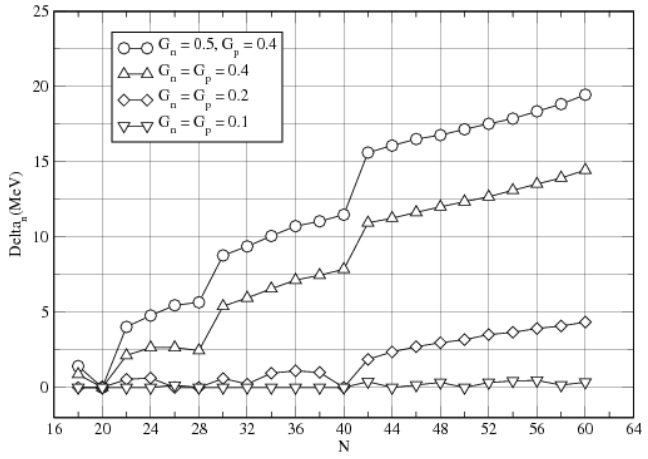


Fig. 4. The variation of pairing gap of $^{38-80}$ Ca isotopes for different strength with NL3 parameter set.

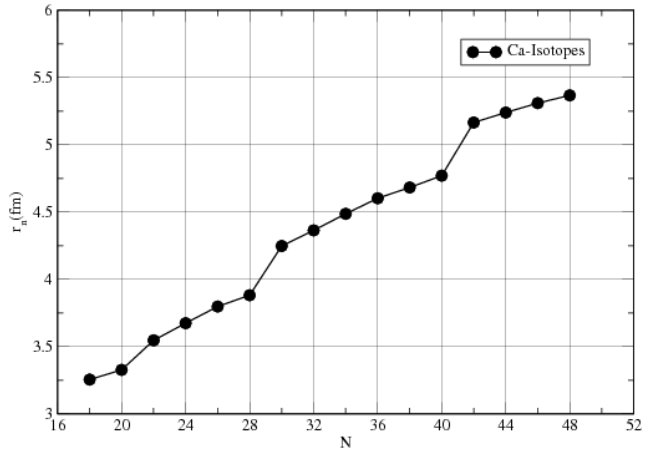


Fig. 5. The neutron radius of $^{38-68}$ Ca nuclei as a function of neutron number using NL3 parameter set.

Conclusion

We have investigated shell structure of the neutron-rich Ca and Ni nuclei using axially deformed Relativistic Mean field Model with NL3 parameter set. At N = 32 a small gap in single particle energy levels is obtained which cannot account for the magic nature, but shell-model calculations using the well established KB3G interaction [18],[19] support a N = 32 shell closure, which is experimentally well established, but not a N = 34 shell closure, whereas the gaps at N = 28 and 40 are considerably large in both the cases considered here. The same is seen in the two neutron separation energy also. The present calculations do not support the N = 34 magic number. Therefore, in RMF calculations with NL3 parameter set show the magicity at N=28, 40 and 50 for the isotopes of Ca and Ni.

References

[1] J. Fridmann et al., Nature, 435, 922 (2005)
[2] P. T. Hosmer et al., Phys. Rev. Lett. 94, 112501 (2005)
[3] R. Kru¨cken, arXiv:1006.2520v1 [phys.pop-ph], 2010.
[4] A. Gade, Th. Glasmacher, Prog.Part. Nucl. Phys. 60, 161(2008); R.V.F. Janssens, Nature 459, 1069 (2009).
[5] R. K. Gupta, M. Balasubramaniam, S. Kumar, S. K. Patra, G.M¨unzenberg and W. Greiner, J. Phys. G.: Nucl. Part. Phys.32, 565 (2006); R. K. Gupta, S. K. Patra and W. Greiner, Mod. Phys. Lett.A12, 1317 (1997).
[6] T. K. Jha, M. S. Mehta, S. K. Patra and R. K. Gupta, PramanaJ. Phys., 61, 517 (2004).
[7] A. Ozawa, et al., Phys. Rev. Lett. 84 5493 (2000); R. Kanungo, I. Tanihata and A. Ozawa, Phys. Lett. B 528, 58(2002).
[8] O. Sorlin and M.-G. Porquet, Prog. Part. Nucl. Phys. 61, 602(2008); O. Sorlin, et al., Phys. Rev. Lett. 88, 092501 (2002).
[9] H. Nakada, arXiv:1003.5720v2 [nucl-th] (2010).
[10] R. Machleidt, Adv. Nucl. Phys. 19, 189 (1989).
[11] S.K. Patra and C.R. Praharaj, Phys. Rev. C44, 2552 (1991).
[12] M. Del Estal, M. Centelles, X. Vi˜nas and S.K. Patra, Phys. Rev.C63, 024314 (2001).
[13] Y.K. Gambhir, P. Ring and A. Thimet, Ann. Phys. 198, 132(1990).
[14] J. Dobaczewski, H. Flocard and J. Treiner, Nucle. Phys. **A422**, 103 (1984); J. Dobaczewski, W. Nazarewicz, T. R. Werner, J. F. Berger, C.R. Chinn and J. Decharge, Phys. Rev. **C53**, 2809 (1996).
[15] D.G. Madland and J.R. Nix, Nucl. Phys. A476, 1 (1988);P. M¨oller and J.R. Nix, At. Data Nucl. Data Tables 39, 213(1988).
[16] D.G. Madland and J.R. Nix, Nucl. Phys. A476, 1 (1988).
[17] R. Broda, et al., Phys. Rev. Lett. 74, 868 (1995).
[18] E. K. Warburton, J. A. Becker, and B. A Brown, Phys. Rev. C41, 1147 (1995).
[19] A. Poves et al., Nucl. Phys. A 694, 157 (2001); E. Caurier etal., Eur. Phys. J. A 15, 145 (2002); A. Poves, F. Nowacki, E.Caurier, Phys. Rev. C 72, 047302 (2005).

Block-Cipher Design with Effective Key Generation Technique Involving the Use of Multiplication Factor in Addition to a Key

S.G. Srikantaswamy[1] and Prof. H.D. Phaneendra[2]

[1]S.G. Srikantaswamy, Research Scholar, National Institute of Engineering, Mysore, Karnataka, India
[2]H.D. Phaneendra, Professor and Research Guide,
Dept. of C.S., National Institute of Engineering, Mysore, Karnataka, India

Abstract
Securing data that is being transmitted from sender to receiver from eavesdropper is a very important aspect of secured communication. Design and development of effective cryptosystems have a great significance in the field of information security. Cryptography deals with the design and development of encryption and decryption algorithms also known as ciphers. Ciphers play a key role to facilitate secret communication between two entities through an existing channel. In cryptography, the original message is called plaintext. The process of scrambling the plaintext message is called encryption and the scrambled message is called ciphertext. The confusion and diffusion properties play a vital role in the design of the cipher. All cryptosystems are based on the key value shared between sender and receiver. The present paper deals with the design and development of a cipher which involves matrix addition, Exclusive- or operation and thus incorporates both arithmetic and logical operations. By combining the effects of arithmetic and logical operations we have tried to make the effect of confusion and diffusion characteristics of the cipher more complex and thus making the system more strong.

Keywords: Cipher, plaintext, cipher text, encryption, Decryption, eavesdropper.

Introduction

Secured communication is one which is characterized by the feature that the sender can able to send message to receiver by preventing the eavesdropper from getting the message contents[4]. Cryptography helps us to keep the message secure from attackers. Encryption is performed on plaintext at the sending end and Decryption is performed on ciphertext at the receiving end. The basic building blocks of all encryption and decryption techniques are substitution and Transposition.[4, 5, 6].A substitution technique is one in which the plaintext characters of the original message are replaced by other characters. Many substitution ciphers are available practically [4]. Hill cipher is a multi-letter cipher developed by Lester Hill in 1929 [4].The Hill cipher takes m successive plaintext characters and substitutes for them m ciphertext characters. Hill cipher is described as follows : C= KP mod 26, where C and P are column vectors representing the plaintext and ciphertext and K is encryption key represented in matrix form. The Hill cipher has been modified by involving Interweaving and Iteration techniques[1]. Another variation to Hill cipher involves the multiplication of plaintext matrix with key matrix and confusion effect has been created by using EX-OR operations between plaintext and key matrices [2]. One more block cipher design shows the use of key on one side and key inverse on other side.[3]. In this paper, we have proposed a cipher which involves the combined effect of arithmetic and logical operations. As Hill cipher, the present cipher treats the plaintext and key values in the form of matrices. The effort of this cipher is to make cryptanalysis more difficult and making the cipher more strong.

In this paper , the ASCII (Decimal) equivalent values of the plaintext characters are considered for analysis purpose. The system is designed such that, it accepts n plaintext characters at a time and produces n ciphertext characters. The plaintext characters are treated as square matrix consisting of n characters. The key k is selected such that, it also contains n characters. The algorithm is different from the other algorithms such that the key value is multiplied by a multiplication factor m before using it to convert the plaintext to ciphertext. m is selected such that preferably it is a prime number. The decryption requires both key and multiplication factor. The drawback is that it is necessary to protect both key and the multiplication factor . The K and m should be distributed to receiver through a secured channel. The actual key value (K1) is obtained by multiplying the key matrix k by multiplication factor m.[K1=K*m]. The actual key value K1 is exclusive –ored bitwise with the plaintext to obtain partial ciphertext c1.The value of c1is added with the key value K1 to produce final Ciphertext C. The decryption is performed exactly in the reverse direction as that of the encryption process to get back the plaintext from the ciphertext at the receiving end. The organization of the paper is as follows. Section 2 provides encryption and decryption process in detail. Section 3 gives the encryption algorithm. Section 4 gives the decryption algorithm. Section 5 provides the encryption results. Section 6 provides decryption results. Section 7 provides the conclusion drawn from the above analysis. Section 8 provides references.

Encryption and Decryption Process Details

Consider the following original message (plaintext) which is to be encrypted and transmitted securely to the receiving end. "**This content is very important. Save this with a separate file name. Don't disclose this to any body. This matter**

should be…….". Our algorithm is designed such that, it takes n characters of plaintext characters at a time and converts it in to Ciphertext. Let us consider the following part of the plaintext message which contains 16 characters. That means 16 characters are processed at a time. It need not be 16, but for calculations purpose , 16 characters are considered here. " **This content is ver**". Now this 16 plaintext characters are treated as 4X4 square matrix. Now each characters of the plaintext are represented by their ASCII Equivalent(decimal) values, for the purpose of analysis. Select a key value consisting of 16 decimal characters. Also select a multiplication factor m which should be preferably a prime number to thwart the attempts to find m. The key value selected is not used directly with the plaintext. The actual key value K1 is calculated by multiplying the value of K with m. The actual key value K1 is the product of K and m. The value K1 is exclusive-ored with the plaintext value bitwise. The resultant value is considered as the partial ciphertext. The final ciphertext is obtained by adding the partial ciphertext matrix (C1) with actual key value matrix (K1). The Encryption process can be expressed as follows :

$C = [(m * K) \oplus P + (m * K)]$.

Note: ⊕ symbol stand for Exclusive-or operation.

The decryption process is a technique of obtaining plaintext from the ciphertext. The decryption process is performed exactly in the reverse direction as that of encryption.

The decryption process is expressed as follows:

$P = [C - (m * K)] \oplus [m * K]$.

Encryption Algorithm

Step 1: Select P, K and m;

P= plaintext message

K= Key value

m= Multiplication factor

Step 2: Calculate the actual key value K1

K1=K * m;

Step 3: Calculate Partial ciphertext C1

$C1 = P \oplus K1$

Step 4: Calculate the final Ciphertext C

C= C1+K1;

$C = [(m * K) \oplus P + (m * K)]$

Step 5: Transmit the ciphertext

Decryption Algorithm

Step 1: Select C, m, K

Step 2: Calculate K1

K1 = k * m

Step 3: Calculate plaintext P

$P = [C - (m*K)] \oplus [(m * K)]$

Encryption Process Results

Consider the plaintext message: "This **content is very important. Save this with a separate file name. Don't disclose this to any body. This matter should be…….".** The portion of the plaintext message considered for encryption at a time is "This **content is ver** ". The Algorithm takes 16 (128 bit) characters at a time and converts it in to plaintext. Select the ASCII decimal equivalent values for the plaintext characters. Select 16 bytes Key value. Also select integer m. Select P, K and m properly. Represent P and K as 4X4 Matrix form. To do this, Replace each plaintext character by ASCII equivalents (decimal).Then represent the plaintext P as a square matrix of order 4X4.

Replacing each character of the above plaintext message by their ASCII decimal equivalents, we get the following set of values.

P =	84	104	105	115	99	111
	110	116	101	110	116	105
	115	118	101	114		

Representing the above set of values in the form of 4 X 4 Matrix, we get P as follows:

$$P = \begin{pmatrix} 84 & 104 & 105 & 115 \\ 99 & 111 & 110 & 116 \\ 101 & 110 & 116 & 105 \\ 115 & 118 & 101 & 114 \end{pmatrix}$$

Select a key value K consisting of 16 decimal characters.

K = { 5 10 121 85 62 17 53 50 41 32 16 08 53 99 86 35 }

Let us represent the above values of K in 4 X 4 Matrix The K in Matrix form appears as shown below.

$$K = \begin{pmatrix} 5 & 10 & 121 & 85 \\ 62 & 17 & 53 & 50 \\ 41 & 32 & 16 & 08 \\ 53 & 99 & 86 & 35 \end{pmatrix}$$

In summary, Encryption involves the following operations.

1. Calculate K1:

K1=m *K;

2. Calculate P1
P1=P ⊕ K1

3. Calculate final Ciphertext C
C=P1 +K1

Note: Symbol⊕ stands for Exclusive -or operation

The actual key value K1 is obtained by multiplying the key value K by a multiplication factor m. Let m=3 (a prime number).

Now K1= K * m.

$$K1 = 3 X \begin{pmatrix} 5 & 10 & 121 & 85 \\ 62 & 17 & 53 & 50 \\ 41 & 32 & 16 & 08 \\ 53 & 99 & 86 & 35 \end{pmatrix}$$

Now calculate P1
P1 = P ⊕ K1

$$P1= \begin{pmatrix} 84⊕15 & 104⊕30 & 05⊕363 & 115⊕255 \\ 99⊕186 & 111⊕51 & 110⊕159 & 116⊕150 \\ 101⊕123 & 110⊕96 & 116⊕48 & 105⊕24 \\ 115⊕159 & 118⊕297 & 101⊕258 & 114⊕105 \end{pmatrix}$$

$$P1= \begin{pmatrix} 91 & 118 & 258 & 140 \\ 217 & 92 & 241 & 226 \\ 30 & 14 & 68 & 113 \\ 236 & 351 & 359 & 27 \end{pmatrix}$$

Final Cipher text , C = P1+K1

$$C= \begin{pmatrix} 106 & 148 & 621 & 395 \\ 403 & 143 & 400 & 376 \\ 153 & 110 & 16 & 137 \\ 395 & 648 & 617 & 132 \end{pmatrix}$$

Thus the resultant cipher text = {106 148 621 395 403 143 400 376 153 110 116 137 395 648 617 132 }.

Decryption Process Results

P= [C-K1] ⊕ K1

The substituting the Values for C , K1 in the above Decryption Algorithm Equation We get Plaintext in terms of ASCII Decimal form as shown below.

$$P = \begin{pmatrix} 84 & 104 & 105 & 115 \\ 99 & 111 & 110 & 116 \\ 101 & 110 & 116 & 105 \\ 115 & 118 & 101 & 114 \end{pmatrix}$$

Now replace each ASCII Decimal Equivalent Values by corresponding character, we get the following characters.

84-T	104-h	105-i
115-s	99-c	111-o
110-n	116-t	101-e
110-n	116-t	105-i
115-s	118-v	101-e
114-r		

Thus, the received message is :"**This content is ver**". The Resultant plaintext message is " **This content is ver** " which is same as the transmitted message from the sender. Now the next portion of the Message is converted to plaintext in the similar manner as described above. The Algorithm has been developed and tested using C language.

Conclusion

Secured Communication is a very desirable aspect in network communication. The information that is being transmitted should be protected from both active and passive attacks. The proposed cipher provides more security to the plaintext message that is being transmitted through the network from eavesdropper. The cipher requires a secured distribution multiplication factor in addition to key value. The length of the key and multiplication factor plays a very important role in providing security. Thus selecting the values for key and multiplication factor is very important to get more security without reducing the encryption speed. The proposed cipher can be applied to messages of any length.

References

[1] *A Modified Hill Cipher Involving Interweaving and Iteration- v.Umakanta Sastry, N. Ravi shankar and S. Durga Bhavani, Director, SCSI, Dean (R&D), sreenidhi Institute of Science and Technology , Hyderabad , India, CSE Department , SNIST , Hyderabad ,India, SIT, JNT University , Hyderabad, India-International Journal of Network Security Vol.11, No.1,PP.11-16, july 2010.*

[2] *A Secure Variant of the Hill Cipher –Mohsen Toorani, Abolfazi Falahati, school of Electrical Engineering , Iran University of Science and Technology, Tehran , Iran – Proceedings of the 14th IEEE Symposium Computers and Communications(ISCC 09)*

[3] *A Block Cipher Having a Key on one side of the Plaintext Matrix and its Inverse on the other side.-Dr. V.U.K.Sastry , Prof.D.S.R. Murthy , Dr.S. Durga Bhavani – International Journal of Computer Theory and Engineering. Vol.2, No.5 , October-2010 1793-8201.*

[4] *Cryptography and Network Security Principles and Practices , Third edition- William Stallings*

[5] *Applied Cryptography Second edition, Bruce Schneier*

[6] *Introduction to Modern Cryptography Jonathan Katz , yehuda Lindell.*

Optimal Congestion with N+1 Label

[1]Ankur Dumka and [2]Prof. Hadwari Lal Mandoria

[1]Ph.D., Uttarakhand Technical University, Uttarakhand, India
[2]HoD, College of Technology G.B. Pant University, Pantnagar, Uttarakhand, India

Abstract
This article discusses the traffic engineering in MPLS in an ISP network. The feedback is send at hop by hop rather then end to end network. A header is attached by the reaceiving header of 1 bit which carry a value of 1, if congestion in the receiving router and its value is 0, if there is no cogestion. Thus, in the case of congestion sending router will send the packet with another longer path, thus preventing the congestion of packets.

Introduction
MPLS is an IETF standard that merges layer 2 and layer 3 protocol that uses label switching in the core network, thus reduces the workout of looking the routing table overhead.MPLS is a traffic switching technology combines the traffic engineering capability of ATM with flexiblity and scalability of IP. This enble fast packet transfer over the network. MPLS establishes a connection-oriented mechanism into a connectionless IP network. MPLS uses short, length-fixed, locally significant labels in the packet header between layer 2 and layer 3 header and the packets are forwarded according to label rather then by routing protocol in the core of network. Egress routers uses the routing table to convert IP packet into MPLS label and back to IP packet while leaving the network.MPLS technology offer services including layer 3 VPN(VPRN), traffic engineering, traffic protection and layer 2 VPN(VPLS).

Traffic Engineering is an advanced feature of MPLS which controls the traffic flows through one's network in order to optimize resource utilization and network performance[2]. The need to use traffic enginnering is because the Interior Gateway

Protocol(IGP) always uses shortest path to forward traffic.Now as all the packets follow the same shortest route then there exist the problem of underutilization of larger or longer path and overutilization of shortest path, which result in congestion in shortest path. To overcome this problem traffic engineering is used.

In IP, the feedback congestion control mechanism is via end to end, which detect and relieve congestion only at endpoint, this has to be improved. Thus, this problem can be overcome, if we extend the feedback congestion control mechanism from end points to routers. This is what we implemented in this paper.

The path the traffic should follow is represented by Label Switch Path (LSP). Thus LSP defines an ingress to egress path through an MPLS network that is followed by all packets assigned to a specific FEC. Label distribution protocol (LDP) and reservation resource routing protocol (RSVP) are used for forwarding of Labels in MPLS network. RSVP provides traffic engineering feature. LSP is unidirectional, thus to perform a two way flowing of packets two LSP have to be established.

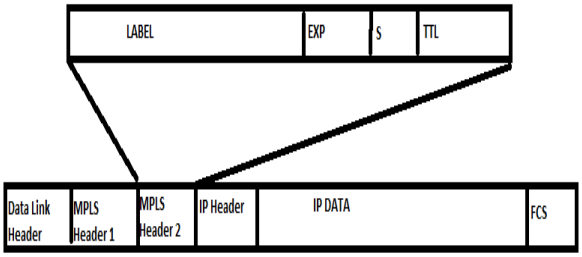
The structure of MPLS shim header, which exist specifically for MPLS is shown above.Each MPLS header is a fixed length of 4 byte (32 bits) in size and contain following fields:
Label :The label value, 20 bits

EXp :experimental use, 3bits, typically used to carry the mapping from Layer 3 TOS or layer 2 COS bits

S : bottom of stack, 1 bit, (0 = additional label follows, 1 = last entry in label stack)

TTL : Time to Live, 8 bits, used for loop prevention similar to traditional IP TTL implementation.

Since the label is a fixed length, packet indexing and look up is much simpler and faster then in conventional IP forwarding.



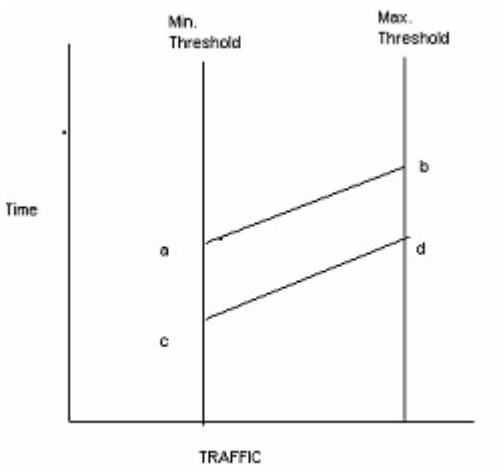
MPLS implements a general model in which a labeled packet may carry any number of labels, organised as a last in, first out sequence. This sequence is referred to as an MPLS label stack. Multiple entry label stacks are used for the implementation of MPLS based services such as VPLS, VPRN, trace, ping and others or traffic engineering application.

Overview

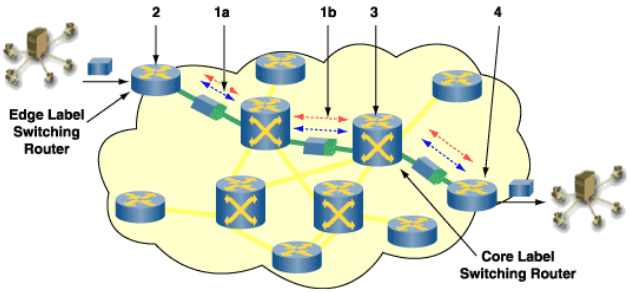
Here in this paper we are proposing a method with an objective to control traffic to achieve a throughput close to that of maximum resource capacity, with very low loss. Here in this network, network notifies user of congestion, end application should reduce traffic accordingly.

Here we are using a mechanism, to enable router participation in both congestion detection and congestion recovery. The packet flow from LER to LSR and if there is a congestion in the LSR then it adds a header to the outgoing label of 1 bit indicating congestion in the network. If the value of bit is 0 then it represent that network is conngestion free and for value 1 it represent congestion in the network.

With the use of this technique the congestion flow within the network will be within the limit that is it doesnot exceeds the threshold limit as show in the diagram below:

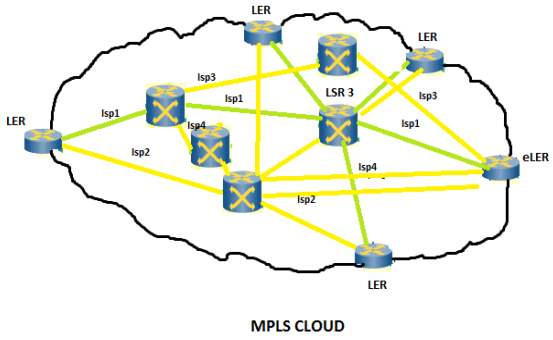


Appending Bits in Label

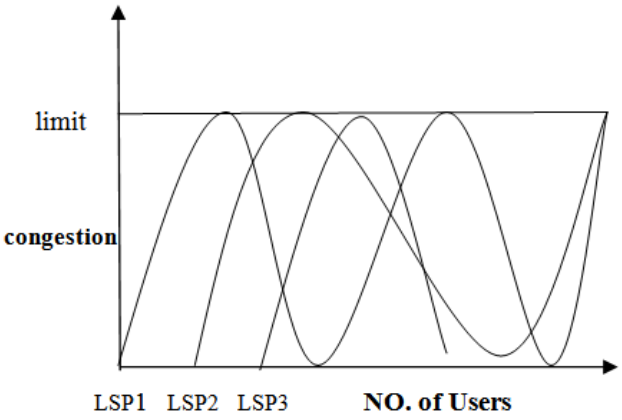


The MPLS network nodes consist of LERs and LSRs. LERs mainly implement the connection between MPLS with the external network or non-MPLS domain, or between different MPLS domains. LERs is a device at the edge of MPLS network, the edge router at the head end of labeled switch path (LSP) are ingress label edge router (iLER) and those at the tail end of LSP are classified as egress label edge router (eLER). The iLER received unlabeled packet from outside the MPLS domain, apply MPLS label to packet and forward labeled packet to MPLS domain. While eLER receive labeled packet from MPLS domain, remove the label and forward unlabeled packet outside the MPLS domain. IP table look up is done only at these two routers in MPLS network. Label switch routers are device internal to MPLS network, with all interfaces inside MPLS domain.In the core of MPLS domain the LSR ignore the IP header and simply forward the packet using MPLS label switching mechanism.

Now here we use RRATE (1) algorithm to number LSP according to the distance as suggested by IGP protocols.Now as the packet enters in the MPLS network, a label is attached at iLER 2 and forwarded to next LSRs as per the numbering. Now since all the packets follow the shortest path and by taking the shortest path all reaches to LSR 3 from different LERs. Then there will be traffic congestion at LSR 3. Now, insted of handling this congestion prevention at the end, we can overcome this congestion by enable LSR to attach a header to the packet flowing through LSP 2 egressing to source router with a bit value. This bit value represent the congestion at the egress router and force the source router to follow a different path rather than the shortest path.



The LSR shown in green represent the path taken by different LER to reach destination router eLER. Since all take their routes from LSR3, hence it will be congested. Since LSP is unidirectional hence their must be two LSP for every path hence we attach a 1 bit to the label to outgoing interface as the congestion is detected by ATCC device (3). As the congestion detected, a bit attached with packet with value 1 representing the congestion in the given path reaching to source router and other packets will take the path LSP2 that is little longer but congested free, again in the similar way if this path also get congested then a label is attached to incomming packet to source and the packets are forced to take path LSP3 and thus reduces the congestion on a specific router and thus maintain a congestion free path over the whole MPLS network.As the congestion on shortest path (LSP1)reduced, a bit with value 0 is attached with the packet and the source router prefer primary (LSP1) route, that is the route with shorter distance.Thus handling the congestion at each routers rather than at end to end, this will make our network more efficient and congestion free.This will also reduces the excessive burden of LER.



Here in this diagram it is shown that as the number of users increases the congestion increases of the LSP with highest priority that's LSP1. Now as the Limit of congestion reached the flow transfer to LSP2 of second highest priority and the congestion in LSP1 begin to decrease. Now, as the congestion in LSP1 reaches to minimum again the traffic diverted to LSP1 from LSP2 and as again the traffic increases in both LSPs then traffic or packets flow diverted to LSP3 till the traffic flow or congestion in LSP1 and LSP2 didn't decreases. Thus the traffic is maintained in the network, by diverting the traffic to less congested paths.

Conclusion

Thus in this paper we proposed a methodology of providing a congestion free path on per router bases rather than end to end. This is done by adding a bit at the end of label as their is congestion in the router. The value of the bit represent weather a path is congested (value 1) or congestion free (value 0). This bit is added to the packet taking path with source router as its destination. Then as the source router look this bit, it send other packets with next path. Thus preventing congestion.

References

[1] B.J.Oomen, S.Misra and O.C.Granmo, Routing Bandwidth-Gauranteed Path in MPLS Traffic Engineering : A Multiple Race Track Learning Approach, IEEE transaction on computers, july 2007, vol 56.

[2] Mahesh Kr. Porwal, Anjulata Yadav and S.V. Charhate, Traffic Analysis of MPLS and Non MPLS Network Including MPLS Signaling Protocols and Traffic distribution in OSPF and MPLS, IEEE 2008.

[3] Zhiqun Zhang, Xu Shao and Wei Ding, MPLS ATCC : An Active Traffic and Congestion Control Mechanism in MPLS, IEEE 2001.

[4] Xipeng Xiao, Alan Hannan and Brook Bailey, Traffic Engineering with MPLS in the Internet, IEEE aprail 2000.

[5] Anupam Gupta, Amit Kumar and Rajeev Rastogi, Exploring the Trade-off Between Label Size and Stack Depth in MPLS routing, IEEE 2003.

[6] Mohammad EL Hachimi, Marc-Andre Breton and Maria Bennani, Efficient QoS implementation for MPLS VPN, 22nd International Conferences on advanced Information Networking and Applications.

[7] Yoo-Hwa Kang and Jong-Hyup Lee, The Implementation of the Premium Services for MPLS IP VPNs, IEEE 2009.

[8] Anupam Gupta, Amit Kumar and Rajeev Rastogi, Exploring the Trade-off Between Label Size and Stack Depth in MPLS Routing, IEEE 2003.

[9] David Applegate and Mikkel Thorup, Load optimal MPLS routing with N+M label, IEEE 2003.

[10] Hooi Miin Soo and jong-Moon Chung, Anlysis of Non-Preemptive Priority Queueing of MPLS Networks with Bulk Arrivals, IEEE.

[11] "www.cisco.com/en/US/docs/internetworking/technology/handbook"from Cisco recognized site available.

Performance Analysis of Various Backoff Algorithms at MAC Layer Based on IEEE 802.11 MANET

Parul Goel[1] and Pooja Saini[2]

[1]M. Tech. Student, [2]Assistant Professor, Department of Computer Science & Engineering,
Ambala College of Engineering and Applied Research, Devasthali, Ambala Cantt-133001, Haryana, India
E-mail: er.parul007@gmail.com, apspst.09@gmail.com

Abstract
The medium access control (MAC) protocol is the main element which determines the system performance in wireless local area networks. The MAC technique of the IEEE 802.11 protocol is called Distributed Coordination Function (DCF). In IEEE 802.11 Wireless Local Area Networks (WLANs), network nodes experiencing collisions on the shared channel need to backoff for a random period of time, which is uniformly selected from the Contention Window (CW). This contention window is dynamically controlled by the Backoff algorithm. Thus, in this paper, we will use different backoff algorithms and analyse them using different parameters and the one which gives best output will be the best suited algorithm.

Keywords: MANET, Backoff Algorithms, Throughput, Packet Delivery Ratio, End to End Packet Delay.

Introduction

Wireless Network

Wireless network refer to any type of computer network that is not connected by cables of any kind and it is commonly associated with a telecommunications network whose interconnections between nodes is implemented without the use of wires.

Benefits*:*

- Convenience
- Mobility
- Easy setup
- Expandable
- Security
- Cost

MANET (Mobile Ad-Hoc Network)

A MANET is a set of wireless nodes which communicate with each other directly connected by wireless links. The nodes are free to move randomly. Thus the network's wireless topology may be unpredictable and may change rapidly. A Mobile Ad hoc Network (MANET) [1] is a dynamic wireless network that is established by a group of mobile stations without necessarily using pre- existing infrastructure. Such networks can be useful in disaster recovery where there is not enough time or resources to configure a wired network.

The routers are free to move randomly and organize themselves arbitrarily; thus, the network's wireless topology may change rapidly and unpredictably.

Due to a lack of infrastructure support, each node acts as a router, forwarding data packets for other nodes.

As nodes in MANETs are mobile, One approach to update and keep the knowledge coherent is by exchanging "hello" packets between neighbouring nodes. The "hello" packets sending process is a broadcast over the network. The broadcast generates extra traffic load over the network, consumes a part of the network resources , causes a longer delay, more control processing, and even gives more work to the backoff algorithm itself.

Rebroadcasting in the wireless network means transmitting packet by a node to all nodes in the transmission radius. This model is called on-to–all model. In this model since every node rebroadcasts the message received to all nodes this causes increasing in the collision in the network which generate a "broadcast storm problem". This problem seriously affects the network performance IEEE 802.11 MAC is a sub-layer of Data Link Layer (DLL) determined in seven layer Open system Interconnection (OSI) model [2].

The main functions provided by MAC are channel access, multiple-access and addressing. Effective medium access control (MAC) is essential to share the scarce bandwidth resource [3].

Various Challenges in MANET

- Multiple users can not able to access the network due to risk of collision.
- Time-varying wireless link characteristics: unreliable
- Broadcast nature of the wireless medium
- Hidden terminal problem and broadcast storms
- Packet losses due to transmission errors
- Mobility-induced route changes
- Mobility-induced packet losses
- Potentially frequent network partitions
- Ease of snooping on wireless transmissions (security issues)
- Limited wireless transmission range

Literature Survey

Backoff Algorithm

Introduction how it is introduced:

Two coordination functions are defined in the IEEE 802.11 MAC standard: the Point Coordination Function (PCF) and the Distributed Coordination Function (DCF).

The DCF sub layer makes use of In the DCF medium access mode, active nodes compete for the use of the channel in a distributed manner via the use of the Carrier Sensing Multiple Access with Collision Avoidance (CSMA/CA) scheme.

CSMA/CA is a layer 2 access method, not a protocol of the OSI model.[4]

Carrier sense multiple access with collision avoidance (CSMA/CA), in computer networking, is a wireless network multiple access method in which:

- a carrier sensing scheme is used.
- a node wishing to transmit data has to first listen to the channel for a predetermined amount of time to determine whether or not another node is transmitting on the channel within the wireless range. If the channel is sensed "idle," then the node is permitted to begin the transmission process. If the channel is sensed as "busy," the node defers its transmission for a random period of time. Once the transmission process begins, it is still possible for the actual transmission of application data to not occur.[1][2]

Collision avoidance is used to improve CSMA performance by not allowing wireless transmission of a node if another node is transmitting, thus reducing the probability of collision due to the use of a random truncated binary exponential backoff time.

Optionally, but almost always implemented, an IEEE 802.11 RTS/CTS exchange can be required to better handle situations such as the hidden node problem in wireless networking.[3]

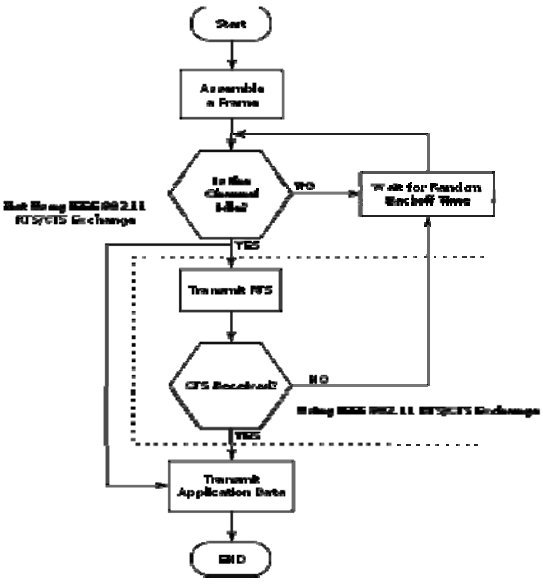


Figure 1: Simplified Algorithm of CSMA/CA.

Terms Used:

RTS- A node wishing to send data initiates the process by sending a Request to Send frame (RTS).

CTS- The destination node replies with a Clear To Send frame (CTS).

A part of the MAC protocol is the backoff algorithm.When a node over the network has a packet to send, it first senses the channel using a carrier sensing technique known as CSMA/CA (Carrier sense multiple access).

Backoff is a mechanism used to avoid collisions in mobile ad hoc networks when more than one node try to access the channel. Collision is avoided in which only one of the nodes is granted access to the channel, while other contending nodes are suspended into a backoff state for some period (BO), before trying to access the channel after a transmission failure [5].

***Binary Exponential Back*off (BEB)**

In the IEEE 802.11 standard MAC protocol, the Binary Exponential Backoff (BEB) is used. This algorithm functions in the following way [6]:

In the BEB algorithm, the contention window is doubled every time a node experiences a packet collision. If a node is successful in its packet transmission, the contention window is reset to the minimum value. In order to avoid the contention window from growing too large or shrinking too small, two bounds on CW are defined: the maximum contention window (CWmax) and the minimum contention window (CWmin). [7]

If the channel is found to be idle and not being used by any other node, the node is granted access to start transmitting. Otherwise, the node waits for an inter-frame space and the backoff mechanism is invoked. A random backoff time will be chosen in the range [0,CW-1]. A uniform random distribution is used here, where CW is the current contention window size.

To calculate the backoff time ie CW:

$$CWnew = (2*CW) * Slot\ Time \quad (1)$$

If the medium is found to be idle the CW is reduced to CWmin.

$$CW = CWmin = 32 \quad (2)$$

If the medium is determined to be busy during backoff , then the backoff timer is suspended . This means that backoff period is counted in term of idle time slots. Whenever the medium is determined to be idle for longer than an inter-frame space, backoff is resumed. When backoff is finished with a BO value of zero, a transfer should take place. If the node succeeded to send a packet and receive an acknowledgment for it, then the CW for this node is reset to the minimum, which is equal to 31 in the case of BEB. If the transfer fails, the node goes into another backoff period. When going for another backoff period again, the contention window size is exponentially increased with a maximum of 1023.

BEB has a number of disadvantages. One major disadvantage the BEB scheme suffers from a fairness problem; some nodes can achieve significantly larger throughput than others. The fairness problem occurs due to the fact that the scheme resets the contention window of a successful sender to CWmin, while other nodes continue to

maintain larger contention windows, thus reducing their chances of seizing the channel and resulting in channel domination by the successful nodes. This behavior causes what is known as "Channel capture effect" in the network.

Modified Backoff Algorithm

Modified BEB algorithm has been proposed. In this, the backoff time is increased exponentially, but with a reduced base value (less than 2) after each unsuccessful transmission until prescribed maximum value (CWmax) is reached. Whenever a node transmits a packet successfully, backoff time is reduced to a specified minimum value (CWmin). [8]

The CW Exponentially increased and then decreased to CWmin every time a node experiences a packet collision or success.

In case of collision CW :

CWnew =1.5 * CW * Slot Time (3)

In case channel is idel:

CW= CWmin = 32 (4)

The used formula provides different outcome for the backoff times, the behavior of the two formulas can be seen in the Figure1.

Multiplicative Increase Linear Decrease (MILD) Backoff Algorithm

MILD is proposed where a node increases its CW l by 1.5 after every unsuccessful transmission and decreases its backoff interval by one after successful transmission. [10]

In case of collision CW

CWnew = 1.5 * CW * slot time (5)

In case channel is idel:

CW = CWnew – 1 (6)

Exponantial Increase Exponantial Decrease (EIED) Backoff Algorithm

EIED is proposed to enhance the performance of DCF. In this scheme, the CW Exponentially increased(2√2, 21/4) and Exponentially decreased every time a node experiences a packet collision or success.[9]

In case of collision CW :

CWnew = CW * 2^1/2 * slot time (7)

In case channel is idel:

CW = CWnew / 2^1/2 (8)

Double Increment and Double Decrement (DIDD) Backoff Algorithm

DIDE has been proposed, in this algorithm, the CW Doubles itself and reduced itself to half every time a node experiences a packet collision or success.

In case of collision CW :

CWnew = 2 * CW * Slot Time (9)

In case channel is idel:

CW = ½ (CWnew) (10)

Logarithmic Backoff Algorithm (LOB)

The CW is Multiplied with the log of the old CW by and respective slot time every time a node experiences a packet collision or success.[10]

In case of collision CW :

CWnew = (log(CW)) CW * Slot Time (11)

In case channel is idel:

CW = CWmin = 32 (12)

Fibonacci Incremental Backoff Algorithm (FIB)

The CW is increased according to values formed from math series called Fibonacci series and decreased by same factor every time a node experiences a packet collision or success.[11]

In case of collision CW :

CWnew = CW * Fn * Slot Time (13)

// (Fn = Fn-1 + Fn-2) -> Fabonacci Series

In case channel is idel:

CW = CWmin (14)

Pessimistic Linear Exponantial Backoff (PLEB)

PLEB is another proposed backoff algorithm which uses a combination of two increment behaviors; Exponential backoff and Linear backoff.In this the stage when the channel is busy i.e we have unsuccessful transmissioin or collision CW increases exponentially as in BEB and as successful transmission comes CW decreases linearly.[12]

Exponential increments give enough backoff time to enhance the network throughput by reducing the number of transmission failures, and the linear increment reduces the average packet delay.

In case of collision CW :

CWnew = CW * 2^1/2 * Slot Time (15)

In case channel is idel:

CW= CW – 1 (16)

Problem Formulation

Performance Analysis of Various Backoff Algorithms at MAC Layer based on IEEE 802.11 MANET

Performance Parameters: **[13]**

Throughput

Packet Delivery ratio

$$PDR = \frac{Total\ number\ of\ packets\ received}{Total\ number\ of\ packets\ sent}$$

Average End to End delay

$$\text{Average end–to-end delay} = \frac{\left[\sum_{i=1}^{m} \text{Sum of average end-to-end delay for each destination}\right]}{m}$$

Introduction to GloMoSim Simulator

Introduction

With GloMoSim we are building a scalable simulation environment for wireless network systems.

Most network systems are currently built using a layered approach that is similar to the OSI seven layer network architecture. The plan is to build GloMoSim using a similar layered approach. The goal is to build a library of parallelized models that can be used for the evaluation of a variety of wireless network protocols. The proposed protocol stack will include models for the channel, radio, MAC, network, transport, and higher layers.

This simulator is used to simulate the performance of environment.

- Select N no of nodes
- CBR(Constant Bit Rate)
- Type of Environment(noisy, smooth)
- Slot Time(no of nodes per slot time)

Conclusion

Analyse the performance using GloMoSim (Global Mobile Information Systems Simulatior) SIMULATOR of different Backoff Algorithms applied on the network at IEEE 802.11 MAC layer using parameters “throughput”, “packet delivery ratio” and “end to end delay” in order to minimize the node or packet collision. And the algorithm having best performance will be the best suited one.

References

[1] Z.Fang, et al., “Performance evaluation of a fair backoff algorithm for IEEE 802.11 DFWMAC.” International Symposium on Mobile Ad Hoc Networking & Computing

[2] S. Xu, And T. Saadawi, “Does the IEEE 802.11 MACprotocol work well in multihop wireless ad hoc networks?” *IEEE Communications Magazine*, pp 130 – 137, 2001.

[3] K. Sakakibara, et al., "Backoff Algorithm with Release Stages for Slotted ALOHA Systems." ECTI Transactions On Electrical Eng., Electronics, And Communications vol.3, no.1 pp 59-70 ,2005.

[4] H. Zhai and Y. Fang,”Performance of Wireless LANs Based on IEEE 802.11 protocols.” 14th IEEE International Symposium on Personal, Indoor and Mobile Radio Communication Proceedings, pp 2586-2590, 2003.

[5] R. Ramanathan and J. Redi, “A Brief Overview of Ad Hoc Networks: Challenges and Directions,” IEEE Communications Magazine, 50th Anniversary Commemorative Issue, pp. 20-22, May 2002.

[6] IEEE Std. 802.11, Wireless LAN Media Access Control (MAC) and Physical Layer (PHY) Specifications, ISO/IEC 8802-11, 1999.

[7] V. Bhargavan , A.Demers , S. Shenker and L. Zhang, “MACAW : A Media Access Protocol for Wireless LANs,” Proc. of the Conference on Communications Architectures, Protocols and Applications, pp. 212-225, ACM Press, 1994.

[8] C. Rama Krishna, Saswat Chakrabarti and Debasish Datta, “A modified backoff algorithm for IEEE 802.11 DCF based MAC protocol in a mobile ad hoc network,” IEEE TENCON 2004, vol. B, vol. 2, pp. 664-667, November 2004.

[9] Nai-Oak Song, ByungJae Kwak and Jabin Song, Leonard E. Miller, “Enhancement of IEEE 802.11 distributed coordination function with exponential increase exponential decrease backoff algorithm,” Proc. of VTC, 2003.

[10] S. Manaseer and M. Ould-kauoa, “Logarithmic Based Backoff Algorithm for MAC Protocol in MANETs,” DCS Technical Report Series, 2006.

[11] P. Chatzimisios, A. C. Boucouvalas, V. Vitsas, A. Vafiadis, A. Economidis and P. Huang, “Fabonacci Increment backoff scheme for the IEEE 802.11 MAC protocol”.

[12] S. Manaseer and M. Masadeh, “Pessimistic Backoff for Mobile Ad hoc Network”. ICIT’09

[13] Pooja Saini “Impact of Mobility and Transmission Range on the Performance of Backoff Algorithms for IEEE 802.11-Based Multi-hop Mobile Ad hoc Networks” International Journal of Advancements in Technology (IJoAT)

Performance Evaluation of VOIP in MultiHop Wireless Mesh Network

[1]Kamal Kumar and [2]Pooja Saini

[1]M.Tech. Student, Assistant Processor, Department of Computer Science & Engineering,
Ambala College of Engineering and Applied Research, Devasthali, Ambala Cantt-133001, Haryana, India
E-mail: kadam.366@gmail.com, apspst.09@gmail.com

Abstract
Wireless Mesh Network (WMN) is considered to be an effective solution to support multimedia services in last miles due to their automatic configuration and low cost deployment. The main feature of WMNs is multihop communications which may result in increased region coverage, better robustness and more capacity. Wireless mesh networks (WMNs) will play an important role in the next-generation wireless communication systems because it can support broadband services with ubiquitous coverage by low power consumption. Voice internet over protocol is an important services of wireless mesh network. Wireless Voip is more popular due to low cost and easy to deploy. Wireless mesh network is a good solution for Voip services deployment, performance evaluation. Performance is main challenges in QOS phase in this paper analyze four parameters to evaluate performance.

Keywords: WMN, VOiP, Throughput, Packet Delivery Ratio, End to End Packet Delay, Zitter.

Introduction

Wireless Mesh Network

Wireless Mesh Networks (WMNs) are an emerging two tier architecture based on multihop transmission. Wireless Mesh Networks (WMN) are gaining attention as a cost-efficient way for providing broadband wireless Internet access. The IEEE 802.11s task group is aimed to form a transparent 802.11 broadcast domain with the same functionality as its wired counterpart. Multi-hop WMNs have several benefits. In comparison to infrastructure networks with single wireless links, multi hop WMNs can extend the coverage of a network and improve the connectivity. The number of fixed Internet access point scan be reduced leading to a cheaper network access as several users share Internet connectivity by multi hopping towards the access routers. Multihop WMNs avoid a wide deployment of cables and can be rapidly deployed in a cost-efficient way. Incase of dense multi hop networks, the use of multi-radio multi channel multi channel mesh nodes increases network capacity, and therefore several paths might become available increasing the network's robustness. The provisioning of VOiP in multi-hop WMNs is an important service for the future wireless Internet. However, VOiP service poses new challenges when deployed over a multihop WMN. Packet losses and an increased delay due to interference in a multiple hop network can significantly degrade the end-to-end VOiP call quality. High traffic leads to high medium contention which increases packet loss rates compared to single hop deployments. [2]

The existence of potential hidden nodes further intensifies this problem. Moreover, the transmission of small (voice) packets imposes a high MAC layer overhead, which leads to a low capacity for VoIP over IEEE 802.11-based WMNs.

The main function of the networking layer is to transfer the packets from the source to the destination over multiple hops. In this respect, WMNs are radically different from 3G systems, WLANs and WMANs. All these technologies use a single wireless link, and hence have no need for a network layer. In contrast, for WMNs and MANETs the source and the destination can be several wireless hops away from each other, and hence the packets have to be routed and forwarded in the wireless network itself. Wireless mesh networks (WMNs) will play an important role in the next-generation wireless communication systems because it can support broadband services with ubiquitous coverage by low power consumption.

Performance evaluation and analysis of different VOiP over WLAN mobility schemes is essential in order to meet adequate QoS levels in VoWLAN deployments. Voice over Internet Protocol (VoiP) service has been every popular and important application over the Internet. Wireless VOiP also becomes more and more popular due to its features of low cost and convenience. Recently, wireless mesh network has been considered as a good solution for VOiP services since it is easy to deploy and provides a larger area coverage. However, the security and performance are the main challenges.

This paper presents a study of real time voice communication QoS in terms of delay, jitter, throughput and packet loss between three different Wireless Mesh Network mobility schemes in a crowded 802.11 environment [1]. The rest of the paper is organized as follows. Section 2 provides an overview of the literature survey. Section 3 explain the problem formulation. Section4 explain the methodology to evaluation of VOiP performance. Finally, Section 5explain the Future scope of the paper.

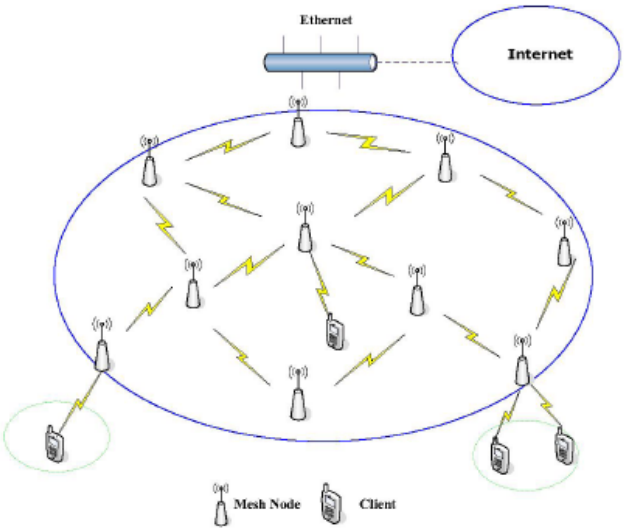


Fig. 1 Architecture of Wireless Mesh Networks

Voice Over Internet Protocol [5]

The provisioning of VOip in multihop WMNs is an important service for the future wireless Internet. However, VOiP service poses new challenges when deployed over a multi-hop WMN. Packet losses and an increased delay due to interference in a multiple hop network can significantly degrade the end-to-end VOiP call quality. High traffic leads to high medium contention which increases packet loss rates compared to single hop deployments. The existence of potential hidden nodes further intensifies this problem. Moreover, the transmission of small (voice) packets imposes a high MAC layer overhead, which leads to IEEE 802.11 MAC is a sub-layer of Data Link Layer (DLL) determined in seven layer Open system Interconnection (OSI) model. The main functions provided by MAC are channel access, multiple-access and addressing. Effective medium access control (MAC) is essential to share the scarce bandwidth resource.

Literature Survey

We divided our study in three types of tests. On each test, the same path from point A to point B was followed during a 90 second walk. We built a small wired LAN that consisted of a PC and the mesh portal from each considered scheme.

The types of tests are presented as follows:

VOiP CALLS

We deployed a small VOiP wired network consisting of a computer running Asterisk and one IP Phone. In this scenario, the IP Phone made calls to a SIP softphone running on a laptop. By using Wireshark we sniffed the incoming packets from the client's side perspective. This data was used to calculate jitter and delay.

Throughput Test(TCP)

We used iPerf to determine that one single client could get by the maximum throughput sending a continuous TCP stream. In order to meet the highest possible throughput, we set the default TCP window size of 56kb with a buffer length of 2Mb.

Packet Loss And Duplicate Packets

For these tests, we simulated a VOiP call by programming a Python script that sent 160-byte length UDP packets every 20 ms, thus generating a 64kbps stream. We implemented a Java script that sniffed the incoming packets in order to calculate packet loss. Because of the redundant multipath routing nature of SMesh generating duplicated packets, we also added a sequence number on the data portion of the packet. Our study determined that the wireless mesh networks presented higher QoS levels compared to a data only 802.11 infrastructure network. Our study also determined that GARP broadcasting schemes provide the lowest delay values, but more unstable throughput. Seamless handoff solutions with redundant multipath routing, such as SMesh, provide the most stable throughput as well as acceptable QOS values for real time voice communication [1].

This Section handles the MeshBed which is a next generation WLAN based Wireless Mesh Network, developed and deployed at T-Systems in Darmstadt, Germany. In its current state the MeshBed consists of 10 Mesh Router Nodes (MRNs) and 2 Mesh Gateways (MGWs) that are all deployed indoors. As hardware platform an embedded AMD Geode SC1100 Systems with 266 MHz CPUs and 64 MB of RAM is used. For nodes that require more processing power, e.g. MGWs, bare bone desktop PCs with 3 GHz Intel Pentium 4 processors and 1 GB of RAM are used. All mesh nodes are equipped with Atheros Wireless Mini PCI WLAN cards as well as Ethernet ports and use operating systems based on Linux together with "madwifi" an open-source WLAN driver. This mesh environment is designed in accordance with a strategy towards 802.11s, which assumes the usage of advanced MAC technique, namely link layer routing. Currently, packets are still routed at the network layer. The testbed emulates dual-radio feature, we call it pre-IEEE 802.11s.The paper addresses the deployment VOiP service in pre- IEEE 802.11s WMN and means for its performance optimization. VOiP, being a part of Triple play service bundle, was chosen as a reference service for extensive measurements. The general finding of the experiments is, that VOiP can be supported with good quality in mesh environment. However, under high load, quality drops and additional mechanisms are needed to overcome these problems. Moreover, it was demonstrated how the VOiP traffic may benefit from the small packet aggregation. A novel hop-by-hop packet aggregation mechanism was proposed. It significantly improves the performance of VOiP traffic in WMNs and reduces MAC layer busy time. [2]

Wireless Mesh Networks (WMNs) are an emerging two tier architecture based on multihop transmission. In Akyildizet al. give a complete overview of WMNs, with an analysis of the challenges and the open issues for this kind of networks. Improving the performance in face of unreliable wireless medium is among the key factors affecting the scalability, and will support WMNs become a very cost-effective solution for wireless ISPs. In such context, routing has a crucial importance and a deep impact on the overall network performance, due to the unpredictable behavior of the wireless channel and the strong layering of the protocol stack. DeCouto et al. in showed that routing in multi-hop wireless networks using the traditional Shortest-Path metric is not a sufficient

condition to construct good paths. By good paths we mean paths able to effectively transport data with reasonable delay, throughput, and reliability. Indeed, the shortest-path metric does not take into account the variable quality of the wireless link. As a consequence, there is the need for the protocol stack to be aware of the nature and capabilities of the lower layers.

Giving to the routing layer the awareness of key parameters of the underlying layers enables the possibility to make smarter path selection. Note that this does not mean to sample some parameters at routing layer, like quality of service (QoS) routing, but to sample some parameters at the layer where they pertain, and allow exchanging this information with the routing layer. Several routing approaches proposed in the literature rely exactly on this idea, i.e. to use a set of parameters from the Physical or Data-Link layer in order to characterize the quality of a link and consecutively make routing decision. The solution proposed and evaluated in this paper is such a cross layer routing approach based on interaction between MAC layer and Network layer. Through carefully study of IEEE 802.11 MAC layer, we find that the retransmission mechanism of lost packets has a strong impact on link throughput and end-to end latency along the path. The expected transmission efficiency (ETE) metric based on it finds path with the highest transmission efficiency required to deliver a packet. ETE estimates the success rate of sending data using per-link measurements of DATA/RTS retransmission. Thus it is a good estimate of both channel condition and congestion. [4]

Scheduling is one of the main issues in maintaining network performance. As the efficiency objectives in MAC layer are throughput and delay, scheduling criteria for wireless networks include:

Efficiency, Applicability, QoS support, and Fairness. Although in order to support QoS, differentiate services are preferred in general, however, this is not feasible in mesh network architectures. Therefore, it seems that necessary design for a QOS centric architecture to support appropriate scheduling services within this framework still requires more investigations. 802.16 Standard defmes two centralized and distributed scheduling schemes to coordinate using simultaneous mini slots (MSs)in data sub-frames. The centralized scheme in mesh mode is managed by Base Stations (BSs) to form a scheduling tree whose root is BS itself. Centralized scheme applies a 2-way handshaking for source transmission requests to be granted in data sub-frames. Centralized scheduler cannot be used for bandwidth allocation to links that are not present in scheduling tree; therefore, in such cases horizontal links will be overlooked. In this model, each node competes with communicational nodes within its two hop distance to acquire bandwidth.

The distributed scheduling itself is classified to coordinated and non coordinated categories. The coordinated distributed scheduling schedulers allow mesh nodes to transmit distributed scheduling signaling messages that contain bandwidth requests in control sub-frames, without any collision. In coordinated distributed scheduling, nodes schedule their transmissions with their two adjacent hopes using a three-way handshaking (request, grant and grant confirmation) for reservation of bandwidth in a link. To perform a non-collision scheduling, each node memorizes (locally) the positions of MSs based on information obtained from its neighbours. 802.16 standard employs an election distributed algorithm to access to transmission opportunities in control sub-frames so that when a node is in transmission, no other node at least within its two hope neighborhood attempts to transmit simultaneously.

This ensures that neighbors of a node are more likely to be able to receive the transmitted control messages properly. In non-coordinated distributed scheduling, similar mechanism to that of coordinated distributed scheduler is used with only difference in mesh election algorithm. This difference lies in the fact that handshaking messages transmitted in data subframes and in MSs, are reserved for specified links instead of being transmitted in control sub-frames. The framework for request process and allocation of bandwidth by 802.16 standard is defined. WMNs had been viewed as networks that fail to meet multimedia QoS requirements due to using multi hop wireless communications. However a number of investigations have led to solve some WMN dmwbacks but still a lot of unsolved problems persist. Mesh network scheduling has also been one of the centers of attention. Some proposed centralized scheduling for inter-mesh flows have been able to provide end to end QoS. But distributed scheduling provides facilities for intra-mesh traffic between client stations. Such a scheduling model does not provide end to end QoS for flows but makes mesh network suitable for situations in which a resistant and scalable ad hoc network is needed [6].

Problem Formulation

Performance Evaluation of VOIP in (MultiHop Wireless Mesh Network)using different performance throughput, packet delivery ratio, end to end delay and “zitter”.

Methodology

The evaluation is carried out with the GloMoSim simulator by performing several experiments that illustrate the performance of the system. The simulation parameters like number of nodes, terrain range etc. Along with their respective values are used to examine the performance of the network. These parameters are available in the “config.in” file present in the GloMoSim. The values can be adjusted according to requirements in this file. After adjusting the values in this file, this file is executed. An output file “glomo.stat” is used to check the various parameters to analyze the performance of network..

The performance metrics used for evaluation are:

Average Throughput

Defined as packets recived successfully to given time interval.

End To End Delay

Network delay is the total latency experienced by a packet to traverse the network from the source to the destination. At the network layer, the end-to-end packet latency is the sum of processing delay, packet, transmission delay, queuing delay and propagation delay.

Packet Delivery Ratio (PDR)

$$PDR = \frac{Total\ number\ of\ packets\ received}{Total\ number\ of\ packets\ sent}$$

ZITTER

Based on different parameters VOiP are simulated and analyzed. All four parameters are evaluated by finding out average throughput, end-to-end delay and PDR, Zitter by varying number of nodes and node mobility. [7]

Future Scope

Using above four parameters evaluate performance of VOiP in WMN and result would be simulate using GloMoSim simulator.By using different scenarios analyze values to given parameters.

References

[1] Basurto J C, Estrada R "An Experiment Study of VoIP Performance in Wireless Mesh Networks Using Different Mobility Approaches", *2nd* International Conference on Software Technology and Engineering(ICSTE)pp 325-329, (march 2010).

[2] Bayer N, Castro M C, Dely P, Kesslar A"VoIP service performance optimization in pre-IEEE802.11s Wireless mesh network", pp75-79, 2008.

[3] Bayer N, Hock D, Roos A, Siebert M, Xu B, Rakocevic V, Habermann J, "VoIP performance in "MeshBed" - a Wireless MeshNetworksTestbed", (BMBF), pp 2218-2222, august 2008.

[4] Jiang L, Feng G, "*A MAC Aware Cross-Layer Routing Approach forWireless Mesh Network"*, (WICCOM), pp 1-6, 2006.

[5] Xian Y, Huang C, " Securing VoIP Services in Multi-Hop wireless Mesh Networks", ISWCS 2007 pp 513-517, 9 july 2007.

[6] Ghazvini M, Movahedinia N "Scheduling Algorithms in Wireless Mesh Networks", Second Pacific-Asia ConferenceonCircuits, Communications and System (PACCS)pp86-89, sept2010.

[7] Singh Y, Chaba Y, Jain M, Rani P, "Performance Evaluation of On-Demand Multicasting Routing Protocols in Mobile Adhoc Networks", International Conference on Recent Trends in Information, Telecommunication and Computing, pp 298-301, october 2010.

Bias Current Effect on Gain of a CMOS OTA

[1]Manoj K. Taleja and [2]Manoj Kumar

[1]Department of Electronics & Communication Engg.,
Guru Jambheshwar University of Science & Technology Hisar, Haryana, (India)
E-mail: manojtaleja@yahoo.com
[2]Department of Electronics & Communication Engg.,
Guru Jambheshwar University of Science & Technology Hisar, Haryana, (India)
E-mail: manoj_jogi10@yahoo.com

Abstract
In this paper, a study in terms of gain for differential CMOS operational transconductance amplifier (OTA) using super cascode transistors has been carried out. A regulated cascode circuit provides high output impedance at low supply voltage. OTA using super cascode transistors shows gain of 68 dB at 100 µA bias current. The input bias current from 1 pA to 100 µA has been varied. The power consumption of OTA varies from 77 nW to 811 mW with bias current variations from 1 pA to 100 µA. The CMOS OTA has been simulated in 0.50µm technology with 1.5V power supply voltage.

Keywords: Bias current, CMOS, gain, low power, transconductance.

Introduction

In the recent years, efforts have been made to reduce supply voltage of integrated circuits. Research efforts also have been made in reducing total power consumption of VLSI systems [1]. The realization of low voltage, high gain and low power amplifiers requires efficient circuit design techniques [2]. A folded cascode configuration has been widely used in CMOS circuits. The reason is that the output capacitance plays the same role as the compensation capacitance [3]. This configuration also increases transconductance (g_0) which in turns increases the gain of the amplifier. Today high performance and low power circuits are required that can be operated at a low supply voltage of 1.5 V or less [4]. A super cascode (SC) transistor technique reported in literature is used to increase the gain of the amplifier. Since amplifiers in super cascode circuits have to drive capacitive loads, so they have to make use of operational transconductance amplifiers (OTAs). The OTA using super cascode transistors has to meet three requirements: high slew rate, large bandwidth and high open loop gain. Gain boosting is another technique which increases output impedance and hence the gain of the amplifier. A single stage folded cascode OTA could be used for further applications [5].

Super Cascode Transistors

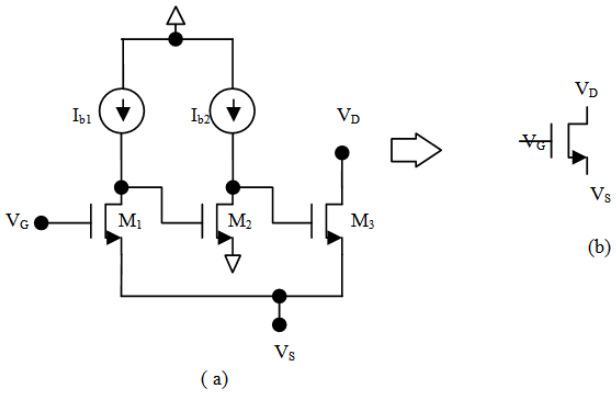


Figure 1: Super Cascode Transistor [4] (a.) Circuit schematic (b) Symbol

Figure 1 shows super cascode transistors. Transistors M_1, M_2, M_3 along with biasing currents I_{b1} and I_{b2} constitute 'super cascode transistors' [6]. The drain of transistor M_1 drives transistor M_3. Since transistor M_1 is common source, hence polarity of drain of transistor M_1 reversed. So an inverting stage is provided by transistor M_2 and biasing current I_{b2} to drive the base of transistor M_3. The biasing current I_{b2} boost the gain which increases the output impedance [7].

OTA Using Super Cascode Transistors

Figure 2 shows OTA using super cascode transistors with cascoding technique which increases the gain without affecting high frequency characteristics. SC transistors used in output stage gives small settling time and high gain. SC stage increases the gain even more than a single stage. The currents in cascode stage and input differential pair are equals. Transistors M_C along with control signal T_C controls the common mode output voltage which is taken from SC transistors [7].

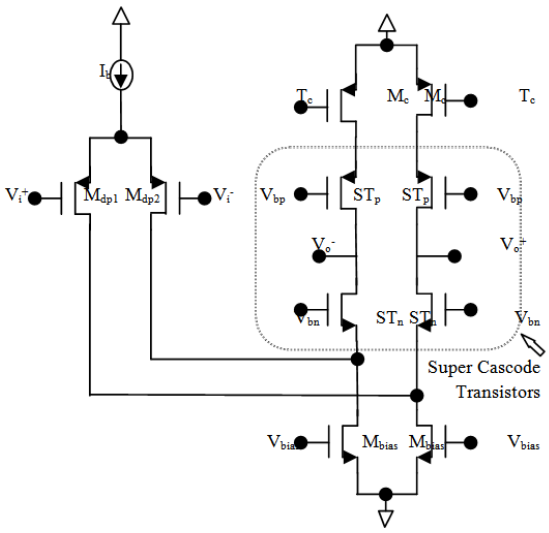


Figure 2: OTA using Super Cascode Transistors [4].

Results and Discussions

The OTA using SC transistors has been simulated using 0.5 µm technology with a supply voltage of 1.5 V. The biasing current (1 pA – 100 µA) that controls the gain of the OTA is shown in Table 1. The biasing current I_{b2} is taken 10 times larger than I_{b1} to satisfy stability criteria [4]. The OTA with SC transistors shows a gain of 68 dB at a biasing current of 100 µA. The results are shown in Table 1. Table 2 shows values of 'biasing current and gain relationship' and 'biasing current and power dissipation across OTA' relationship. The graph showing biasing current and gain relationship is shown in figure 3 and the graph showing biasing current and power dissipation across OTA relationship is shown in figure 4.

Table 1: Performance characteristics for CMOS amplifiers.

Design parameters	OTA using Super Cascode Transistors [4]	Conventional amplifier [8]
Power supply	1.5 V	2.5 V
DC voltage gain	68 dB	48 dB
Input bias current	1 pA – 100 µA	1 mA

Table 2: Power and gain Vs biasing current.

Biasing current	Gain of OTA using SC Transistors	Power dissipation across this OTA
1 pA	63.52	77.565 nW
500 pA	63.53	98.790 nW
1 nA	63.54	124.317 nW
500 nA	63.57	1.418 mW
1 µA	63.63	2.934 mW
25 µA	65.60	111.926 mW
50 µA	67.00	289.032 mW
75 µA	68.04	524.248 mW
100 µA	68.82	811.280 mW

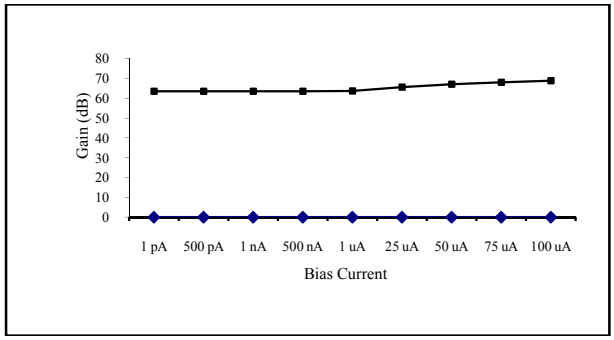


Figure 3: Gain with bias current.

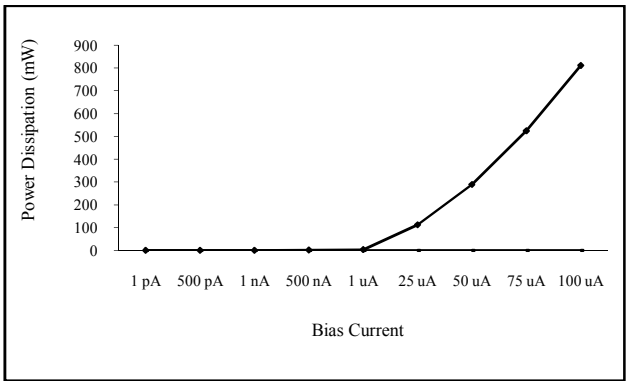


Figure 4: Power consumption with bias current.

Conclusions

A low voltage, high gain differential OTA using SC transistors has been studied using 0.5µm technology. The circuit has been simulated using SPICE. The OTA shows a gain of 68 dB at a biasing current of 100 µA. The gain of OTA using SC transistors varies from 63.52 dB to 68.82 dB with a biasing current varying from 1 pA to 100 µA as shown in figure 3. The power consumption of OTA using SC transistors varies from 77.565 nW to 811.280 mW with a biasing current varying from 1 pA to 100 µA as shown in figure 4.

References

[1] L.Bouzerara, M.T.Belaroussi, and B.Amirouche, "Low-voltage, low-power and high gain CMOS OTA using active positive feedback with feed forward and FDCM techniques," in 23rd international conference proc. on microelectronics, vol 2, NIS, Yugoslavia, 12-15 May, 2002.

[2] Subhajit Sen and Bosco Leung, "A class-AB high-speed low-power operational amplifier in BiCMOS technology," IEEE journal of solid – state, vol. 31, no. 9, pp 1325-1330, Sept. 1996.

[3] F. Op't Eynde and W.Sansen, "A CMOS wideband amplifier with 800 MHz gain-bandwidth," in IEEE custom integrated circuit's conference proc., pp 9.1.1-9.1.4, June 1990.

[4] R.G. Carvajal, B. Palomo, A. Torralba, F. Munoz and J. Ramirez-Angulo, "Low voltage high gain

differential OTA for SC circuits," Electronic Letters, vol. 39, no. 16, 7th August 2003.

[5] Bult. K. and Gallen. J.G.M., "The CMOS gain – boosting technique," Analog Integrated Circuits Signal process, pp 119-135, 1991.

[6] Torralba. A., Carvajal. R.G.,Ramirez-Angulo. J. and Munoz E., "Output stage for low supply voltage, high-performance CMOS current mirrors," Electronic Letters, 38, (24), pp 1528-1529, 2002.

[7] Bult. K. and Gallen. G., "A fast-settling CMOS op-amp for SC circuits with 90 dB dc gain," IEEE journal of solid–state circuits, 25, (6), 1990.

[8] Katsufumi Nakamura and L. Richard Carley, "An enhanced fully differential folded cascode op amp," IEEE journal of solid - state circuits, vol. 27, no. 4, pp 563-568, April 1992.

Efficient Grid Resource Selection based on Performance Measures

[1]Anjali, [2]Savita Khurana and [3]Meenakshi Sharma

[1]M.Tech., Pursuing, SSCET, Pathankot, India
[2]Assistant Professor, JMIT, Radaur, India
[3]Professor, SSCET, Pathankot, India
E-mail: kalra.anjali@gmail.com, savu.khurana30@gmail.com, sharma.minaxi@gmail.com

Abstract
Grid computing system is the one where individual entities share their resources with others in their own domain as well as with other domain. Resource Discovery and Resource Selection are the crucial tasks in grid scheduling and resource management. The goal of resource discovery and selection is to identify list of authenticated resources that are available in the grid for job submission and to choose the best node. The resources should be shared efficiently without losing their features and performance. Performance prediction in Grid computing is an important challenge because of volatiles, heterogeneous and unreliable Grid environments. Performance prediction, intends to provide real-time forecast of important performance metrics (such as application run time and queue wait time) which can support Resource selection decisions. For a Grid middleware to perform resource allocation, prediction models are needed, which can determine how long an application will take for completion on a particular platform or configuration. The predictions can include both estimations of the execution time of an application for a range of problem sizes, and predictions of how an application performs on different Grid resources. We proposed an algorithm for resource selection in grid computing using the concept of performance prediction. Various performance prediction tools are usage count, cost, availability, reliability, trust value.

Keywords: Usage Count, Performance, Reliability, MTBF, MTTR, GPP.

Introduction

Grid computing system is a collection of distributed computing resources available over a local or wide area network that appears to an end user or application as one large virtual computing system. The aim of grid system is to create virtual dynamic organizations through secure, coordinated resource sharing among individuals, institutions, and resources. Grid computing is to provide unlimited power, collaboration, and information access to everyone connected to grid. Among the key challenges of current Grid technologies is the problem of resource selection and brokering. The set of resources chosen for a particular job can vary strongly depending on the goals of the researchers, and might involve minimizing the costs, optimizing the location of the computational resources in order to access a large data set or minimizing the overall execution time of the job. Resource selection is an important issue in a grid environment where a consumer and a Service Provider(SP) are distributed geographically across multiple administrative domains. In such a scenario, it is essential that the selection of a SP is not only based on service functionality alone, but also on the process behavior such as trustworthiness and Quality of Service of the SP.

Grids will develop to deliver high performance computing capabilities with flexible resource sharing to dynamic virtual organizations [1]. Accurate predictions of the performance of Grid jobs are important, e.g., to improve resource selection (Grid brokering), coordinate task scheduling of workflows, and increase the utilization of individual clusters. Essential to this growth is the development of fundamental infrastructure services that will integrate and manage large heterogeneous pools of computing resources, and offer them seamlessly to differentiated users[2]. A good prediction service not only provides grounds for such decision making about execution of application through Grid middleware but also facilitates to improve the reliability of execution by providing reliability predictions about resources. The Grid Performance Prediction (GPP) requirements span over multiple dimensions. They range from simple execution time prediction of an activity/task, prediction about a machine and different resources within a machine, network level prediction, workflow level prediction, and cluster-wide prediction to the Grid-wide prediction. Grid applications utilize high performance distributed resources similar to high performance systems, networks, databases, etc. Much of the work was done on finding an optimal selection of resources in Grid computing environments. The selection schemes are divided into two main categories; conventional and economical. The conventional strategies consider the overall performance of the system as a metric for determining the system quality. It does not take the cost as factor for scheduling jobs on resources and treat all resources as the same at all. Some examples are SmartNet, AppleS Project, Condor-G, NetSolve etc. In economic strategy, cost is considered as essential factor for scheduling jobs. The user is charged based on the utility of the resources in the Grid system[3].

In a grid environment, resource objects submitted by users may contain hardware resources (CPU, memory, bandwidth etc) and software resources （JRE, Loadlever, Gcc etc）. Those grid resources that meet the minimization

demands of the user will be found, and the best resource objects will be selected and be provided for the user. Once the list of possible target resource objects is known, the second phase of the scheduling process is to select those resource objects that best suit the constraints and conditions imposed by the user [4]. The result of resource selection is to identify a resource object list of candidates in which all resources can meet the minimum requirements for a submitted job or a job list. The relationship between resources available (Ravailable) and resources selected (Rselected) is: Rselected⊆Ravailable It's different for hardware resources and software resources in the selection process. The application of hardware resources can always be assured by comparing with candidate one based on a quantitative evaluation, while software resources can be assured only by if it is present. Rest of the paper is organized as: section 2 presents the related work. Section 3 describes the Resource selection algorithm. Section 4 includes the experimental Results and discussion. Section 5 concludes the paper.

Related Work

Our work is inspired by a number of previous works related to resource selection in grid environment. These related works are reviewed below:

Thomas Kevin et al presents a resource selection mechanism using price prediction technique. In this paper author presented an integrated Grid market of computational resources based on combining a market-based resource allocation system, Tycoon, and a Grid meta scheduler and job submission framework. Ren et al. [3] leverage host CPU utilization and resource contention traces to develop a model for resource availability prediction. The model utilize state transition based prediction and produce a Markov chain using the previous days of a resource's history.

Farag Azzedin and Muthucumaru Maheswaran [11] proposed a formal definition of both trust and reputation and discussed a model for incorporating trust into Grid systems. Rajkumar Buyya and Srikumar Venugopal [12] proposed an overview of an open source Grid toolkit, called Gridbus, whose architecture is fundamentally driven by the requirements of Grid economy. Gridbus technologies provide services for both computational and data grids that power the emerging eScience and eBusiness applications. Ernesto Damiani et al. [9] proposed a self-regulating system for P2P network using robust reputation mechanism. In their system reputation sharing is realized through distributed polling algorithm. Chuang Liu et al. [4] proposed a general-purpose resource selection framework by defining a resource selection service for locating Grid resources that match application requirements and evaluated them based on specified performance model and mapping strategies, and returned a suitable collection of resources, if any are available. Yao Wang and Julita Vassileva [7] proposed a Bayesian network-based trust model and a method for building reputation based on recommendations in peer-to-peer networks. Sepandar D. Kamvar et al. [5] proposed a reputation management system, called EigenTrust, which can effectively reduce the number of downloads of inauthentic files in a P2P system. The reputation value of each peer is determined by the number of successful downloads and the "opinions" of other peers. Shanshan Song and Kai Hwang [8] proposed a new fuzzy-logic trust model for securing Grid computing across multiple resources sites. They have developed a new Grid security scheme, called SARAH supported by Spooner, Jarvis, Caoy, Sainiz and Nudd et al presents development of a multi-tiered scheduling architecture (TITAN) that employs a performance prediction system (PACE) and task distribution brokers to meet user-defined deadlines and improve resource usage efficiency. This paper focuses on the lowest tier which is responsible for *local* scheduling. They uses just in time approach for evaluating the performance models of an system. Noorisyam Hamid et al.[10] proposes a resource selection technique based on resource ranking. They emphasizes on improving the current resource selection technique in grid scheduler by taking into account the quality and reliability of both users and resources. Their Quality-based Grid Resource Discovery aims at providing a grid efficient resource discovery. Their resource selection technique i.e Resource Ranking is based on Google's search technique PageRank. Google's PageRank is a numeric value that represents the importance of a page on the web. A page has a high rank if the sum of its backlinks is high. Similar idea can be applied to grid where submission of jobs to resources indicates the backlinks. Resource can obtain higher ResourceRank score, if many users from different organizations submit jobs to that resource or there exist users with high Resource Rank using the resource. They calculate the Resource Rank using the damping factor. After that they incorporates Resource Rank into the rank equation in Condor ClassAd. Hence, Resource Rank becomes a new constraint that must be considered when matchmaking is performed. Zhou and Lu et al[6]. proposed an algorithm of resource evaluation and selection based on Multi-QoS constraints applying multiple attribute decision making (MADM) and hierarchical analysis methods. It has a hierarchical and scalable QoS model. Three main matrix-Decision matrix, Normalized decision matrix, fuzzy preference matrix are used to store the information of resource selection and resource scheduling.

Resource Selection Algorithm

We have implemented a general-purpose resource selection algorithm. It accepts user resource requests and finds a set of resources with highest rank based on resource information from a Grid information service. An open interface allows users to customize the resource selector by specifying an application-specific mapping module. The Resource Selection Service (RSS) comprises three modules. The *resource monitor* for querying Monitoring and Discovery Service to obtain resource information and for caching this information in local memory. The *set matcher* uses the set-matching algorithm to match incoming application requests with the best set of available resources. The performance of some applications, such as Cactus, is tightly related to resource topology and workload allocation. The *mapper* decides the resource topology and the allocation of application workload to resources.

Algorithm

begin:

1. Input length of the job, minimum amount of RAM required, and the maximum job cost from the user.
2. Calculate the success rate of the machine.
3. Calculate availability of the machine.

Avail= MTBF/(MTBF+MTTR).
Where avail=availability of the machine.
MTBF=mean time between failure.
MTTR= mean time to repair.

4. Calculate the reliability of the machine.

Reliability= availability * Feedback * MTBF/3.

5. Calculate trust value(T.V.)

T.V.=fb*Rel*security*usage/4
Where fb=User feedback of the machine.
Rel=Reliability of the machine.
Security is the machine security.
And usage is calculated by calculating how many times the job is successfully submitted on resource.

6. Calculate machine performance.

Performance=cost*security*fb*avail*rel*tv/6.

7. Select the resource based on the rankingaccording to performance of the resource**.**

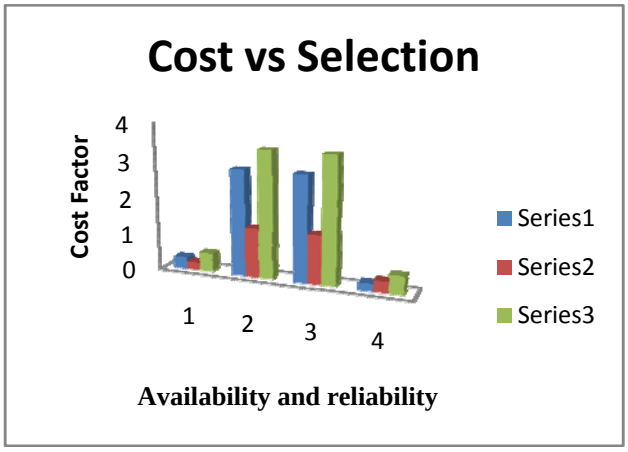
Experimental Study and Results

In this section, we first describe the experimental setup and present the analysis of our experimental results. The proposed algorithm is implemented in Java. The experimental setup consists of multiple Grid entities considered as separate resources. First the user enters the job requirements and the proposed system calculates the availability of the resource, trust value, reliability and the performance of the resource and rank them. User choose the best resource available based upon the performance of the resource. After the completion of job, the user is asked to provide feedback about the entity on some security attributes. The selected entity has provided high security for the job execution.

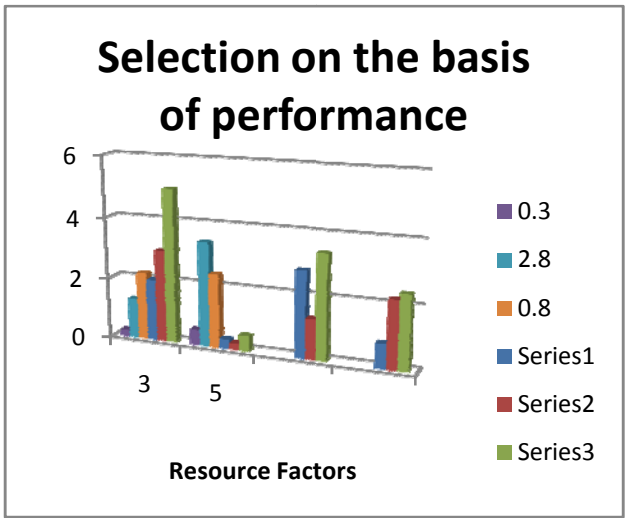
The work is simulated on a JAVA Based discrete-event Grid simulation toolkit for Grid environment called GridSim. This toolkit provides facilities for modeling and simulating Grid resources and Grid users with different capabilities and configuration. To simulate the resource selection mechanism in the GridSim environment requires the modeling and creation of GridSim resources and applications that model tasks.

Results

Machine	Cost	Usage Count	Availability	Reliability	Trust value	Final Perf.
2	200	0.1	0.307	2.8717	0.5	0.083333336
1	300	0.3	0.200	1.3333	0.1	2.2222223
3	500	0.1	0.5	3.4313	0.9	2.4444447



The above results shows the resource selection on the basis of cost factor only. But if we consider the Performance factor then the resource selection is efficient on the basis of cost, usage count, availability, reliability and many factors.



Conclusion

Besides that the Grid technology is getting more matured day by day, different policies for resource availability in the Grid, and resources' working stability raise serious issues about their suitability for different jobs. In this work we compare resources based traditional and new metrics- their MTBF, MTR in general In this paper, we have described a framework for calculating performance of a system in Grid environment. Various factors are considered for calculating the performance of the resource these factors are trust value, machine security, feedback of the user, success rate, reliability, availability of the resource. As future work we have decided to implement the advanced resource reservation for resource selection considering performance prediction as a main factor.

References

[1] D.P. Spooner_, S.A. Jarvis_, J. Caoy, S. Sainiz and G.R. Nudd, "Local Grid Scheduling Techniques using

Performance Prediction",

[2] Chuang Liu* Lingyun Yang* Ian Foster*# Dave Angulo* "Design and Evaluation of a Resource Selection Framework for Grid Applications", Proceedings of the 11 th IEEE International Symposium on High Performance Distributed Computing HPDC-11 2002 (HPDC'02).

[3] X. Ren, R. Eigenmann, "Empirical studies on the behavior of resource availability in fine-grained cycle sharing systems", Proceedings of 35th International Conference on Parallel Processing, Columbus, USA, August, 2006.

[4] C. Liu, L. Yang, I. Foster and D. Angulo, "Design and Evaluation of a Resource Selection Framework for Grid Applications", in Proceedings of HPDC-11, 2002.

[5] S.D. Kamvar, M.T. Schlosser and H. Garcia-Molina, "The Eigentrust Algorithm for Reputation Management in P2P Networks", in Proceedings of ACM WWW 2003.

[6] Jiantao Zhou, Min Yan, Xinming Ye, Haiyan Lu, "An Algorithm of Resource Evaluation and Selection Based on Multi-QoS Constraints", Seventh Web Information Systems and Applications Conference, 2010.

[7] Yao Wang and Julita Vassileva: Trust and Reputation Model in Peer-to-Peer Networks. In Proceedings of the 3rd IEEE International Conference on Peer-to-Peer Computing. Linkŏping: IEEE Computer Society (2003), 150–158.

[8] Shanshan Song and Kai Hwang, Dynamic Grid Security with Trust Integration and Optimized Resource Allocation, Internet and Grid Computing Laboratory, University of Southern California, Los Angeles, CA. 90089 USA.

[9] E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, P. Samarati and F. Violante, "A Reputation-Based Approach for Choosing Reliable Resources in Peer-to-Peer Networks", in Proceedings of ACM CCS 2002.

[10] *Noorisyam Hamid, Fazilah Haron and Chan Huah Yong* "Resource Discovery using PageRank Technique in Grid Environment" Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGRID'06)

[11] Farag Azzedin, Muthucumaru Maheswaran, "Towards Trust-Aware Resource Management in Grid Computing Systems, " ccgrid, p. 452, 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID'02), 2002.

[12] R. Buyya and S. Venugopal, The Gridbus Toolkit for Service Oriented Grid and Utility Computing: An Overview and Status Report, Proceedings of the First IEEE International Workshop on Grid Economics and Business Models (GECON), 2004.

Study of Reactive Solutions for Web Application Performance Enhancement during Overload

[1]Charulata S. Bonde and [2]Prof. A.A. Sawant

[1]M.Tech (Computer Engineering), [2]Professor Emeritus,
Department of Computer Engineering and IT, College of Engineering, Pune, Maharashtra, India
E-mail: bondcs09.comp@coep.ac.in, charu.bonde@gmail.com, aas.comp@coep.ac.in

Abstract
In today's world, almost all of the E-commerce websites or applications are available over the Internet. Such websites are often faced with incoming load of requests that exceeds their capacity, i.e. they are subjected to overload. Most existing web applications show severe drop in throughput and thus degraded performance at high overload.

There are multiple reactive solutions being suggested for such overload situations to enhance the performance of the web servers viz. Enhancing the caching ability at the web server, applying compressed caching for the underlying OS, applying optimum timeout for the TCP connections, effective overload handling etc.

The approach taken for this project is to enhance the web server performance by effective overload handling (overload detection, control etc)

This paper attempts to make a survey of the available suggestions / solutions about web server performance enhancement. It also comments on the advantages and limitations of these suggestions/solutions.

Index Terms: Web application; performance enhancement; reactive solutions;

Introduction

Overload

Cause for performance degradation Online services of today, such as banking, shopping, stock market trading are supported by Web-based multi-tiered server systems. Such services are exposed to variable load. The possible reasons for overload

Peak hour usage phenomenon

Events viz. sales, holiday shopping, or headline events. Peak load during such events can sometimes be orders of magnitude higher than average load, and can exceed the capacity of the system. When the incoming request rate on a system exceeds its processing capacity, it is said to be overloaded.

Overload consequences

Most server systems display unstable behavior when overloaded. Although ideally a system should operate at its maximum capacity when overloaded, many systems experience a drop in throughput (successful request completion rate), which is often drastic. The consequences of overload

- Increase in the response time
- Timing out of requests
- Abandoning the server system after being serviced for some amount of time.
- Request abandonments (either manual or protocol-triggered) lead to retries which further elevate the effective load on the servers.
- The overloaded server ends up being busy serving a large number of requests which timeout, resulting in a severe drop in throughput. This 'feedback phenomenon' further deteriorates the performance of the website.

Approaches for overload handling

The problem of overload could be partially eliminated by proper server capacity planning through server duplication, data redundancy and request redirection; it cannot be fully eliminated. Unexpected peak hour usage of a Website can always happen; e.g. due to a major breaking news event for a popular news Web-site, or server-side failures that reduce total capacity. Since it is not prudent to size server systems for such occasional overload situations, a mechanism is required which specifically aims to keep the server system operating in a stable manner, even in the presence of overload. Overload can be controlled using two broad approaches; Pro-active and Reactive.

1. **Pro-active approach:** In this approach, the control mechanism prevents the system from getting overloaded by exercising admission control for the incoming requests. A fair amount of knowledge of the system's capacity, a request's resource needs and monitoring of system resources is required to be able to make an accurate decision about request admission. Such complex mechanisms are best employed when user QoS requirements (viz. max response delay, max tolerable precision loss etc.) are exclusively expressed, and when the server system is required to be a QoS-aware system, that provides specific and differentiated performance guarantees.

Many of the existing approaches depend on the knowledge of the utilization of a resource or estimation of the resource demands of the resource which has been identified as the bottleneck. This seems reasonable if in a system the bottleneck remains unchanged with the varying workloads.

However, in a complex multi-tiered server system, determining which resource is the bottleneck resource can be very difficult.

Furthermore, the bottleneck resource itself may vary with changing workloads or software and hardware configurations.

2. **Reactive approach:** Most Web-sites aim for a simple "best-effort" service, where the users do not express any explicit QoS targets - thus the system goals are those of ensuring stability on overload, maintaining the throughput near capacity, and response times that do not result in a large number of request abandonments. Such systems can activate an appropriate overload control mechanism only upon overload detection - a reactive overload control. For a reactive approach, two components are required: an overload detection mechanism and an overload control mechanism.

A number of existing approaches [2, 3, 4] use overload detection mechanisms based on resource utilization. These mechanisms assume that the potential bottleneck resource is known and can be monitored i.e. high utilization of this resource can indicate an overload. However, system bottlenecks may not be known a-priori; they may vary based on

- Type of workload
- CPU intensive
- Network I/O intensive etc.

- Machine hardware configuration
- CPU speed
- Network bandwidth
- System cache and memory sizes etc.

- Software configuration
- Thread pool size
- Buffer size
- Object pool size etc.

Thus, determining the bottleneck resource is nearly impossible in the case of multi-tiered heterogeneous systems which support varying workload mixes. This motivates the need for an overload detection mechanism that does not require the knowledge of the bottleneck resource, and therefore does not need to monitor it.

Proactive Solutions

There are many solutions to enhance the performance of the web application in overloaded situations viz.

- Increasing the processing power i.e. CPU capacity or adding multiple servers
- load balancing of the servers
- Increasing memory size
- Increasing storage capacity
- Optimizing database in terms of data organization, number of connections, query optimization etc.
- Code optimization i.e. utilizing the resources properly, removing memory leakages etc.

All the above mentioned solutions deal with performance enhancement by hardware up gradation or software optimization. These solutions have some practical difficulties viz. Hardware upgrade increases the cost of implementation and software optimization can be time consuming or may not be possible in certain situations e.g. application is already implemented, running in production environment.

Reactive Solutions

To overcome these difficulties, numbers of solutions are being proposed for enhancing the web application performance in overloaded situations. These are the reactive solutions that could be applied on top of the existing application without disturbing it. Few significant solutions are mentioned below.

Proxy based Self-tuned Overload Control Mechanism

The paper [1] claims that an absolute indicator of a system in overload is when its throughput (rate of successful completion of requests) is lower than its request arrival rate. As long as requests arrive at a rate that the system can process them, the completion rate has to be close to the arrival rate. If the completion rate (smoothed and averaged, to ignore transient effects) drops below arrival rate, it is a clear indicator, that the server cannot process the requests at the rate they are arriving, and is hence, overloaded.

The paper [1], proposes a proxy-based, reactive overload control mechanism which uses the ratio of the throughput to the arrival rate as an indicator of overload. Overload is flagged by the proxy when this ratio is lower than 1 by a given amount (determined by a threshold value). On overload detection, the proxy uses a self-clocked admission control on incoming requests that are queued at the proxy. The request at the head of the queue is admitted into the server system, only when a request is seen successfully exiting from the server, indicating that there is room for a new request. The mechanism is similar in concept to window-based flow control mechanisms used in networking.

Thus, the mechanism is self-tuned, and requires no knowledge of the system hardware. It enables the server to operate at its maximum capacity while keeping response times within acceptable bounds even at very high overload.

Advantages:

- In overloaded situation, serves more number of requests with reduced response time than web server without proxy.
- Avoids web server crashing due to overload.
- Can be used for already deployed applications without changing a single line of code in the application.
- Web proxy is the only additional component needed.

Many web proxies are available as FOSS.

Limitations

- Does not guarantee that in overloaded situation, all the requests coming to the web server will be served with reduced response time.

Delayed caching at Web Cache Server

Mechanism

Web caching is an economic and efficient solution for the problem of degraded web server performance at the time of heavy load. Many of the existing solutions focus on increasing the caching capacity of the cache server instead of really enhancing the cache server throughput with available capacity. The paper [5] proposes delayed caching as a solution to enhance web server performance.

In case of overload, the caching proxies do not cache the data and just focus on serving the request. This means for every new request, response is provided but the caching of the response is avoided to save time and to serve more requests in available time. Because of this, when the server moves to normal condition after overload, the cache does not have latest served information and loading the cache with latest information when the request arrives; takes time. The delayed caching solution involves recording the meta-information of the requests at the time of overload and loading the cache with this information when the server moves back to normal condition.

Delayed caching can be seen as the method to improve the performance of the server in order to improve system reliability and provide a quick service to users' service requests. The module for delayed caching can be implemented inside the caching web proxies viz. SQUID

Advantages

- Caching reduces the time to fetch data from the repository.
- Delayed caching loads the cache after server exits from overloaded condition to normal condition and hence can serve new requests from cache itself.
- Web proxy is the only additional component needed. Many web proxies are available as FOSS.
- Can be used for already deployed applications without changing a single line of code in the application.

Limitations

- Maintaining the cache could be expensive especially in distributed scenario where web server(s) and web proxy are deployed on separate hosts.

Site based caching at the web cache server

Mechanism

The existing designs and solutions for web caching systems commonly make caching decisions based on document or uniform resource locator (URL) information. This site-based approach makes caching decisions based on the website that an object belongs to, rather than the object itself. The paper [6] shows that this new approach can benefit different scopes of cache design, ranging from internal operation of a single proxy (host level), mapping of proxy array in a local area network (LAN level), to load reduction in the global cache hierarchy (wireless area network (WAN) level). Since disks are usually the performance bottleneck in a proxy, to overcome this, site-based cache architecture is proposed that tries to store web objects belonging to the same site in nearby disk blocks. This new architecture reduces disk access time as compared to the conventional URL-based cache architecture. Besides, in the LAN-level design, a site-based mapping scheme can be used to map all requests targeting the same website to the same proxy, resulting in reduction in the total transmission control protocol connection overhead.

Advantages

- Caching reduces the time to fetch data from the repository.
- Reduces disk access time. Web proxy is the only additional component needed.
- Many web proxies are available as FOSS.
- Can be used for already deployed applications without changing a single line of code in the application.

Limitations

- Maintaining the cache could be expensive especially in distributed scenario where web server(s) and web proxy are deployed on separate hosts.

Main memory compression

Mechanism

Current web servers are highly multithreaded applications whose scalability benefits from the current multi-core/multiprocessor trend. However, some workloads cannot capitalize on this because their performance is limited by the available memory and/or the disk bandwidth, which prevents the server from taking advantage of the computing resources provided by the system. To solve this problem, paper [7] proposes the use of main memory compression techniques to increment the available memory and mitigate the disk bandwidth problem, allowing the web server to improve its use of CPU system resources.

This solution is proposed for Linux Operating system where the memory management subsystem is changed to use the compressed page cache (CPC). Compressed memory systems are based on the reservation of some physical memory to store compressed data, virtually increasing the amount of memory available to the applications. This extra memory reduces the number of accesses to the disk and allows the execution of applications with larger working sets without trashing.

Advantages

- Maximum data can be cached with the available physical memory.
- Faster as reduces disk accesses.
- Can be used for already deployed applications without changing a single line of code in the application.

Limitations:

- Deep knowledge about the operating system i.e. Linux is needed.

Active TCP Connection management and delay prediction

Mechanism

The paper [8] proposes a forward neural network model to predict the optimum TCP active connection timeout. This model and the calculations can be used for deciding the optimum session timeouts so that a web server can serve maximum requests in given time.

Active TCP connection deals with setting connection main-tenance time when a TCP connection is established or a request arrives. During this period of time, the connection is always effective, and can be used to serve the HTTP requests that follow. Whenever a new request arrives, the server resets this connection maintenance time, once the connection maintenance time expires, the server closes the connection.

Advantages

- The unused connection can be closed depending on the timeout period suggested / predicted by the model and new connections can be made to serve newly arriving requests.
- Can be used for already deployed applications without changing a single line of code in the application.

Limitations

- The connection timeout period suggested / predicted by the model does not guarantee correct value in all the situations.

Conclusion

This paper surveys various suggested solutions for web server performance enhancement during overload. Implementing any of these mechanisms can improve the performance of the web application without changing the code of the already implemented web application. Multiple mechanisms can also be implemented in parallel to give significant performance enhancement.

The following table summarizes the features of the above mentioned approaches:

Feature	Proxy based self-tuned overload control	Delayed caching at web cache server	Site-based caching at web cache server	Main memory compression	Active TCP connection management and delay prediction
Technique used for performance enhancement	Request admission control	Caching	Caching	Modification of memory page structure	Artificial intelligence
Component available as FOSS	Yes	Yes	Yes	Yes	Yes
Additional components needed	Web proxy	Web cache server	Web cache server	-	-
Software modules implemented or embedded	Load detection, overload control modules in web proxy	Load detection, delayed caching module in web cache server	Site based caching module in web cache server	Compressed page cache module in memory management subsystem of the Linux OS	Active TCP connection management module in web server
Is OS specific	No	No	No	Yes	No

References

[1] Rukma P. Verlekar, Varsha Apte, A Proxy-Based Self-tuned Overload Control for Multi-tiered Server Systems. HiPC 2007, LNCS 4873, 285-296, 2007.

[2] Chen, H., Mohapatra, P., Overload Control in Qos-Aware Web Servers. Computer Networks 42(1), 119-133, 2003.

[3] Cherkasova, L., Phaal, P., Session-Based Admission Control: A Mechanism for Peak Load Management of Commercial Web Sites. In: IEEE Transactions on Computers 51(6), 669-685, 2002.

[4] Elnikety, S., Nahum, E., Tracey, J., Zwaenepoel, W., A Method for Transparent Admission Control and Request Scheduling in E-commerce Web Sites. In: WWW 2004: Proceedings of the 13th International Conference on World Wide Web, 276-286, ACM Press, New York, 2004.

[5] Daesung Lee; Kim, K.J., A Study on Improving Web Cache Server Performance Using Delayed Caching. In: Information Science and Applications (ICISA), 2010 International Conference, 1-5, 2010.

[6] Wong, K.-Y.; Yeung, K. H., Alternative web caching design: a site-based approach. In: Communications, IET, 1504-1515, 2010.

[7] Beltran, V., Torres, J., Ayguade, E., Improving Web Server Performance through Main Memory Compression. In: 4th IEEE International Conference, 303-310, 2008.

[8] Guofang, Yu; Xianglu, Tan, QoS Control of Web Server Based on Active TCP Connection Management and Delay Prediction. In: Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference, 955-958, 2010. 06

Images Compression and Encryption Method using VLSI Implementation

Neeraj Shrivastava[1], Ashish Surywanshi[2]
Megha Shrivastava[3] and Pushpendra Sharma[4]

[1 & 2]IPS-IES College, Indore, India, [3]J.N.C.T., Bhopal, M.P., India, [4]ITM College, Bhilwara, India
E-mail: neeraj0209@gmail.com, ashishsurywanshi@gmail.com,
yachna_2008@yahoo.co.in, pushpendra0285@yahoo.co.in

Abstract
In this Paper, we describe fully pipelined single chip architecture for implementing a new simultaneous image compression and encryption method suitable for real-time applications. The proposed method exploits the DCT properties to achieve the compression and the encryption simultaneously. First, to realize the compression, 8-point DCT applied to several images are done. Second, contrary to traditional compression algorithms, only some special points of DCT outputs are multiplexed. For the encryption process, a random number is generated and added to some specific DCT coefficients. On the other hand, to enhance the material implementation of the proposed method, a special attention is given to the DCT algorithm. In fact, a new way to realize the compression based on DCT algorithm and to reduce, at the same time, the material requirements of the compression process is presented. Simulation results show a compression ratio higher than 65% and a PSNR about 28 dB. The proposed architecture can be implemented in FPGA to yield a throughput of 206 MS/s which allows the processing of more than 30 frames per second for 1024x1024 images.

Introduction
Reconfigurable hardware in the form of Field Programmable Gate Arrays (FPGAs) have been proposed to obtain high performance and economical price to implement image processing applications like face recognition, detector or airport security [8]. For these applications, we need to use communication systems with a good security level (encryption) and an acceptable transmission rate (compression rate). In the literature, several encryption and compression techniques can be found. However, for some applications such as detectors, the encryption and the compression techniques cannot be deployed independently and in a cascade manner without considering the impact of one technique over another [2]. To solve this problem, we developed a new technique to simultaneously compress and encrypt multiple images [3]. The main idea of our approach consists, firstly, in multiplexing the spectra of different transformed images (to be compressed and encrypted) by a Discrete Cosine Transform (DCT) and secondly in implementing the proposed system in FPGA. Consequently, special attention is given to the DCT algorithm implementation in the context of image compression. In fact, the DCT is the heart of the proposed compression and encryption method. It has been widely used in speech and image compression due to its good energy compaction [13]. However its computational requirement is a heavy burden in the real time simultaneous compression and encryption application. Different DCT architectures have been proposed to exploit signal proprieties to improve the tradeoff between computation time and hardware requirement. Among these, the DCT algorithm proposed by Loeffler [16], has opened a new area in digital signal processing by reducing the number of required multiplications to the theoretical limit. In this paper we use the DCT architecture for image compression and we demonstrate that the number of arithmetic operators can be reduced without dramatically decreasing the compressed image quality. In fact, by exploiting the spacial correlation of input images, we can reduce the number of arithmetic operators from 11 multipliers and 29 adders to 4 multipliers and 14 adders. Simultaneously, in order to perform the security level, a second stage a using random number generator is applied to some specific DCT outputs. This paper is organized as follows: the description of the proposed simultaneous compression and encryption method is presented in section II. Section III is dedicated to the optimization of the DCT architecture. Implementation results using FPGA are illustrated in the last section before conclusion.

Method Principle
We proposed a new technique, based on our methods presented in [3] and [5], which can carry out compression and simultaneous encryption using random number generator and Discrete Cosine Transform (DCT). The main idea of our approach consists in multiplexing the spectra of different transformed images separately by a DCT. The choice of the DCT is justified by the use of the DCT in many standards such as JPEG [14], MPEG [15] and ITU-T H261 [12]. Moreover, we need fewer DCT coefficients than DFT coefficients to get a good approximation to a typical signal [11]. In fact, by applying the DCT, the most of the signal information tends to be concentrated in a few low frequency Components. Consequently, the higher frequency coefficients are small in magnitude and can be ignored in the compression and encryption process. Fig. 1 presents the synoptic diagram of the proposed compression and encryption system. In the left side, 4 input gray level images are presented (P1, P2, P3, P4). To apply to each of these images a full parallel DCT algorithm,

we need to parallelize each image by blocks of 8 pixels. This operation can be done by a serial to parallel block composed by 8 flip-flops. Then, 4 DCT blocks are used to transform the 4 images. These DCTs are employed to regroup lower frequency components of the DCT.

Input

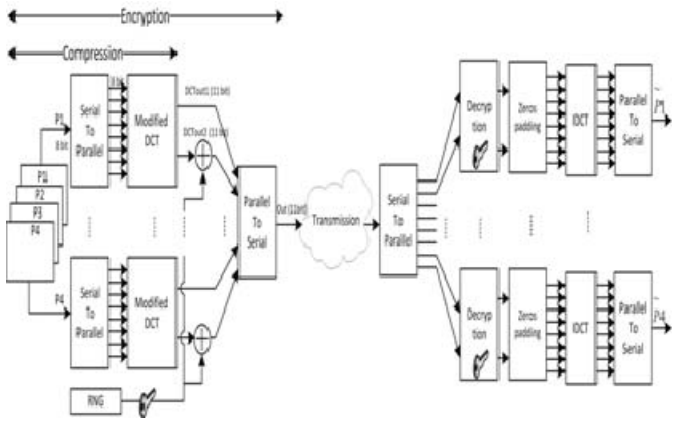


Figure 1: Synoptic diagram of the proposed compression and encryption system.

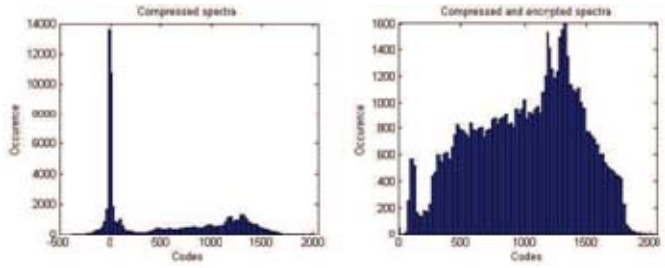


Figure 2: Histogram of DCT outputs.

In fact, by taking into account only the first and the second DCT outputs among 8, we get a good approximation of input pixels. Hereafter, we use the following notations: DCTout1 for the first DCT output and DCTout2 for the second one. Concerning the encryption process, the algorithm is based on the next observation: the multiplexed spectrum plane presents alternatively one high value DCTout1 followed by one low value DCTout2. In terms of security, we can imagine that behind the histogram on the left of Fig. 2 we can find DCT coefficients. In fact, as it will be explained in the next section, the low value of the DCTout2 is due to the spatial correlation between 8 successive pixels presented in input images. In order to ensure a good encryption level against any hacking attempt, we propose to add to DCTout2 a positive random value to have a data values close to DCTout1. As mentioned in Fig. 2, the addition of a random number can drastically modify the characteristic spectral distribution of the DCT. The security key will be sent separately as a private encryption key. Once secure and compressed information safely reach the authorized receiver, the image extraction can be easily done by reversing the various steps used in the whole process:

- Subtract the received image by the security key;
- Add 6 zeros to each block (zeros padding);
- Run an Inverse DCT (IDCT).

DCT Architecture

The DCT is the heart of the proposed compression and encryption method. Therefore, an optimization of the whole proposed method requires a DCT optimization. In this section, we present the modified DCT architecture and the data encoding of DCT outputs in order to allow an acceptable compression ratio and a relatively high image quality. A. Related Work The N-point DCT of N input samples x(0),...,x(N − 1) is defined as:

$$X(n) = \sqrt{\frac{2}{N}}\, C(n) \sum_{K=0}^{N-1} x(k) \cos\frac{(2k+1)n\pi}{2N}$$

where C (0) = 1/√2 and C (n) = 1 if n 6= 0.

In literature, many fast DCT algorithms are reported. In [17], the authors show that the theoretical lower limit of 8-point DCT algorithm is 11 multiplications. Since the number of multiplications of Loeffler's algorithm [16] reaches the theoretical limit, we use this algorithm as the reference to this work. A modified signal flow graph of 8 inputs 2 outputs DCT is presented in Fig. 3. We will explain the reasons of modifications in the next section. In [9] one realization based on Loeffler algorithm is shown. A low power design is obtained with this algorithm. In [10] use the recursive DCT algorithm and their design requires less area than conventional algorithms. The authors of [10] use Distributed Arithmetic (DA) multipliers and show that N-point DCT can be obtained by computing N N/2-point inner products instead of computing N N-point inner products. In [7], a new DA architecture called NEDA is proposed, aimed at reducing the cost metrics of power and area while maintaining high speed and accuracy in digital signal processing (DSP) applications. Mathematical analysis proves that DA can implement inner product of vectors in the form of two's complement numbers using only additions, followed by a small number of shifts at the final stage. Comparative studies show that NEDA outperforms widely used approaches such as multiply/accumulate (MAC) and DA in many aspects. In this paper, we will not optimize the arithmetic operators but we present a new algorithm based on Loeffler one and makes dependency between the compression ratio and the material complexity. Consequently, optimizations in [9], [10] or [7] can be used with the presented new algorithm.

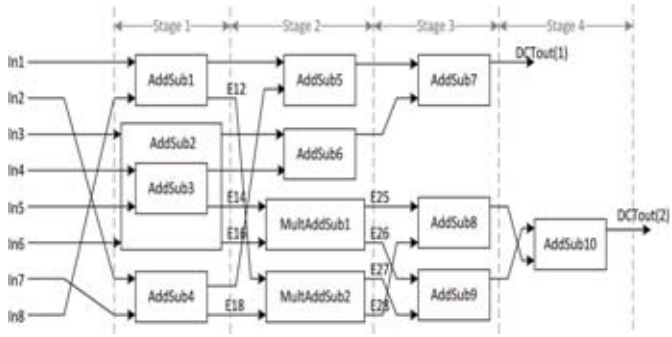


Figure 3: Signal flow graph of 8 inputs 2 outputs DCT.

Proposed DCT Architecture

The circuit of the proposed algorithm of the DCT is inspired from the Loeffler one. Therefore, some similarities exist between these two circuits. For example, we propose to compute DCT outputs on four stages as shown in Fig. 3. Each stage contains some arithmetic operations. The first stage is performed by 4 AddSub (adder and subtracter) blocks while the second one is composed of 2 AddSub and 2 MultAddSub (Multiplier, Adder and Subtracter) blocks. The details of these blocks are shown in Fig. 4. Moreover, some important differences should be mentioned. Since the proposed DCT circuit accepts 8 pixels per clock cycle and delivers 2 outputs against 8 outputs in the original Loeffler algorithm, we decide to change the DCT architecture to compute only necessary DCT coefficients DCTout1 and DCTout2. It should be outlined that traditionally, one possible manner for compression based on DCT algorithm consists in computing all DCT outputs (8 outputs for 8 pixels) and after that some special points are selected. The whole computation time and latency are therefore very high. The changes in DCT architecture are as follows: First, only necessary paths to compute DCTout1 and DCTout2 are kept as shown in the Fig. 3. Thus, we can economize 5 multipliers, 2 adders and 2 subtracter compared to the Loeffler architecture. Then, we can notice that in Fig. 3 only the first outputs of AddSub5 to AddSub10 are used. Therefore, the AddSub5 to AddSub10 blocks are reduced to 1 adder per block. Consequently, 6 additional subtracters can be saved. Finally, the DCTout2 can be written as follows:

$$\begin{aligned} DCTout2 &= (E25 + E28) + (E26 + E27) \\ &= (E18 * \cos(\prod/16) + E12 * \sin(\prod/16)) \\ &+ (-E16 * \sin(3\prod/16) + E14 * \cos(3\prod/16)) \\ &+ (-E18 * \sin(\prod/16) + E16 * \cos(\prod/16)) \\ &+ (E16 * \cos(3\prod/16) + E14 * \sin(3\prod/16)) \end{aligned} \quad [2]$$

After factorizations, DCTout2 can be written as follows:

$$\begin{aligned} DCT\ OUT2 &= E18 * \underbrace{(\cos(\pi/16) - \sin(\pi/16))}_{C1} \\ &+ E16 * \underbrace{(\cos(3\pi/16) - \sin(3\pi/16))}_{C2} \\ &+ E14 * \underbrace{(\cos(3\pi/16) + \sin(3\pi/16))}_{C3} \\ &+ E12 * \underbrace{(\cos(\pi/16) + \sin(\pi/16))}_{C4} \end{aligned} \quad [3]$$

According to these equations, the MultAddSub blocks of Fig. 3 can be replaced by more simple blocks. In fact, the original block requires 1 adder, 1 subtracter and 4 multipliers to compute the outputs. Loeffler reduces the number of arithmetic operators to 3 multipliers and 3 adders per block. In this work, as presented in Fig. 4 the MultAddSub block can be replaced by only two multipliers. Like this, we economize 6 adders and 2 multipliers. Using these three optimization levels, the proposed DCT architecture requires 4 multipliers and 14 adders to compute relevant and representative data outputs for image compression against 11 multipliers and 29 adders proposed by Loeffler.C. Data encoding The minimization of data length implies less computation, and consequently, lower power consumption and higher speed. On the other hand, truncating introduces errors at the outputs and degrades the PSNR (Peak Signal to Noise Ratio). Thus a trade-off between power and PSNR is made. In the input side of the proposed method, the pixels of input images are encoded using unsigned 8-bit values. In the output side, DCTout1 contains the major part of the information, so this value must be encoded by the maximum number of bits. DCTout1 results in 3 successive additions of input pixels. Consequently, and considering the carry of each addition, the DCTout1 is encoded by using 11 bits. For the constant ci, i ∈ [1, 4] of (3) we can employ the coefficients encoding used in [6] and detailed by the next equation:

$$\tilde{C}_i = round\left(C_i * 2^{8-1} - 1 / C_{i,max}\right) \quad [13]$$

For DCTout2 encodage, we can take into account the spatial correlation of images. In fact, we can suppose that for image size of 256*256 pixels or higher, the block of each 8 adjacent pixels of the same line are very correlated and have a very close value. Consequently, signals E12, E14, E16 and E18 from Fig. 3 which are the subtraction of input image pixels from In1 to In8 have a very low value. In the same way, the signals E25 to E28 also have a lower value compared to input pixel images. Consequently, we can limit the DCTout2 by $-FS \leq DCTout2 \leq FS$ where FS = 28 is the full scale of

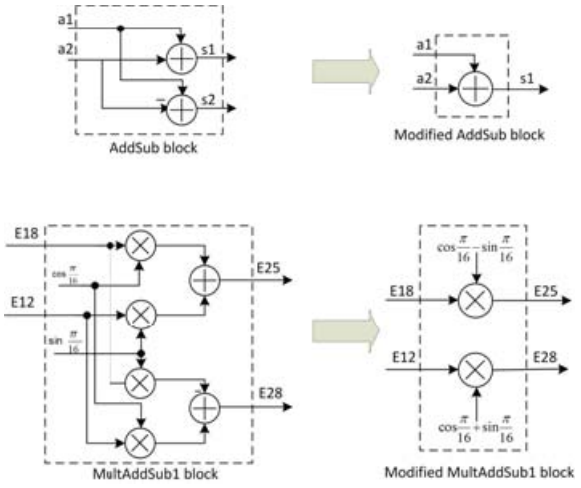


Figure 4: Arithmetic operator blocks.

the input images. On the other hand, for encryption process, the encrypted DCT out2 have to be close to DCTout1 which is in 0, 23 ∗ FS_. Consequently, the DCTout2 is added to ˜r, a generated random number expressed by the next equation:

$$\tilde{r} = \text{round}(FS + 6 * FS * r) \quad [16]$$

where r is an uniformly distributed pseudorandom number, $0 \leq r \leq 1$. Finally, DCTout1, DCTout2 are encoded using 11 bits. The data encoding allows a high compression ratio. In fact, since the 4 spectra will be regrouped in a single plane, the consequent compression ratio for input image sizes of S is:

$$R = \left(1 - \frac{S * 11 \text{ bits}}{4 * S * 8 \text{ bits}}\right) * 100\% = 65.62\% \quad [6]$$

Moreover, for higher compression rate, we can use the correlation between the neighboring pixels to encode the second DCT coefficient using only 7 bits. The obtained compression ratio can achieve a value of about 72%

Figure 5: Input images.

Validation

Methodology

A fixed point Matlab Simulink model has been established to validate the proposed method. This step is very important to validate the the algorithm structure before the material implementation. Concerning the description language, we decide to use VHDL rather than DIME-C and Mitrion-C which produce less efficient hardware design. In fact, DIME-C and Mitrion-C are much easier to program than VHDL, but visibility to hardware details allowing optimizations is lost due to abstraction [4]. In addition, the VHDL standard language gives the choice of implementing target devices (FPGA family, CPLD, ASIC) at the end of the implementation flow. It means that the models reported here are synthesized and may be implemented on arbitrary technologies [1]. Simulation results of the VHDL model are reported in Fig. 5 and Fig. 6 and show that original images are rebuilt correctly with a PSNR average between four images about 28 dB.

FPGA implementation

The original DCT Loeffler architecture and the proposed one in this article have been implemented in the same kind of FPGA boards, that is, Virtex 5 of xc5vlx330t. In order to illustrate the differences in hardware consumption, the FPGA implementation results are presented in Table 1. From this comparison we can notice that the proposed DCT architecture reduces the area consumption (slices and Look Up Tables, LUTs) at a rate higher than 50 %. Furthermore, the throughput, expressed in Millions of Samples per second (MS/s), presents a light increase compared to the Loeffler architecture. The throughput of 206 MS/s allows the processing of more than 30 frames per second. Finally, it should be pointed out that the modified DCT and the proposed compression and encryption method have the same throughput: the proposed method is for sure fully pipelined.

Figure 6: Output images.

Table I: Synthesis Results.

Characteristics	Loeffler	Modified DCT	Compression method
Slice registers	507	247	1536
Slice LUTs	1293	492	2058
Fully used LUT	316	162	955
Throughput (MS/s)	191.867	206.423	206.423

Conclusion

In this manuscript, a new method of simultaneous compression and encryption based on a DCT transformation is presented. An optimized DCT algorithm is proposed to reduce real time application requirements. This algorithm needs only 4 multiplications to compute relevant DCT output data. The FPGA implementation of the whole method shows improvements in terms of throughput, area and power consumption. To prove the good performances, the proposed algorithm is compared favorably with several existing methods.

References

[1] M. Jridi and A. AlFalou , Direct digital frequency synthesizer with CORDIC algorithm and Taylor series approximation for digital receivers, European Journal of Scientific Research, Vol. 30, No. 4, pp. 542-553, August 2009.

[2] A. Alfalou and C. Brosseau, Image Optical Compression and Encryption Methods, OSA: Advances in Optics and Photonics, vol 1, pp. 589-636,2009.

[3] A. Alfalou, M. Elbouz, M. Jridi and A. Loussert, A new simultaneous compression and encryption method for images suitable to optical correlation, Optics and Photonics for Counterterrorism and Crime Fighting V, edited by Colin Lewis, Proc. of SPIE Vol. 7486, 74860J-1-8, 2009.

[4] S. H. Park, D. R. Shires and B. J. Henz, Coprocessor computing with FPGA and GPU, 3rd ed. DoD HPCMP Users Group Conference. Seattle, WA, pp. 366-370, July 2008.

[5] A. Loussert, A. Alfalou, R. El Sawda, and A. Alkholidi, Enhanced System for image's compression and encryption by addition of biometric characteristics, International Journal on Software Engineering and Applications, pp. 111-118, 2008.

[6] E. Darakis and J.J. Soraghan, Reconstruction domain compression of phase-shifting digital holograms, Journal of Applied Optics, Vol. 46, pp. 351-356, January 2007.

[7] A. Shams , A. Chidanandan, W. Pan and M.A Bayoumi, NEDA : A low-power high-performance DCT architecture, IEEE transactions on signal processing, Vol. 54, No.3, pp, 955-964, 2006.

[8] M.W. James, An Evaluation of the Suitability of FPGAs for Embedded Vision Systems, CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, pp. 131-137, 2005.

[9] C.Y Pai, W.E. Lynch and A.J. Al-Khalili, Low-Power data-dependant 8x8 DCT/IDCT for video compression, IEE, Proceedings. Vision, Image and Signal Processing, Vol. 150, pp. 245-254, August 2003.

[10] S. Yu and E.E. Swartzlander Jr, DCT implementation with distributed arithmetic, IEEE Transactions on Computers, Vol. 50, No.9, pp, 985-991, September 2001.

[11] K. F Blinn, What's the deal with the DCT?, IEEE Computer Graphics and Applications, pp. 78-83, July 1993.

[12] ISO/IEC DIS 10 918-1, Digital compression and coding of continuous- tone still image, 1992.

[13] K. Rao and P. Yip, Discrete cosine transform algorithms advantages applications, Academic Press, New York, 1990.

[14] ISO/IEC JTC1/SC2/WG8, JPEG-8-R8, JPEG technical specification,1990.

[15] ISO/IEC JTC1/SC2/WG11, MPEG 90/176, Coding of moving picture and associated audio, 1990.

[16] C. Loeffler and A. Lightenberg and G.S. Moschytz , Practical fast 1-D DCT algorithm with 11 multiplication, IEEE, ICAPSS, pp. 988-991, May 1989.

[17] P. Duhamel and H. H'mida, New 2n DCT algorithm suitable for VLSI implementation, IEEE, ICAPSS, pp. 1805-1808, November 1987.

Clock-less Design of Reconfigurable Floating Point Multiplier

Yogesh Kumar and R.K. Sharma

ECE Department, NIT Kurukshetra, Kurukshetra, Haryana, India
E-mail: [1]yogi.yogeshkumar@gmail.com,[2]mail2drrks@gmail.com

Abstract
Floating point multiplication is a common function in signal, image processing, filters and real time data processing. This paper presents a new approach to design a high speed multiplier with a reconfigurable path for IEEE 754 single precision and two half precision numbers in parallel. The design is simulated on ModelSim and synthesized on XST of Xilinx ISE tool and implemented on board vertex 2pro FPGA. The multiplier is designed in two units Multiply - Add unit and Aligner - normalizing unit. This design can work up to 362.8 MHz while tested on Vertex 2pro FPGA.

Keywords: Asynchronous, Reconfigurable, Floating point, FPGA, pipelining and parallel Architecture.

Introduction

The use of FPGAs increasing day by day due to advancement in technology and fabrication process. The density of programmable block has increased as result large number of functional units can synthesized on a single board. The use of floating point operation is dominant in the field of digital signal processing, image processing, multimedia and real time data extraction. So, floating point multiplier had become a common element in the units of common applications, this element is also responsible for the accuracy and speed of the unit as it use normalizing process in the final result computation. The proposed design is divided into two units Multiply - add unit and Aligner - Normalizing unit.

To meet the requirement of high speed of computation in the first unit the pipelined architecture is used, where partial results are computed in steps. Mantissa of the number is divided into small units to achieve high speed and fully utilize the hardware in reconfigurable approach; the results are then added with proper shifting to get the final result of multiplication [1], for exponent part proper modification is done as per the shift in mantissa part. In the second unit we are using parallel architecture is used to achieve high performance. The design is based on leading one detection; this task is more time consuming so it is done with parallel processing approach. Multiplication result is checked in small groups to get leading one then the group is further checked for the final position of 'One' there after the mantissa part is shifted and proper correction is made to the exponent part.

To attain energy efficiency asynchronous approach [6] is used to operate the unit in parts, after calculating the product result multiplication unit will give a request signal to the normalizing unit and that is acknowledged by asserting the acknowledge signal after that the multiplication unit will compute the next product of incoming operands. The final acknowledgment signal is given by the normalizing unit and starting request signal is given to the multiplying - add unit from the external block. This makes the circuit to be active while computing the results only and after that the block will be shut down due to low request signal. Mular C-element is used to control the working of asynchronous design signals i.e. request and acknowledgement signals between the functional units [7] [8].

Verilog HDL is used to design the modules, Xilinx ISE tool is used to synthesize it and model sim is used for the simulation. The design is in HDL and can be optimized for any FPGA platform. The results are then compared with the previously proposed synchronous designs.

The rest paper in organized as: Section 2 explain IEEE 754 standard for floating point representation [5]. Section 3 and 4 explain Muller C-element and floating point multiplication respectively. Section 5 and 6 describe the proposed design and its implementation and results followed by conclusion in section 7.

Floating Point ieee 754 Standard

Floating point is refers to that radix point (decimal point or binary point) which can 'float' i.e. it can placed with respect to significant digits of number and its position is separately represented internally. The advantage of floating point representation over fixed point number is that it allows a wider range of values to be represented with the same number of bits that will not be allowed by fixed point number. This range is attained at the expense of precision of the value.

IEEE 754 Floating Point Representation

These standards gives the way that how to represent a floating point number and the way to carry out arithmetic operation on the operands. The mantissa is having a implied '1' as MSB for all the precisions in case the number is normalized. For subnorm or de-normalized numbers [4] it is zero, this MSB is not shown in the mantissa representation and rest of the following bits are shown in the mantissa part of floating point representation.

In this representation a decimal number 'n' is represented using a sign bit 's' a biased exponent part 'e' biased with excess-127 code(for single precision) and a mantissa part 'm' where the decimal number is given by

$$n = (-1)^s \times m \times 2^{e\text{-bias}(127)} \quad (1)$$

where $0 \le m < 2$

$0 \le m < 1$ for subnorm or de-normalized number

$1 \le m < 2$ for normalized number

'S' is a single bit represent the sign of the number

'S' = 0 for positive number and

= 1 for negative number

'e' i.e. exponent part is represented in exess-127 code format (for single precision) to compare exponents easily for arithmetic operations.

IEEE 754 standards

Floating point is represented by using different number of bits in different precisions is shown in table-1, given below:

Table 1: IEEE description of Floating point Number.

Precision↓	Sign	Exponent	Mantissa	Total	Exponent bias
Half	1	5	10	16	15
Single	1	8	23	32	127
Double	1	11	52	64	1023
Quad	1	15	112	128	16383

Muller c-Element

For asynchronous processing the request signal and acknowledgement signal will plays a significant part as the successive unit will show it's working by a acknowledgement signal and the preceding unit will give the operands by asserting the request signal. To give a request signal the unit must have to complete its work by that time so that it can go for further computation which this is indicated by the acknowledgment signal, so before giving request signal to the second unit, first unit have to wait for the acknowledgment signal. In case of 4-phase hand shake protocol we can Muller C-element [9] can be used. The behaviour of Muller C-Element is shown in table-2.

Table 2: Truth table for Muller C-element.

A	B	Out
0	0	0
0	1	No change
1	0	No change
1	1	1

So the request signal is only asserted when both the request signal from first unit and acknowledgment signal from second unit is asserted, and similarly the acknowledgment signal will be propagated.

Floating point Multiplication

The multiplication of floating point involves multiplication of mantissa part with addition of exponent part and sign determination of the final product, then normalizing the mantissa product and proper modification to the exponent part. The product 'n (s, e, m)' will be calculated from two operands $n_1(s_1, e_1, m_1)$ and $n_2(s_2, e_2, m_2)$ as:

$$s = s_1 \text{ xor } s_2 ; \quad (2)$$

$$e = e_1 + e_2; \quad (3)$$

$$m = m_1 \times m_2; \quad (4)$$

Proposed Design

The block diagram of the proposed design is shown in figure 1. The design is a reconfigurable approach for single precision floating point multiplier that can be reconfigured to get a single precision multiplication or two half precision multiplication at a time and for other applications which are not using IEEE standard the output can be truncated to 8-bit and 16-bit mantissa part while exponent part is fix i.e. 5-bit and 8-bit for two half and single precision number respectively. This 8 and 16-bit output can be used in image processing blocks as they have the standards of using byte dependant computing. This reconfigurable approach is controlled by a 4-bit 'select' input port and depending on its value the final path of operands is determined inside the design. The output bits are decided by value given on select line, shown in table-3, given below:

Table 3: Modes of selection of output Mantissa bits and Precision.

Select	Input operands	Output mantissa bit
0000	Two half precisions	Full 10-bits of half precision
0001	Two half precision	8-bit of half precision
1000	One single precision	Full 23-bits of single precision
1001	One single precision	8-bits of single precision
1010	One single precision	16-bits of single precision

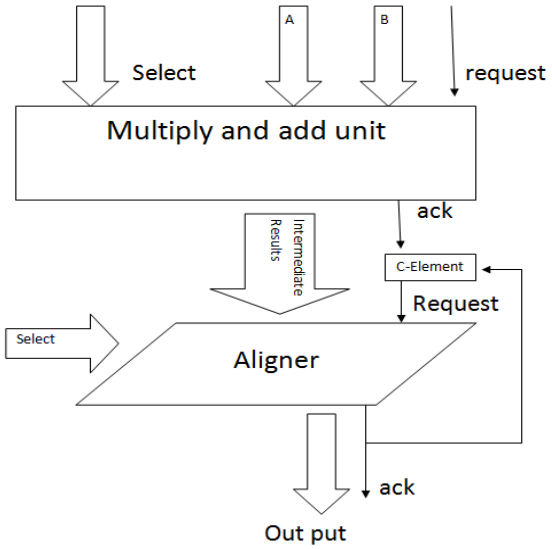


Figure 1: Block diagram of Proposed design.

Multiplier Unit

The multiplier unit is designed by using four stage pipelining structure. The operands are divided into three parts of 8-bit and then the multiplication is done to get three partial products that are then manipulated to get the final result of the product terms [9]. The pipelined structure increase the operating frequency of the unit to compete with the parallel process unit i.e. normalizing unit.

Add Exponent

The exponent of each computation is added in parallel to the multiplication by using the respective excess-127 and excess-15 code arithmetic operations depending on single precision and half precision respectively. Form the result of each result an extra excess term must be detected to get the corresponding correct result i.e. 127 and 15 for respective precisions.

Muller C-element

The multiplication unit after computing the result will assert a request signal for normalizing unit this element will prevent the mismatching between these unit and multiplication unit will not interfere with the normalizing unit in case it done fast computation. It will assert the request signal for normalizing unit after seeing the acknowledgement signal from its side. So, the request signal to the normalizing unit can be asserted only when the both signals are present i.e. request form multiplication unit and acknowledgement from the normalizing side. This unit will make the functioning of 4-phase handshake protocol.

Normalizing unit

This unit use parallel processing to get faster computation. The leading one detection is done by firstly checking each nibble i.e. the 48-bit result of multiplication is 12 parts for checking 'one' in it then real position is determined by further checking the nibble. Then depending on this one position the exponent part is modified; in parallel to it rounding is done by checking the nibble 23-bit position in the right of this leading one. Thus the whole process is done with a much improvement in the speed of processing. The 48-bit number 'a' is divided in 12 nibbles i.e. from 'a_1, a_2,.....a_{12}'

These are checked in parallel for rounding and normalizing so the separate attention is not needed for these and hence the target of high speed is achieved.

This leading one detection will enable the unit to compute the subnorm number computation as these are not have MSB equals to 'one'.

The output result of simulation is shown in figure 2. First is the request signal then four port of 16-bit each, then select line to make the reconfigured path followed by acknowledgment and error signal respectively, then the results of product after simulation.

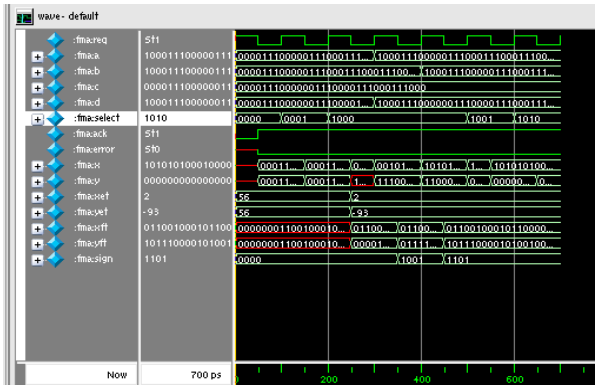


Figure 2: Simulation waveform of the design on ModelSim.

Implementation and Results

The design is implemented in Verilog HDL in top-down approach using Xilinx ISE tool for synthesis and simulation is done using ModelSim simulator. The results are verified on vertex 2p FPGA board. The simulation output is shown in figure 2.

The proposed design can work to maximum frequency of 360.878 MHz, with number of slices used is 541 and 792 4-input LUTs are used. The worst case combinational delay is 19.31 ns out of which 68% is logical delay and 32% is routing delay, with latency of 8 cycles. The results are compared with single precision non-reconfigurable design in the following section.

Comparison of results

The results for vertex 2p are compared with the other designs [1] shows 27% improvement in operating frequency, 48% improvement in area used on FPGA, 62% improvement in latency as compared below:

Table 4: Comparison of proposed design with reference [1].

Parameter↓	Reference [1]	Proposed design
Freq. (MHz)	284	362.8
Area (slices)	1049	541
Latency	23	8

The comparison of result of proposed design for vertex E with design [2] gives 36% improvement in area used by the design this improvement comes with an increase in delay, so this design gives a better solution where delay can be tolerated by occupying less area on board and multi precision floating point multiplication is required. In comparison with design [3] the proposed design shows 73% improvement in area requirement and 4 % improvement in delay. The results are compared below:

Table 5: Comparison of proposed design with [2],[3].

Parameter↓	Reference [2]	Reference [3]	Proposed design
Area (slices)	1331	3149	850
Delay (ns)	22.859	41.203	39.1

Conclusion

The proposed design is better in solution for multi precision multiplication of floating point number. The proposed design can work up to 362.8 MHz with an overall improvement in the area occupied on FPGA.

References

[1] Lamiaa S. A. Hamid, Khaled A. Shehata, Hassan El-Ghitani, Mohamed ElSaid, "Design of generic Floating Point Multiplier and Adder/Subtractor Units," UKSIM 2010, pp. 117.

[2] Saroja. V Siddamal, R. M Banakar and B. C. Jinaga, "Design of High-Speed Floating point Multiplier, " DELTA 2008, pp. 19.

[3] Himanshu Thapliyal and M. B. srinavas, "A Novel Time-Area-Power Efficient Single Precision Floating Point Multiplier," Proceeding MAPLD 2005.

[4] G. Govindu, L. Zhuo, S. Choi and V. Prasanna, "analysis of high performance floating point arithmetic on FPGAs," in Proceeding of the 18th International Parallel and Distributed Processing Symposium, April 2004, pp. 149-156.

[5] IEEE standard board, IEEE standard for floating-point arithmetic, 2008.

[6] Zain-ul-Abdin and B. Svensson, "Evolution in Architecture and Programming Methodologies of Reconfigurable Computing," Microprocessors and Microsystems, 2008.

[7] J. Teifel and R. Manohar, "An Asynchronous Dataflow FPGA Architecture," IEEE Transaction on Computers, vol. 53, no. 11, pp. 1376-1392, 2004.

[8] C. G. Wong, A. J. Martin, and P. Thomas, "An Architecture for Asynchronous FPGAs," In IEEE International Conference on Field-Programmable Technology, 2003.

[9] E. Rodriguez- Villegas, G. Huertas, M.J. Avedillo, J.M. Quintana, and A. Rueda, IEEE Trans, on Circuits and Systems-II, Special Issue on "Floating gate Circuit and Systems", Vol. 48, Issue 1, January 2001, pp. 102-106.

[10] The IEEE website. [Online]. Available: http://www.ieee.org/

Disease Detection using Analysis of Voice Parameters

Sonu and R.K. Sharma

Department of Electronics and Communication
National Institute of Technology Kurukshetra, Haryana, India
E-mail: sonu_27jan@yahoo.co.in, mail2drrks@gmail.com

Abstract
This paper investigates the adaptation of automatic speech recognition to disease detection by analyzing the voice parameters. The analysis of the voice allows the identification of the diseases which affect the vocal apparatus and currently is carried out from an expert doctor through methods based on the auditory analysis. This paper presents a novel method to keep track of patient's pathology: Easy to use, fast, non invasive for the patient and affordable for the clinician. This method uses parametric method (jitter, shimmer, harmonic to noise etc…) to evaluate the pathological voice. The method for this task also relies on Mel Frequency Cepstral Coefficient (MFCC) as feature extraction and Dynamic Time Warping (DTW) as feature Matching. The aim of the study is to evaluate the voice quality in patients with mild-to-acute asthma by parametric method and non parametric method. Comparative analysis is also done between parametric and non-parametric methods.

Keywords: MFCC, DTW, Dysphonia, Jitter, shimmer, HNR, Acoustic parameters

Introduction

Asthma as a chronic inflammatory disorder of the airways associated with increased airway hyper-responsiveness, recurrent episodes of wheezing, breathlessness, chest tightness, and coughing, particularly at night/early morning.Airway inflammation caused by allergies and asthma can hurt the sound quality of the voice. The vocal cords cover the larynx, the top part of the trachea. These mucus-covered muscular bands are the vibrating "strings" that produce voice sound, which is then filtered and shaped by the resonating cavities of the throat, nose, and mouth. Inflammation along the passageways from the nose down to the larynx can impair vocal quality. Bronchial asthma, labored breathing and wheezing, and allergies can also cause sore throat and inflammation around the vocal cords. Swollen, inflamed cords don't vibrate efficiently and can make the voice sound hoarse or scratchy [2]. Nearly half of the patients complain about permanent voice disorders. Any modification of above system may cause a qualitative and/or quantitative alteration of the voice, defined as dysphonia. Dysphonia can be due to both organic factors (organic dysphonia) and other factors (dysfunctional dysphonia)[7]. Spectral "noise" is strictly linked to air flow turbulences in the vocal tract, mainly due to irregular vocal folds vibration and/or closure, causing dysphonia. Such symptom requires a set of endoscopic analysis (by using videolaryngoscope, VLS) for accurate analysis [3]. However, clinical experience has pointed out that dysphonia is often underestimated by patients and, sometimes, even by family doctors. But early detection of dysphonia is of basic importance for pathology recovering. Several methods for assessing speech pathologies have been introduced. In general, objective speech quality measures are usually evaluated in the spectral, time or cepstral domains. In the spectral analysis methods, researchers have tried to keep track of the spectral variations of signal such as amplitude, bandwidth and frequency of formants including sub-band processing methods. In time domain, method based on temporal measurements of signal and their statistics, such as average pitch variation, jitter, shimmer, etc. to distinguish between normal and pathological speech is used. Moreover, Speech processing based on cepstral analysis has proved to be an excellent tool for voice disorder detection. In this paper, we investigate both time domain methods and the adaptation of Automatic Speaker Recognition for dysphonic voice assessment. Mel-frequency cepstral coefficients (MFCC) have traditionally been used in speaker identification applications. In this paper we have used MFCC for the feature extraction from the speech signals provided in the database and dynamic time warping (DTW) is used for feature matching in order to discriminate non-asthmatic persons from the asthmatic's patients. This paper is organized as follows: the speech and subject database and methods for feature extraction and feature analysis are described in section II. The results are presented in Section III and conclusion in Section IV.

Materials and Methods

Speech and subject database:

For this analysis, the speech record database consisted of a sustained phonation of the vowel /a/.The asthmatic group consisted of 21 patients with asthma's disease, 13 males (aged between 26 and 82 years, mean of 51.923 years) and eight females (aged between 46 and 62 years, mean of 53.4 years) and duration of the disease from 1 month to 30 years, with an average of 9.5 years. The control group was composed of 21 individuals non-asthma, four males (aged from 17 to 45 years, mean of 31 years) and twelve females (aged from 20 to 72 years, mean of 45.9 years) and five children (aged from 6 to 10, mean of 8 years).Acoustic assessment was performed by analysis of vowels phonated in isolation and in a constant linguistic test starting with the following words :(" hum sab ek

hein"). Acoustic analysis was performed with PRAAT Software programme. The following parameters were analyzed: F0, F1, F2, F3 Formants frequency levels, degree of voice break in isolated vowels, constant, fundamental frequency, Fo (Hz), Jitter (frequency perturbation – local, %), Shimmer (amplitude perturbation –local, %), Harmonic to noise ratio (HNR – dB), Intensity(dB).

Acoustic Parameters:

Jitter (local): This is the average absolute difference between consecutive periods, divided by the average period. MDVP calls this parameter *Jitt,* and gives 1.040% as a threshold for pathology.

Shimmer (local): This is the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude. MDVP calls this parameter *Shim*, and gives 3.810% as a threshold for pathology.

Harmonics-to-Noise Ratio (HNR) : A Harmonicity object represents the degree of acoustic periodicity, also called Harmonics-to-Noise Ratio (HNR). Harmonicity is expressed in dB. Harmonicity can be used as a measure for voice quality. For instance, a healthy speaker can produce a sustained a or i with a harmonicity of around 20 dB, and an u at around 40 dB; Hoarse speakers will have an a with a harmonicity much lower than 20 dB.

Degree of Voice Breaks DVB /%/ [15] - the ratio of the total length of areas representing voice breaks to the time of the complete voiced sample; and number of voice breaks NVB. The criteria for voice break area can be a missing impulse for the current period or an extreme irregularity of the pitch period.

Formant Frequency Measures: Frequency component amplified by resonator (vocal tract) and acoustic properties that distinguish speech sounds. Typically measured by LPC (Linear Predictive Coding)or spectrographic analysis. F_1 related to tongue height; F_2 related to tongue advancement. F_2 transition: change in frequency value of formant over time; reflects change in position of articulators.

Mel Frequency Cepstral Coefficient :

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behaviour. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency [8-10]. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech.

The speech input is recorded at a sampling rate of 16000Hz.This sampling frequency is chosen to minimize the effects of aliasing in the analog-to-digital conversion process.

The MFCC processor consists of seven computational steps is shown in the figure1. Each step has its function and mathematical approaches as discussed briefly in the following:

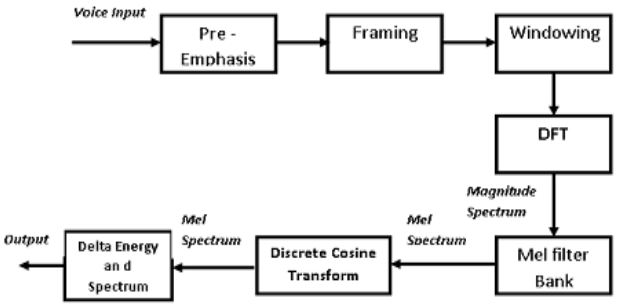


Figure 1: Block diagram of MFCC Processor.

Step 1: Pre–emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.[8]

Step 2: Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 512.

Step 3: Hamming windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines.

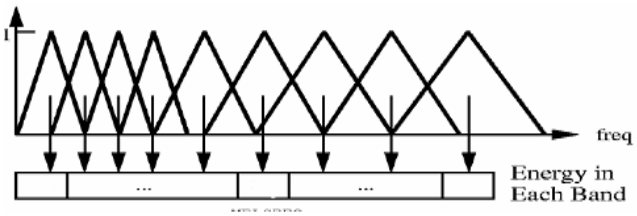


Figure 2: Mel scale filter bank from (young et al 1997).

This figure shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [7, 8]. Then, each filter output is the sum of its filtered spectral components.

Step 4: Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT).The result of the conversion is called Mel Frequency Cepstrum Coefficient.

Step 5: Delta Energy and Delta Spectrum
The voice signal and the frames changes, such as the slope of formant at its transitions. Therefore, there is a need to add feature related to the change in cepstral features over time. 13 delta velocity features (12 cepstral features plus energy), and 39 features double delta or acceleration feature are added.

Dynamic time warping
DTW algorithm is based on Dynamic Programming techniques as describes in [11]. This algorithm is for measuring similarity between two time series which may vary in time or speed. This technique also used to find the optimal alignment between two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series. Figure 4 shows the example of how one times series is 'warped' to another [12].

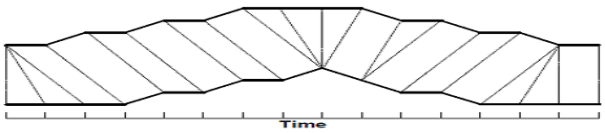


Figure 3: A Warping between two time series[12].

Suppose we have two time series *Q* and *C*, of length *n* and *m* respectively, where:

$Q = q1, q2, \ldots, qi, \ldots, qn \ldots.$ (1)

$C = c1, c2, \ldots, cj, \ldots, cm \ldots..$ (2)

To align two sequences using DTW, an *n*-by-*m* matrix where the (*i*th, jth) element of the matrix contains the distance *d* (*q*i, *c*j) between the two points *q*i and *c*j is constructed. Then, the absolute distance between the values of two sequences is calculated using the Euclidean distance computation:

$d(qi, cj) = (qi - cj)^2 \ldots.$ (3)

Each matrix element (*i*, *j*) corresponds to the alignment between the points *q*i and *c*j. Then, accumulated distance is measured by:

$D(i, j) = \min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i, j)$

Using dynamic programming techniques, the search for the minimum distance path can be done in polynomial time P(t), using equation below:

$P(t) = o(N^2V)$ (4)

where, N is the length of the sequence, and V is the number of templates to be considered[8].

Methodology

1. Database of 21 asthmatic patients (13 Male, 8female) undergoing treatment in the military hospital and 21 non-asthmatic person, is collected.
2. Vowels uttered by each person are extracted from the sentence with sampling frequency 16KHz, Mono, 8bit PCM.
3. Programming is done in Matlab to calculate Mel Frequency Cepstral Coefficient (MFCC) as feature extraction and dynamic time warping as feature matching.
4. Acoustic parameters such as fundamental frequency, jitter, shimmer, harmonic to noise ratio, formant frequency, intensity are extracted using PRAAT software.
5. Voice features of asthmatic patient and non-asthmatic persons are compared.

Algorithm for Proposed Architecture
An analysis of acoustic feature of asthmatic patient and an attempt to relate the variation in the voice characteristics of asthmatic. Recognition experiments is done by database of asthmatic patients recorded from the Military Hospital. Block Diagram illustrated in figure-4 describes the speech processing step to diagnose asthmatic patients

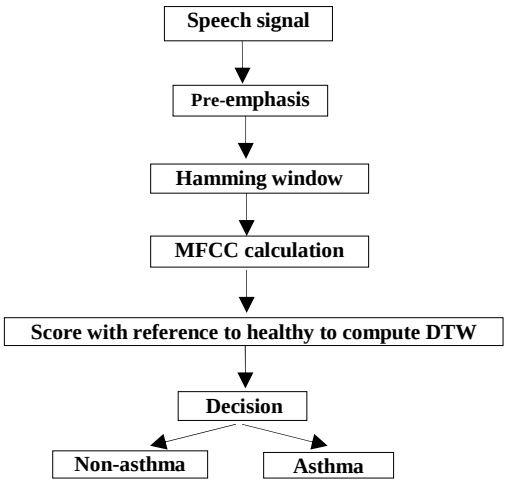


Figure 4: Algorithm for DTW score calculation.

Results
We recorded 21 phonation uttered by the asthmatic patients and 21 phonation by non-asthmatic group. For the acoustic analysis all 21 phonation of vowels is considered and for programming part 16 asthmatic and 16 non asthmatics voices are taken. Table 1 and Table 2 shows the results obtained by acoustic analysis using Praat software. Figure-5 shows the DTW Scores calculated and plotted.

Table 1: Acoustic Parameters of Asthmatic Group

SNo.	Parameters	Asthmatic group				
		min.	max	mean	overall result/success	
1	Pitch(Hz)	97.44	213	161	slighty higher	
2	Std. Devi(Hz)	4.5	53.67	24.8	higher	
3	UVF	15.09	70.11	29	mix value	
4	DVB(%)	5.14	62.75	31.1	higher	
5	Intensity(db)	67.34	90.51	80.4	lower	
6	Shim(db) local	4.912	17.94	8.44	<=3.8	85.71%
7	Jitter(db) local	0.83	4.27	1.73	<=1.04	95.23%
8	H/N ratio(db)	8.26	19.3	14.4	< 20	
9	Formant1	329	846	604	higher	
10	Formant 2	1697	2250	1972	lower	
11	Formant3	2503	3276	2967	mix value	

(CALCULATIONS BY PRAAT)

Table 2: Acoustic Parameter of Non-Asthmatic Group.

SNo.	Parameters	Non-asthmatic group				
		min.	max	mean	Overall result /success	
1	Pitch(Hz)	106	296	197	lower	
2	Std. Devi(Hz)	0.68	10.3	5.54	lower	
3	UVF	11.7	54.2	30.6	mix value	
4	DVB(%)	6.7	55.5	26.4	lower	
5	Intensity(db)	70.8	89.2	86.9	higher	
6	Shim(db)local	1.72	6.76	3.7	>3.8	71.28%
7	Jitter(db)local	0.12	1.53	0.62	>1.04	95.23%
8	H/N ratio(db)	12.5	21.8	17.6	nearly 20	
9	Formant1	363	857	458	lower	
10	Formant 2	1692	2771	2309	higher	
11	Formant3	1992	3369	2966	mix value	

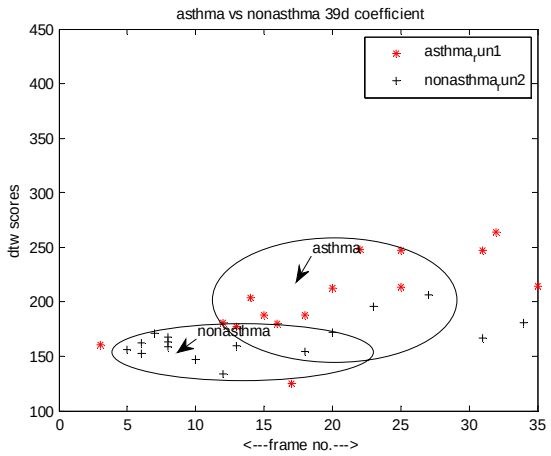


Figure 5: Comparison of asthma and non-asthma group (Vowel 'a' extracted from continuous speech).

Table 3: Rate of Classification From Both Groups

Method	Asthma	Nonasthma	Rate of classification
Acoustics analysis (using praat)	21	21	85%
MFCC/DTW	16	16	62.5%

Conclusion

Acoustic analysis of voice signal is showing better outcome though it is time consuming process. The application of cepstral analysis for the clinical evaluation of voice function has been qualitatively reviewed on asthmatic patients. The mathematical transformations involved in the analysis have been described as well as the suitability of the analysis for this application is described.

References

[1] [My paper]S Hackenberg, T Hacki, R Hagen, N H Kleinsasser, "Voice Disorders in Asthma" Laryngorhinootologie. 2010 Aug460-4. Epub 2010.

[2] Lee L, Chamberlain LG, Loudon RG, Stemple JC, " Speech segment durations produce by healthy and asthmatic subjects", J Speech Hear Disord. 1988 May;53(2):186-93.

[3] L Lee, R G Loudon, B H Jacobson, R Stuebing, " [My paper]Department of Communication Sciences and Disorders, University of Cincinnati, Ohio 45221.

[4] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, "Speaker Identification Using Mel Frequency Cepstral Coefficients", 3rd Proceedings of International Conference on Electrical & Computer Engineering, ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh, pp 565- 568

[5] S. B. Davis, "Acoustic characteristics of normal and pathological voices, " Speech and Language: Advances in Basic Research and Practice", vol. 1, pp. 271–335, 1979.

[6] P.Yu, M.Ouaknine, J.Revis, and A.Giovanni, "Objective Voice Analysis for Dysphonic Patients: A Multiparametric Protocol Including Acoustic and Aerodynamic Measurements", Journal of Voice Vol. 15, No. 4, pp.529–542, 2001

[7] B. Boyanov, S.Hadjitodorov: "Acoustic analysis of pathological voices: a voice analysis system for screening of laryngeal diseases", Proc. IEEE Engineering in Medical and Biology, (1997), vol. 16, no. 4, 74-82.

[8] Lindasalwa muda, Mumtaj Begam, I. Elamvazuthi, "Voice Recognition Algorithms using MFCC & DTW Techniques", Journal of Computing, volume2, issue3, march 2010.

[9] Jamel Price, Ali Eydgahi, "Design of Matlab based Automatic Speaker Recognition Systems", 9th International Conference on Engineering Education, July2006.

[10] Rosalyn J. Moran*, Richard B. Reilly, "Telephony-Based Voice Pathology Assessment Using Automated

Speech Analysis'. ieee transactions on biomedical engineering, vol. 53, no. 3, march 2006.
[11] By Jackie Ehrlich, "Asthma, Allergies, Articulation: a Speech Therapy Perspective"
[12] Dirk Michaelis, "Selection and combination of acoustic features for the description of pathologic voices", 1998,
[13] Tripti kapoor, Dr. R.K.Sharma, " Parkinson's disease Diagnosis using Mel Frequency Cepstral Coefficients and Vector Quantization", International Journal of Computer Applications (0975 – 8887) Volume 14– No.3, January 2011
[14] Batler, "Inhaled Beclomethasone Dipropionate Improves Acoustics Measure of Voice in Asthmatic Patient", Journal of Speech and Hearing Research, Volume 39, 126-134, February 1996
[15] Lingyun Gu, John G. Harris, " Disordered Speech Assessment Using Automatic Methods Based on Quantitative Measures", Journal on Applied Signal Processing, 2005
[16] "Subjective and objective evaluation of voice quality in patients with asthma", Journal of Voice, 2007
[17] L. Rabiner, and B.H. Juang, "Fundamentals of Speech Recognition". Prentice Hall, 1993.
[18] Risa J Robinson, Richard L Doolittle, John N Diflorio, "Use of asthmatic pulmonary function test data to predict lung deposition", [My paper]J Aerosol Med. 2007 ;20 (2):141-62
[19] Kosztyła-Hojna Bb, " Objective analysis of voice quality in asthma patients", Int. Rev. Allergol. Clin. Immunol., 2009; Vol. 15, No. 1-2
[20] Praat software www.praat.org.
[21] Han Su Kim, Jin Wook Moon, " A Short-Term Investigation of Dysphonia in Asthmatic Patients Using Inhaled Budesonide, '' Journal of Voice, Vol. 25, No. 1, pp. 88-93.
[22] http://www.mathsworks.com

Carrier to Noise Ratio Analysis of Radio over Fiber System based on Optical Single Side Band

[1]Abhimanyu, [2]Keshav Dutt, [3]Manisha and [4]Amit Mahal

[1]Assistant Professor, Indus Institute of Engineering & Technology, Jind, Haryana, India
E-mail: nainabhi@gmail.com
[2]Scholar, Somany Institute of Technology & Management, Rewari, Haryana, India
E-mail: keshavdutt88@gmail.com
[3]Scholar, Sri Sukhmani Institute of Engineering & Technology, Derabassi, Punjab, India
E-mail: manisha.guleria@gmail.com
[4]Assistant Professor, Indus Institute of Engineering & Technology, Jind, Haryana, India
E-mail: ad.indus@gmail.com

Abstract
Carrier to Noise Ratio (CNR) has been investigated for radio over fiber systems including the effects of chromatic dispersion, phase noise due to RF oscillator and electrical filter bandwidth in this paper. Optical Single sideband signal is studied as it has tolerance for power degradation due to dispersion effects over a length of fiber. Investigations have been made out for Radio frequency of 30 GHz, with a continuous wave (CW) laser source of 1550 nm. CNR has been evaluated using Power Spectral Density(PSD) function.CNR is studied for varying electrical receive required filter power to total RF power ratio over 10 km fiber with chromatic dispersion D=17 ps/km nm.

Keywords: *CNR, MZM, OSSB, Power degradation*.

Introduction

For satisfying the increasing capacity and data rate of subscribers, wideband communication systems are necessary in both wired and wireless link. A Radio over fiber (RoF) system is one of the most attractive systems for future broadband wireless communication having a high data rate at a microwave or millimeter wave frequency band because of the advantages of an optical fiber including the low-transmission loss and ultra wide bandwidth [1]. The volume of data traffic is ever increasing due to the demand of subscribers for voice, data, and multimedia services that require the access network to support high data rates at any time and in any place inexpensively. Generally, RoF systems transmit an optically modulated radio frequency (RF) signal from a central station (CS) to a base station (BS) via an optical fiber. The RF signal recovered using a photo detector (PD) at the BS arrives at a mobile station (MS) through a wireless channel as shown in Fig. 1.This architecture provides a cost-effective system since any RF oscillator is not required at the BS. However, the performance of RoF systems depends on the method used to generate the optically modulated RF signal, power degradation due to fiber chromatic dispersion, nonlinearity due to an optical power level, and phase noises from a laser and an RF oscillator. Therefore, it has been a matter of concern and interest to investigate parameters that degrade the performance of RoF system.

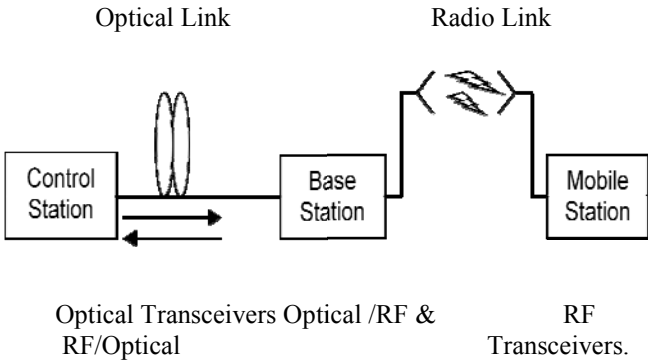


Figure 1: Optical Link in Radio over Fiber System.

Single sideband (SSB) modulation scheme is an effective way to eliminate the dispersion effects in RoF system. The power degradation due to fiber dispersion can be overcome by employing an optical single sideband modulation scheme [2].The nonlinear effect of an optical fiber can be managed by the modulation format and control of a launched optical power [3] [4].Unlike those parameters, a phase noise is one of practical and decisive factors in high quality services which require high carrier to noise ratio (CNR) because it result in the bit error rate (BER) floor at high carrier to noise ratio (CNR) values [5].This phenomenon is serious to RoF systems because the purpose of RoF systems is to provide high data rate and high quality service requiring a large carrier to noise ratio. Thus the system performance can be much sensitive to the phase noise in such services.

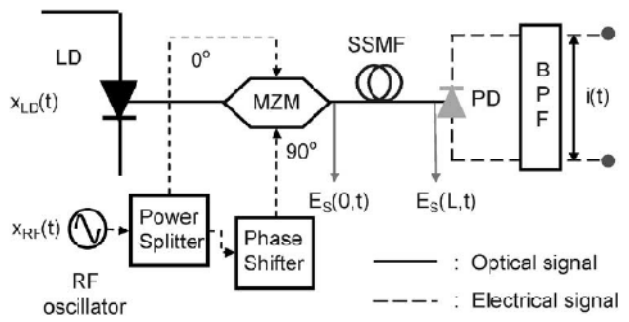


Figure 2: Radio over Fiber system using an OSSB modulation and direct detection scheme.

Here, we investigate the CNR penalty due to fiber chromatic dispersion and phase noises due to laser line width using an Optical Single Side Band (OSSB) signal and a direct-detection scheme. For the analysis of the Carrier to noise ratio penalty, the autocorrelation and the power spectral density function of a received photocurrent are evaluated. The bandwidth of an electrical filter is dealt in the CNR penalty since the phase noises result in an increase of the required bandwidth and the increased bandwidth causes an additional Carrier to noise ratio penalty. It is shown that the phase noise due to the laser line width is the dominant parameter in a large optical distance.

RoF System Based on Optical Single Side Band & Direct Detection

An Optical Single Side Band (OSSB) signal is generated by using Dual electrode mach zehender modulator (MZM) and a phase shifter. An RF signal from an oscillator is split by a power splitter and a 90 ◦ phase shifter. This RF signal is optically modulated by the Laser Diode (LD) with an MZM. The optically modulated signal is transmitted to the PD and the photocurrent corresponding to the transmitted RF signal is extracted by the BPF as in Fig. 2. First, the optical signals from the optical source, laser diode and the RF oscillator are modeled as:

$$x_d(t) = A^d.\exp\ j(w_d t + \Phi_d(t)) \tag{1.1}$$

$$x_o(t) = V_o.\cos(w_o t + \Phi_o(t)) \tag{1.2}$$

Where A^d and V_o define amplitudes from the optical source and the RF oscillator signal, ω_d and ω_o define angular frequencies of the signals from the LD and the RF oscillator, and $\Phi_d(t)$ and $\Phi_o(t)$ are phase-noise processes. The OSSB signal generated using Dual electrode MZM is modeled in equation (3).

$$E_{SS}(0,t) \cong A^d.L_{MZM}\begin{Bmatrix} J_0(\alpha\pi)\exp j\left[w_d t+\Phi_d(t)+\frac{\pi}{4}\right]-\sqrt{2}J_1(\alpha\pi) \\ \exp j\left[w_d t+\Phi_d(t)+w_o t+\Phi_o(t)\right] \end{Bmatrix} \tag{1.3}$$

After the transmission of signal over L km fiber, the signal can be represented as equation (4) & in this equation L_{add} denotes an additional loss in the optical link, α_{fiber} is the SSMF loss, L_{fiber} is the transmission distance of the SSMF, and $\tau 0$ and $\tau +$ define group delays for a center angular frequency of ω_d and an upper sideband frequency of $\omega_d + \omega o$. $\varphi 1$ and $\varphi 2$ are phase-shift parameters for specific frequencies due to the fiber chromatic dispersion.

$$E_{SS}(L,t) \cong \begin{bmatrix} A^d.L_{MZM}.L_{add}.10^{-\frac{\alpha_{fiber}L_{fiber}}{20}}J_0(\alpha\pi) \\ \begin{Bmatrix} \exp j\left[w_d t+\Phi_d(t-\tau_0)-\phi_1+\frac{\pi}{4}\right]\frac{\sqrt{2}J_1(\alpha\pi)}{J_0(\alpha\pi)} \\ \exp j\left[w_d t+\Phi_d(t-\tau_+)+w_o t+\Phi_o(t-\tau_+)-\Phi_2\right] \end{Bmatrix} \end{bmatrix} \tag{1.4}$$

To evaluate the CNR, we utilize the autocorrelation function and the PSD of the photocurrent.

$$i(t) \cong \eta\left|E_{SS}(L,t)\right|^2 \tag{1.5}$$

Where η defines the responsivity of the PD and $|.|^2$ is the square-law detection.

$$i(t) \cong \eta\left|A_1^d\right|^2\left\{B+2\alpha_1\cos\begin{bmatrix}\Phi_d(t-\tau_+)-\Phi_d(t-\tau_0) \\ +w_o t+\Phi_o(t-\tau_+)-\Phi_2+\Phi_1\end{bmatrix}\right\} \tag{1.6}$$

Where

$$A_1^d = A^d.L_{MZM}.L_{add}.10^{-\frac{\alpha_{fiber}L_{fiber}}{20}}J_0(\alpha\pi)$$

$$\alpha_1 = \frac{\sqrt{2}J_1(\alpha\pi)}{J_0(\alpha\pi)} \text{ and } B = 1+\alpha_1^2$$

The autocorrelation function $R_I(\tau)$ is obtained as

$$R_1(\tau) = \left\langle i(t).i(t+\tau)\right\rangle \tag{1.7}$$

Now we will evaluate PSD function which is Fourier transform of $R_I(\tau)$

$$S_1(f) = F\left\langle R_1(\tau)\right\rangle \tag{1.8}$$

Where

$$S_1(f) = R_1(\tau)\int_{-\infty}^{\infty} R_1(\tau)d\tau * \exp(-j\tau w) \tag{1.9}$$

In equation (9), the first term represents a dc component, the second and third is the broadening effects due to the fiber chromatic dispersion and the line widths of the laser and the RF oscillator. the second term was only a carrier to noise penalty due to the fiber chromatic dispersion.Now the received RF carrier Power P1 is approximately represented as follows

$$P_1 = 2\int_{f_o-\frac{B_o}{2}}^{f_o+\frac{B_o}{2}} S_1(f)df \tag{1.10}$$

By using (9), received RF carrier power P_1 as

$$P_1 \cong \frac{4\eta^2 A_1^{d4}\alpha_1^2}{\pi}\exp(-2\Upsilon_t|\tau|)\tan^{-1}\left(\frac{\pi.B_o}{2\Upsilon_o}\right) \quad (1.11)$$

The CNR induced by the differential delay from the fiber chromatic dispersion and the line widths from the laser and the RF oscillator is found

$$CNR \cong \frac{P_1}{2B_o.\left(\frac{N_o}{2}\right)} \quad (1.12)$$

$$CNR \cong \frac{2\eta^2 A_1^{d4}\alpha_1^2 p}{N_o.\left(\frac{\Upsilon_o}{\pi}\right)\tan\left(\frac{\pi.p\exp(-2\Upsilon_t|\tau|)}{2}\right)} \quad (1.13)$$

Where $\eta = responsivity$, A_1^d= constant related to the laser light amplitude A and the losses in fibre, MZM and the joint and splices given by $\alpha_1 = \frac{\sqrt{2}J(\alpha\pi)}{J_{o(\alpha\pi)}}$ J= Bessel function of 1st kind, of order n and α_1 =normalized RF voltage given by $\alpha_1 = \frac{V_{rf}}{V_\pi}$Where A_1^d is the amplitude of laser light, L_{MZM} is the lose in the MZM, L_{add} is the factor accounting for the additional loss in the fiber, α_{fiber} is the loss in the fiber and L_{fiber} is the length of fiber. V_{rf} is the input RF voltage and V_π is the MZM switching voltage, p is the ratio of the power required for a particular filter used to the total carrier power.This parameter incorporates the effect of the bandwidth of the filter being used.And N_o is the additive white Gaussian noise power spectral density. The parameters $2_{\gamma LD}$= $2_{\pi\Delta VLD}$ and $2_{\gamma RF}$ = $2_{\pi\Delta VRF}$, define the angular full-linewidth at half maximum(FWHM)of the Lorentzian shape for the laser and the RF oscillator.And $2_{\gamma t}$ = $2_{\pi\Delta VLD}$+ $\pi\Delta VRF$ gives the total linewidth.$\tau = \tau = \tau \pm \tau_o$ is the differential delay due to the fiber chromatic dispersion and is given by$\tau = D.L_{fiber}.\lambda^2.\frac{f_{RF}}{c}$ Where D is the fiber chromatic dispersion parameter, L_{fiber} is the fiber length, f_{RF} is the RF frequency and c is the speed of light.

Result and Discussion

CNR is evaluated and resulting values are simulated to study and realize the importance of electrical filter bandwidth on CNR which can be seen from the resulting plot..The effect of the filter bandwidth dependent factor p on the CNR of the system is shown in the figure. Here, CNR is plotted against the parameter p for two different values of laser line width viz. 10 MHz, 300 MHz and the RF oscillator line width is assumed to be 0.8 Hz. The first parameter is the photodiode responsivity$\Re$. For most of the photo diodes its values is between 0.6 to 0.8. Taking the value of R as 0.7 [11]. Now the second constant is A1 which in tem depend upon other parameters given as

$$A_1 = A.L_{MZM}.L_{add}.10^{\frac{-\alpha_{fiber}L_{fiber}}{20}}.J_o(\alpha\pi)$$
$$= V_{rf}/V_\pi$$

Here LMZM is the loss of the DE- MZM. Now considering the MZM as a integrated waveguide power splitter and combiner, its value can be assumed to be negligible (which is true for small lengths of the waveguide).Ladd is the additional loss caused by the fiber components such as the splices, joints etc.Its value for an 10 Km fiber link can be taken as approximately 3 dB. αfiber is the loss per Km of the fiber and is around.2dB/Km for SSMF. Lfiber is the length of the fiber and is equal to 10 Km for this case. α is the modulation index of the MZM and is equal to α=Vrf/Vπ Now taking Vrf=1mV and Vπ=2.2V, we obtain α=.00045 then the modulation index is given as απ=0.0014. It gives J0(απ) equal to 1 approximately. From above all, the value of A1 is calculated as 0.1342. N0 is the power spectral density of the AWGN for very low noise case, it can be taken as 10^{-11}. Now α1 depends upon the first harmonic of the photo detector and the fundamental component. So the value of α1 is 0.001. Thus all the constants terms are evaluated, then CNR is studied including the effects of the laser and RF oscillator line width.

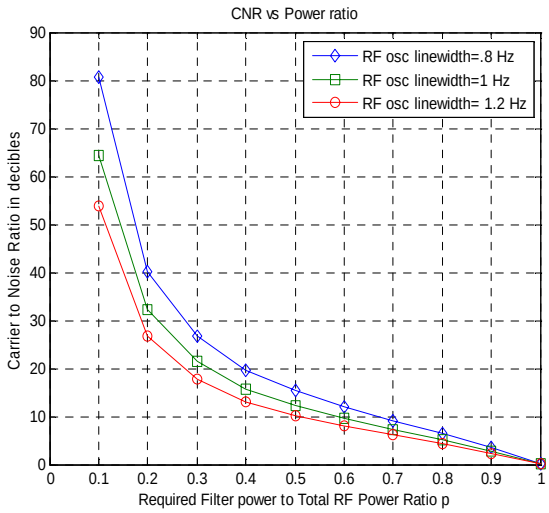


Figure 3: CNR in dB vs. required filter power to total RF power

As the result, in Fig. 3, CNR consists of the Laser-linewidth effect γo and the ratio p. The effect of γo is linearly proportional to CNR, as shown in Fig. 3. CNR decreases as p becomes large since the increment of the noise power is greater than that of the received signal power as the bandwidth increases. Thus, the bandwidth should be considered carefully for p > 90%, since the CNR decreases drastically over the point as a result. For example, the CNR of p = 0.99 is 15.1 dB as compared to p = 0.1. The received RF signal power will decrease less than 50% at p=0.5. Thus, the minimum required power to detect the signal should be carefully considered before we choose the filter bandwidth.

Conclusion

CNR has been investigated due to the phase noise from the laser for various line widths over different lengths of fiber. Here a direct detection for cost effectiveness has been used. It

is evident that the CNR decreases as the power ratio increases following the exponentially decrement.

References

[1] P. Smolders, "Exploiting the 60 GHz Band for Local Wireless Multimedia Access: Prospects and Future Directions, " *IEEE Commun. Mag.*, Jan. 2002, pp. 140–147.

[2] G. H. Smith and D. Novak, "Overcoming chromatic-dispersion effects in fiber-wireless systems incorporating external modulators, " IEEE Trans. Microw. Theory Tech., Vol. 45, No. 8, Aug. 1997, pp. 1410–1415.

[3] J. Leibrich, "CF-RZ-DPSK for suppression of XPM on dispersion managed long-haul optical WDM transmission on standard single-mode fiber, " IEEE Photon. Technol. Lett., Vol. 14, No. 2, Feb. 2002, pp. 155–157.

[4] Y. J. Wen, "Power level optimization of 40 Gb/s DWDM systems with hybrid Raman/EDFA amplification, " in Proc. Conf. Optical Internet/Australian Conf. Optical Fibre Tech. (COIN/ACOFT), Melbourne, Australia, Jul. 2003, pp. 309–312.

[5] J. R. Barry and E. A. Lee, "Performance of coherent optical receivers, " *Proc. Inst. Elect. Eng.*, Vol. 78, No. 8, Aug. 1990, pp. 1369–1394.

[6] P. B. Gallion and G. Debarge, "Quantum phase noise and field correlation in single frequency semiconductor laser systems, " *IEEE J. Quantum Electron.*, vol. QE-6, no. 4, pp. 343–349, Apr. 1984.

Effect of Four Wave Mixing in WDM Optical Fiber Systems

[1]Shelly Garg, [2]Keshav Dutt, [3]Abhimanyu and [4]Manisha

[1]Associate Professor, Indus Institute of Engineering & Technology, Jind, Haryana, India
E-mail: s_singla428@rediffmail.com
[2]Scholar, Somany Institute of Technology & Management, Rewari, Haryana, India
E-mail: keshavdutt88@gmail.com
[3]Assistant Professor, Indus Institute of Engineering & Technology, Jind, Haryana, India
E-mail: nainabhi@gmail.com
[4]Scholar, Sri Sukhmani Institute of Engineering & Technology, Derabassi, Punjab, India
E-mail: manisha.guleria@gmail.com

Abstract
This paper introduces the non linear optical effect known as four wave mixing (FWM). In wavelength division multiplexing (WDM) systems four wave mixing can strongly affect the transmission performance on an optical link. In this paper, we investigate the effect of input signal power on the optical powers of the signals generated corresponding to signals for the four wave mixing effect. Here power signals P_{321} and P_{332} have been investigated and it has been found that P_{332} is almost proportional to the input power emphasizing that the power signals as small as 20pW could be observed for input signal powers at receiver unit.

Keywords: Four-wave mixing (FWM), Wavelength division multiplexing (WDM), nonlinear effects.

Introduction

Four wave mixing (FWM) is one of the major limiting factors in wavelength division multiplexing (WDM) optical fiber communication systems that use low dispersion fibers or narrow channel spacing. Shibata et al. [1] stated that estimation of the FWM efficiency is very important for both the design and evaluation of wavelength division multiplexed (WDM) system. Song et al. [3] reported that the generation of a new frequency of radiation due to FWM has applications in the development of tunable sources and wavelength conversion in all-optical routing systems. However, generation of light through four-wave mixing has serious implications for the rapidly expanding telecommunications field of wavelength division multiplexing (WDM). If two or more channels interact with each other through four-wave mixing, optical power will be generated with new frequencies at the cost of a reduction of power in the original channels. This power loss makes it more difficult to correctly detect the digital data in these channels at the far end of the fiber, making errors more likely. Kyong et al. [7] reported that a more damaging consequence that is the FWM between two or three channels generating light at a frequency that coincides with one of the other allocated channels. The FWM generated light then acts as noise on this channel and leads to even greater degradation of the overall system performance. It is therefore important to take steps to avoid four-wave mixing in multichannel optical communication systems. Four-wave mixing in WDM systems can be minimized by ensuring that phase matching does not occur [8]. This can be achieved by using a number of methods including spacing channels unequally and operating at wavelengths where channels propagate at different speeds. In multi-channel systems, a signal channel suffers from FWM, which generates various combinations of different channel frequencies and causes crosstalk degradation. For any three co-propagating optical signals with frequencies f_i, f_j, f_k the new frequencies f_{ijk} generated by FWM are represented by $f_{ijk} = f_i + f_j - f_k$ for i,j,k. Considering all the possible permutations, N co-propagating optical signals will give rise to M new optical signals as

$$M = N^2(N-1)/2$$

Some of these new frequencies fall onto the N original channels, while others are found in other new frequency locations. Those FWM signals, which overlap with the original ones, are considered as crosstalk and will interfere with the normal operation of the WDM channels. This crosstalk between neighboring channels places a lower limit on the wavelength separation between adjacent channels and an upper limit on the input power in each channel. In the present paper, the optical power signal for FWM in optical fiber systems has been studied and investigated and sensitivity of WDM systems to input powers has been found.

Theoretical Background

Through an FWM process, three waves of frequencies f_i, f_j, f_k ($j \neq k$) generate the frequency $f_{ijk} = f_i + f_j - f_k$ (subscripts i,j and k select 1, 2, and 3). Fig. 1(a) and (b) illustrates schematically three different signal frequenciesf_1, f_2 and f_3, and nine new frequencies f_{ijk} The identification of the origin of the newly generated frequencies is clearly understood from Fig. l(a). Three signal frequencies and the frequencies generated through the FWM are overlapped for the situation of equal frequency separation $\Delta f = f_2 - f_1 = f_3 - f_2$, as shown in Fig. l(b). This leads to the crosstalk problem in frequency multiplexed coherent communication systems. The time-averaged optical power $P_{ijk}(L, \Delta\beta)$ generated through the

FWM process for the frequency component f_{ijk} is written as [4-5]

$$P_{ijk}(L) = \left(\frac{1024\pi^6}{n^4\lambda^2c^2}\right)(Dx_3)^2\left(\frac{P_iP_jP_k}{A_{eff}^2}\right)$$

$$X[\{\exp(i\Delta\beta - \alpha)L - 1\}/(i\Delta\beta - \alpha)]^2$$

where L is the fiber length, n is the refractive index of the core, λ is the wavelength, c is the light velocity in free space, D is the degeneracy factor, which can select D = 1, 3, and 6**,** x_3, is the third-order nonlinear susceptibility, A_{eff} is the effective area for the guided HE_{11}mode, P_i,P_j,P_k are the input powers launched into a single mode fiber, $\boldsymbol{\alpha}$ is the fiber attenuation coefficient, and Δβ is the propagation constant difference. Here β indicates the propagation constant.

D_c is the fiber-chromatic dispersion value given by

$$D_c = -\left(\frac{\omega_k^2}{2\pi c}\right)\left[\frac{d^2\beta(\omega_k)}{d\omega^2}\right]$$

The propagation constant difference is given by

$$\Delta\beta = \left(\frac{2\pi\lambda_k^2}{c}\right)\Delta f_{ik}\Delta f_{jk}$$

$$X[D_c + \left(\frac{\lambda_k^2}{2c}\right)(\Delta f_{ik} + \Delta f_{jk})\left\{\frac{dD_c(\lambda_k)}{d\lambda}\right\}]$$

where $\Delta f_{mn}=[f_m-f_n]$ ***(m, n = i, j , k) .*** Generally, D_c dominates, and the contribution of $dD_c/d\lambda$ can be neglected at the wavelength far from zero chromatic dispersion wavelengths around 1.3 and 1.55 μm. At the zero chromatic dispersion wavelength D_c , = 0 and the dispersion slope $dD_c/d\lambda$ must be included. The generated wave efficiency $\boldsymbol{\eta}$ with respect to phase mismatch ΔβL can be expressed as

$$\eta = \left[\frac{\alpha^2}{\{\alpha^2+(\Delta\beta)^2\}}\right]\left[\frac{1+4\exp(-\alpha L)sin^2(\Delta\beta L)}{\{1-\exp(-\alpha L)\}^2}\right]$$

For the operating wavelength far from the zero chromatic dispersion wavelength, the efficiency η is described as a function of the equivalent frequency separation $\Delta f_{eq}(ijk)$ defined by

$$\Delta f_{eq}(ijk)=(\Delta f_{ik}\Delta f_{jk})^{1/2}$$

From above equations, we can evaluate η as a function of $\Delta f_{eq}(ijk)$for a given fiber length L.

Here it is noted that the degeneracy factor D = 1, 3, or 6 depends on whether three, two, or none of the frequencies f1 , f2, and f3 are the same. For respective cases of $f_i=f_j\neq f_k$ $f_i\neq f_j\neq f_k$,and $P_{ijk}(L,\Delta\beta)$ is then rewritten as

$$P_{ijk}(L) = \eta\left(\frac{1024\pi^6}{n^4\lambda^2c^2}\right)(6x_3)^2$$

$$X(\frac{L_{eff}}{A_{eff}})^2P_iP_jP_k\exp(-\alpha L)$$

Where L_{eff} is the effective interaction length given as $L_{eff}=[1-\exp(-\alpha L)]/\alpha$.

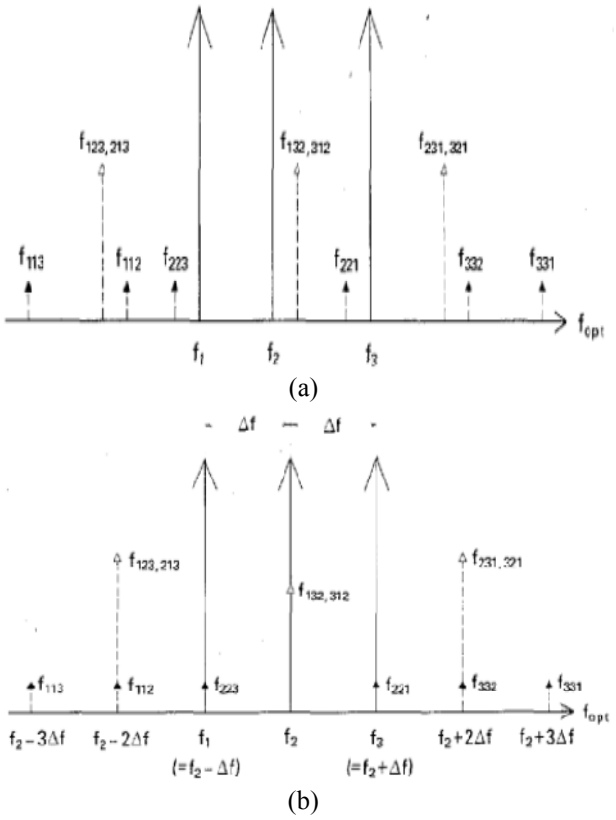


Figure 1: Sketch of the three input waves and the nine waves generated through the four wave mixing process. (a) For different frequency separation(b)For equal frequency separation with respect to f_3-f_2 and f_2-f_1.

Simulation and Analysis

First, optical powers of the generated waves have been studied as a function of the input signal power for frequency components f_{231},f_{321},f_{332} under various conditions.

Simulation Parameters

Attenuation(α)	0.2dB/km
Speed of light (c)	$3x10^8$ m/sec
Effective core area (A_{eff})	70μm^2
Length of fiber (L)	10 km
Third-order nonlinear susceptibility (x_3)	$6x10^{-14}$m^3/J
Degeneracy factor (D)	6 for $i \neq j \neq k$

Consider f_2-f_1=17.2GHz, f_3-f_2=11.0GHz P_1=0.43mW, and P_2 = 0.14 mW. The input signal power P_3 was varied from 0.15 to 0.60 mW. Fig. 2 shows the dependence of P_{ijk} on the input signal power P_3. Circles indicate the measured values of the generated power for the frequency components of f_{321},f_{332} respectively. It was found that P_{321} increases slightly with input power P_3, while P_{332} varies linearly with the input power, P_3, which can also be predicted from its expression. It should be emphasized that optical powers **as** small as 20 pW could be observed for input signal powers below 1 mW by the heterodyne detection technique.

Conclusion

The optical output powers of the newly generated waves through the FWM process are studied theoretically and simulative investigation is carried out to study the impact of input power variation on the respective output components (P_{331}, P_{332}). From the investigations it can be stated that the very small optical powers could be observed for input signal powers below 1 mW resulting into performance degradation of optical fiber systems.

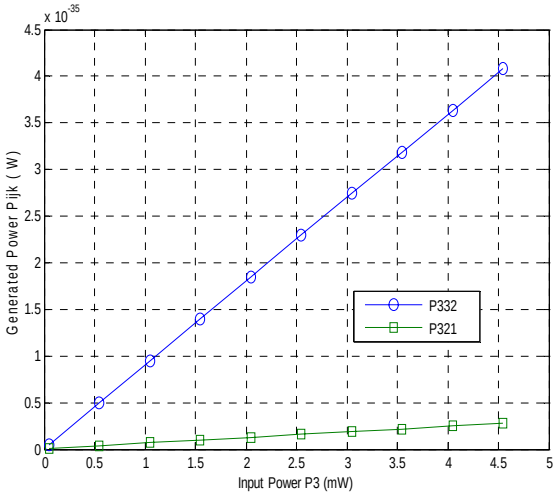


Figure 2: Output powers with frequency components of f_{321} and f_{332} as a function of input power.

References

[1] N. Shibata, R.P.Braun and R.G.Waarts, “Crosstalk due to three wave mixing in a coherent single mode transmission line” Electron Lett., Vol. 22,pp. 675-677, 1986.

[2] R.G.Waarts and R.P.Braun, “Crosstalk due to stimulated brillouin scattering in monomode fiber” Electron Lett., Vol. 21,pp. 1114-1115, 1985.

[3] Shuxian Song, Christopher T. Allen, Kenneth R. Demarest and Rongqing Hui, "Intensity-Dependent Phase-Matching Effects on Four-Wave Mixing in Optical Fibers," J. Lightwave Technol., Vol. 17, No. 11, pp. 2285-2290, 1999.

[4] Jeff Hecht, Light Nonlinear Effects: Understanding Fiber Optics, Prentice Hall, 2004.

[5] K. 0. Hill, D. C. Johnson, B. S. Kawasaki, and R. I. MacDonald, "CW Three-Wave Mixing in Single-Mode Fibers." J. Appl. Phys., Vol. 49, No. 10, pp. 5098-5106, 1978.

[6] N. Shibata, “Phase-Mismatch Dependence of Efficiency of Wave Generation through Four-Wave Mixing in a Single-Mode Optical Fiber," IEEE J. Quantum Electron., QE-23, pp. 1205-1210, 1987.

[7] Kyong Hon Kim, Hak K;yu Lee, Seo Y. Park and El-hang Lee, "Calculation of Dispersion and Nonlinear Effect Limited Maximum TD)M and FDM Bit Rates of Transform-Limited Pulses in Single-Mode Optical Fibers," J. Lightwave Technol., Vol. 13, No. 8, pp. 1597-1605, 1995.

[8] Liren Zhang and Junhua Tang, "The Effect of IP Traffic on WDM-Based Networks," IEEE Commni. Lett. Vol. 4. No. 9, pp. 295-297, 2000.

Quantum Cryptography & its Comparison with Classical Cryptography: A Review Paper

Aakash Goyal[1], Sapna Aggarwal[2] and Aanchal Jain[3]

[1]M.Tech. Student (CSE) JIET-Jind, India
[2]Assistant Professor, CSE Department, JIET-Jind, India
[3]M.Tech. Student (ECE), BMIET-Sonipat, India
E-mail: aakash.goyal99@gmail.com[1], sapna.ruby@gmail.com[2], aanchal.always@gmail.com[3]

Abstract
Cryptography long had been a valuable, essential tool for defensive computer security. A technique whether classical or modern must be well-built and also must be practically viable. With the application of quantum mechanics principles to cryptography, a new dimension to secure communication can be given; such system so developed can detect eavesdropping and assure that it should not occur at all. Work presents the brief review of the existing state of quantum cryptography science. Core principles of the quantum cryptography are demonstrated by giving the example of BB84 protocol. Work principally presents quantum cryptography's comparison with classical cryptography.

Keywords: BB84 protocol;, Quantum cryptography; Classical cryptography; Polarization states;

Introduction

Cryptography is the science in which the use of mathematics occurs to encrypt and decrypt data .Cryptography enables user to store sensitive information or transmit it across Insecure networks (like the Internet) so that it cannot be read by anyone other than the intended receiver. To attain secure transmission, an algorithm is used which unite the message with additional information to produce a cryptogram. The algorithm is called as cipher and the additional information is known as the key. This technique is termed as encryption.

Whereas cryptanalysis is the science of analyzing and breaking secure communication. Classical cryptanalysis involves an interesting combination of analytical reasoning, application of mathematical tools, pattern finding, patience, determination, and luck. Quantum cryptographic devices in general make use of individual photons of light and take benefit of Heisenberg's uncertainty principle, according to which attempt to compute a quantum system disturbs it and yields partial information about its state before the measurement. Eavesdropping on a quantum communications channel thus causes an unavoidable disturbance, alerting the legal users. Quantum techniques also support the achievement of cryptographic objectives such as enabling two mutually suspicious parties to make combined decisions based on private information, while compromising its confidentiality as little as possible. Physical devices with these specialized cryptographic protocols can invoke up streams of random bits whose values will remain unknown to third parties. When we use these bits as key material for Vernam ciphers, we can get Shannon's ideal of perfect secrecy—cheaply and easily.

The development of quantum cryptography was inspired by the short comings of classical cryptography methods. In classical cryptography communicating parties need to share a secret sequence of random numbers, the key, that is exchanged by physical means and thus open to security loopholes. The classical cryptography does not detect eavesdropping like quantum cryptography, also with increase in computing power and new computational techniques are developed, the numerical keys will no longer be able to provide satisfactory levels of secure communications. These flaws led to the development of quantum cryptography, whose security basis is quantum mechanics [2]. This paper presents the comparison of quantum and classical cryptography on several backgrounds.

Historical Timeline

In 1917, Gilbert S Vernam, an AT&T employee, created a machine that makes a non-repeating, virtually random sequence of characters called as one-time pad. Using an encryption key the same length as the message and never using that key again is the only proven method of securely communicating. In the 1940s, Claude Shannon provided the information-theoretic basis for secrecy; the amount of uncertainty that can be commenced into an encoded message can't be greater than that of the cryptographic key used to encode it. In the early 1970s, or possibly earlier, numerous researchers, including Whitfield Diffie, Martin Hellman, Ralph Merkle, Ron Rivest, Adi Shamir, Leonard Adleman, James Ellis, Clifford Cocks, and Malcolm Williamson, invented cryptographic techniques based on computational complexity. Quantum cryptography was first proposed in 1984 by Brennet and Brassard [1] based on the No-Cloning theorem. It relies on fact that we should base security on known physical laws not on mathematical complexities.

Classical Cryptography

Classical cryptography makes use of several mathematical techniques to restrict eavesdroppers from knowing the contents of encrypted messages. The most popular among them that are adopted globally have been described below.

Throughout the paper, the transmitter is referred as 'Alice', the receiver as 'Bob', and an eavesdropper as 'Eve'.

Data Encryption Standard (DES)

The data encryption algorithm developed by IBM for NBS was based on Lucifer, and it became known as the Data Encryption Standard, although its proper name is DEA (Data Encryption Algorithm) in the United States and DEA1 (Data Encryption Algorithm-1) in other countries. In DES [4] the encrypting and decrypting algorithms are publicly announced; the security of the cryptogram depends entirely on the secrecy of the key and the key consist of randomly chosen, sufficiently long string of bits. This algorithm ensures that the output bits have no apparent relationship to the input bits and spreading the effect of one plaintext bit to other bits in the cipher text. Once the key is established between the sender and the receiver, subsequent communication involves sending cryptograms over a public channel which is vulnerable to total passive eavesdropping. However with the purpose of establishing the key between two users, who share no secret information initially, must at a certain stage of communication use a reliable and a very secure channel. A random key must first be send through a secret channel before the transfer of actual message. The major drawback of DES is that like other classical cryptographic mechanism it also cannot guarantee ultimate security of a communication channel.

Public Key Cryptographic (PKC) Systems

With a conventional symmetric key system, each pair of users needs a separate key. As the number of users grows, the number of keys increases very rapidly. An n-user system requires n * (n - 1)/2 keys, and each user must track and remember a key for each other user with which he or she wants to communicate. We can reduce the problem of key proliferation by using a public key approach. In a public key or asymmetric encryption system, each user has two keys: a public key and a private key. The user may publish the public key freely because each key does only half of the encryption and decryption process [3]. The keys would be inverse functions. If 'Alice' wants to send a secret message to 'Bob', he would encrypt his message with Bob's public key and send it via an insecure channel. 'Bob' receiving the message would then decode it using his private key. This methodology ensures that the sender can't decode his own message once encrypted. These systems make use of the fact that certain mathematical operations are simple to do in one direction than the other. For example, multiplication of two large prime numbers is easy but factoring the result would be infeasible if the number is large and computational assets are poor. RSA (Rivest-Shamir-Adleman encryption Algorithm), the first PKC cryptosystem [5] obtains its security from the straightforward fact that factoring of large numbers is extremely tough. The disadvantage of classical cryptosystem is that it provides no method for detecting eavesdropping. Also, with the building of feasible quantum computer Shor's algorithm could easily break RSA in polynomial time.

One-time pad (OTP) Cryptosystem

The one-time pad cryptosystem [6] created by Gilbert Vernam in 1917 is very simple and yet, very effective. The system ensures perfect secrecy. A one-time pad is sometimes considered as ideal cipher. The system is named after encryption method in which a large, non repeating set of keys is written on sheets of paper, attached together into a pad. For the encryption to work, the receiver needs a pad identical to that of the sender. The major disadvantage with one-time pad is even though its security it is very impractical. For every message encoded with the system, the participants have to to exchange a secret key that has at least the same length. The one-time pad method needs absolute synchronization between sender and receiver and no key should be used twice.

Quantum Cryptography

Quantum channel construction requires a pair of polarizing filters at both sender and receiver ends. So, that at sender end we can send photos of selected polarization and at receiver end to measure the polarization of photons. There are two types of polarization filters rectilinear and diagonal; in rectilinear filter we have horizontal and vertical orientation of photons whereas in diagonal we have 45 and 135 degree of orientation of photons. The two directions can be detected by vertically oriented calcite crystal and two detectors like photomultiplier. If the photon is horizontally polarized it will be directed to upper filter and to vertical detector if it is vertically polarized. If similar apparatus is rotated at 45 it will record diagonal directions. Thus the rotated apparatus is useless for rectilinear direction and vertical apparatus for diagonal direction .hence we cant measure both simultaneously thus verifying Heisenberg uncertainty principle.BB84 Protocol was developed by Charles H, Bennett of the IBM Thomas J. Watson Research Centre and Gilles Brassard of the University of Montreal, quantum cryptography is based on the fact that measuring a quantum system such as a photon irreversibly changes its state and wipes out information about the aspects before measurement [1]. It uses two channels-quantum by which Alice and bob send polarized photons second is classical public channel by which they send ordinary messages such as comparing and conferring the signals sent through quantum channel. In Quantum Key Distribution sequence of operations are as follow:-

First, Alice generates and forwards Bob a sequence of photons with polarizations that are chosen randomly (0, 45, 90 or 135 degrees). Bob receives the photons and chooses randomly whether to measure its rectilinear or diagonal polarization for each photon. Next Bob publicizes which kind of measurement he has made (either rectilinear or diagonal) but not the measurement result for each photon. Alice tells him openly, whether he has made the correct type of measurement for each photon. Alice and Bob then discard all cases in which Bob has made the incorrect measurement or in which his detectors have failed to record a photon.

Bit sequence	0	1	2	3	4	5
Alice logic sequence	0	0	1	1	1	1
After passing a polarizing filter	↑	↖	→	→	↗	↗
Bob's polarization states	↑	↖	↖	→	↑	↗
Bob's correct states tested by Alice	V	V		V		V
Quantum key	↑	↖		→		↗

Figure 1: Sequence of operations

If none has eavesdropped on the quantum channel, the left over polarizations should be shared as secret information between Alice and Bob. Alice and Bob next test for eavesdropping, for example, by openly evaluating and discarding a randomly selected subset of their polarization data. If the evaluation shows proof of eavesdropping, Alice and Bob abandon all their data and start again with a fresh lot of photons. Otherwise they adopt the left over polarizations as shared secret bits, reading 0 or 45-degree photons as binary 0's and 90 or 135-degree photons as binary 1's. If she makes the incorrect measurement, then she resends Bob a photon reliable with the result of her measurement, she will have forever randomized the polarization originally sent by Alice for a particular photon, which causes errors in one fourth of the bits in Bob's data that have been subjected to attack since one has no information of Alice's secret choice, 50% of the time (probability 1/2) one will estimate exactly and 50% of the time (probability 1/2) one will estimate incorrectly. If one estimates exactly, then Alice's transmitted bit is received with probability 1. On the other hand, if one estimates wrongly, then Alice's transmitted bit is received correctly with probability 1/2[7]. Overall the probability of accurately receiving Alice's transmitted bit is

P=1.1/2+1/2.1/2=3/4

The BB84 scheme was customized to produce a working kit of quantum cryptography at IBM. The modifications were done to handle with practical problems like noise in the detectors. In BB84 scheme encoding is done as single polarized photon for each bit but this kit encodes each bit in a dim flash of light. This initiates a fresh eavesdropping danger to the system, if Eve taps into the beam; splitting each flash into two flashes of lesser intensity can be done, evaluating one for her while letting the other move to Bob. If Eve diverts only a meek fraction of the beam, Bob may not see the abating signal, or may take it as expected losses in the channel. This attack can be successfully let down by reducing data transmission rate, by sending very dim flashes of an intensity less than one photon per flash on average. Another problem is that available detectors for a moment produce a reaction even when no photon has been arrived which sources errors even when there has been no eavesdropping. An added fragile end is key storage. Once Alice and Bob have recognized the key, they must store it until it is required. But the longer they keep the key in, the more they are vulnerable to unauthorized check. It is possible to build a cryptosystem based on the well-known Einstein-Podolsky-Rosen (EPR) effect. Employing the EPR effect, Ekert recently developed a cryptosystem that gave assurance of security of both key storage and key distribution. It also cannot be used practically due to the technical infeasibility of stocking up photons for more than a tiny portion of a second [8].

Classical V/S Quantum Cryptography

Both quantum cryptography and classical cryptography can be compared on following dimensions:

Fundamental dimension

In theory, any classical private channel can be easily monitored inertly, without the knowledge to sender or receiver that the eavesdropping has been done. Classical physics is the theory of macroscopic bodies and phenomena such as radio signals that allows a physical property of an object to be measured without disturbing other properties. Cryptographic key like information is encoded in computable physical properties of some object or signal. Thus there is open possibility of passive eavesdropping in classical cryptography. Quantum theory which is basis of quantum cryptography is believed to direct all objects, but its consequences are mainly noticeable in individual atoms or subatomic particles like microscopic systems. As far as classical cryptography is concerned there is frequent requirement of using longer keys as computational power doubles in every 18 months and cost of computation is reducing rapidly with time [moors law]. Thus an algorithm using k bit key which is secure may not be secure in future, i.e. it needs regular updating. On the other hand,, security in quantum cryptography is based on the basic principles of quantum mechanics, so the possibilities of major changes requirements for future are almost negligible.

Commercial dimensions

Commercial solutions for QC that already exist; they are only suitable for point-to-point connections. On the other hand, crypto chip made by the siemens and Graz technical university [11] makes possible the creation of networks with many participants, and cost of €100,000 per unit, the system is very expensive and requires a lot of work. On other hand classical cryptography can be implemented in software and its cost for consumer is almost zero. Also, cryptographic system based on classical cryptography can be implemented on small hardware component like smart card , but this is major issue in case of quantum cryptography shrinkage to such a level require too much development.

Application dimensions

The digital signatures reveal the authenticity of the digital data to the receiver. A digital signature assures recipient that the message was formed by a known sender, and it was not changed in transit. The three main algorithms are key generation, signing, and key verification. But we know that

algorithms cannot be implemented in QC very easily. Therefore QC lacks many critical features like digital signature, certified mail etc.

Technological dimensions

Chinese scientists accomplished the worlds most long-distance of quantum communication transmission (teleportation), or as "instant matter transmission technology" technology. From the China University of Technology and researchers at Tsinghua University, Hefei National Laboratory in their free-space quantum communication experiments, and effectively enlarges the communication distance to 10 miles [9]. But classical cryptography can be used to communication distance of several million miles. According to the latest research, Toshiba achieve new record bit rate for quantum key distribution, that is, 1 Mbit/s on average [10]. On the other hand the bit rate of classical cryptography depends on the computational power largely.

Other dimensions

Communication medium is not an issue in classical cryptography because its security depends only on the computational complexity. Thus, this removes the need for excessively secure channels. On the other hand communication of quantum cryptography require a quantum channel like optical fiber or through air (wireless), also, there is constantly a likelihood of modification in polarization of photon due to Birefringence effect or rough paths that cause change in refractive index due to damage sometimes. Also, an n bit classical register can store at any moment exactly one n-bit string. Whereas an n-qubit quantum register can store at any moment a superposition of all 2^n n-bit strings.

Comparison of Quantum and Classical Cryptography

Features	**Quantum cryptography**	**Classical cryptography**
Basis	Quantum mechanics	Mathematical computation
Development	Infantile & not tested fully	Deployed and tested
Existing Infrastructure	Sophisticated	Widely used
Digital Signature	Not present	Present
Bit rate	1Mbit/s avg.[10]	Depend on Computing power
Cost	Crypto chip €100,000[11]	Almost zero
Register storage (n bit) at any moment	one n-bit string	2^n n-bit strings
Communication Range	10 miles max.[9]	Million of miles
Requirements	Devoted h/w & communication. lines	S/w and portable
Life expectancy	No change as based on physics laws	Require changes as computing power increases
Medium	Dependent	Independent

Conclusion

Quantum cryptography is based on mixture of concepts from quantum physics and information theory. The security standard in QC is based on theorems in classical information theory and on the Heisenberg's uncertainty principle. Experiments have demonstrated that keys can be exchanged over distances of a few miles at low bit rate. Its combination with classical secret key cryptographic algorithms permits increasing the confidentiality of data transmissions to an extraordinary high level. From comparison, it's obvious that quantum cryptography (QC) is having more advantage than Classical Cryptography (CC) though some issues are yet to be solved. This is mainly due to the implementation problems but in future there exist possibilities that most of the problems in quantum cryptography will get resolved.

Challenges & Future Direction

In the future, enhancing the performance of practical QKD systems and further improvements, both in key rate and secure transmission distance, are necessary for some applications. Another vital point is that, in real life, that is, quantum signals may share the channel with regular classical signals. The final goal is to achieve a client affable QKD system that can be effortlessly included in the Internet. To achieve a higher QKD key rate, one can consider other QKD protocols. Continuous variable QKD is projected to get a higher key rate in the small and medium transmission distance. Still, the scalability is a big challenge, as no one knows how to build a large scale quantum computer, which is interesting subject to be worked out.

References

[1] C. Bennett and G. Brassard, in Proceedings of IEEE, International Conference on Computers, Systems.

[2] Hughes, Richard J., D.M. Alde, P. Dyer, G.G. Luther, G.L. Morgan, and M. Schauer, Quantum cryptography, Contemporary Physics, Vol. 36, No. 3 (1995).

[3] Applied Cryptography, Second Edition: Protocols, Algorithms, and Source Code in C (cloth) Author(s): Bruce Schneier.

[4] FIPS. 46-3, '"Data Encryption Standard," Federal Information Processing Standard (FIPS), Publication 46-3, National Bureau of Standards, US. Department of Commerce, Washington D.C., October 25, 1999.

[5] R.L. Rivest, "Dr. Ron Rivest on the Difficulty of Factoring," Cipher text: The RSA Newsletter, v. 1, n. 1, fall 1993, pp. 6-8.

[6] G. R. Blakley, "One Time Pads Are Key Safeguarding Schemes, Not Cryptosystems Fast Key Safeguarding Schemes (Threshold Schemes) Exist." , Proceedings of the 1980 IEEE Symposium on Security and Privacy, 1980, pp. 108-113.

[7] Charles H. Bennett, Gilles Brassard, and Artur K. Ekert "Quantum Cryptography", Scientific American 267:4, (October 1992).

[8] Einstein, A., B. Podolsky, N. Rosen, Can quantum,

mechanical description of physical reality be considered complete?, Phys. Rev. 47, 777 (1935).

[9] Quantum communication transmission experiments http://www.waybeta.com/news/10441/quotquantum-communication-transmissionquot-experiments-in-china-_-the-success-of-huawei-wireless-network-card-news/

[10] **Cambridge Lab of Toshiba** http://www.physorg.com/news191010509.html

[11] Affordable Quantum Cryptography http://www.siemens.com/innovation/apps/pof_microsite/_pof-spring-2009/_html_en/interview-christian-monyk.html

A Neural Controller for Electron Beam Welding Power Supply Unit

[1]Jagannath Malik, [2]Anil Kumar, [3]Pravanjan Malik and [4]M.L. Mascarenhas

[1]Department of Electronics and Communication Engineering, Indian Institute of Technology, Roorkee, Roorkee-247667, India
E-mail: jags.mallick@gmail.com
[2]Department of Electrical Engineering, Indian Institute of Technology Roorkee, Roorkee-247667, India
[3]Laser and Plasma Division, Bhaba Atomic Research Centre, Mumbai, India
[4]Laser and Plasma division, Bhaba Atomic Research Centre, Mumbai, India

Abstract
Welding is an unavoidable unit practically for every manufacturing industry. Electron Beam welding (EBW) are very important unit of some specific manufacturing processes where high degree of accuracy and flawless welding is highly desirable like aerospace engineering. The power supply unit (PSU) used for EBW are very important unit, for which high degree of stability is a must. Since EBW is absolutely non-linear system, for better performance, adopting non-linear control methods could be a good solution. In the current study a robust adaptive controller based on multi-layer feed-forward neural network is developed for real-time voltage regulation. Simulation shows a better characteristic in terms of maximum overshoot and maximum undershoot for the neural controller compared to that of a conventional PI controller based on Ziegler Nichols [1,2,3,4] frequency response tuning method. The controller has the unique advantages of nonlinear mapping and adaptive learning.

Index Terms: Artificial Neural Network (ANN), Error back propagation algorithm, EBW.

Introduction

Electron *Beam Welding (EBW)* is widely used in industry, most notable for manufacture of aero engines. The equipment consists of a vacuum chamber, work piece manipulator and electron gun. The gun typically operated in the range of few kilovolts to several hundreds of kilovolts, which will accelerate the electron beam to increase the kinetic energy. In the advancement of the high voltage solid state device SMPS are used to produce such a high voltage for electron gun. The primary design requirements for the power supply are

- To produce a stable, low ripple accelerating voltage (DC) at the gun
- To be able to adjust that voltage over the working range
- To maintain the voltage from virtually no load to full load
- To reduce the stored energy at output filter elements

EBW system is a complex nonlinear system due to flash over and sparks occurs inside welding chamber. Due to its strong nonlinear behavior, the problem of identification and control of EBW power supply is always a challenging task for control systems engineer. Usually in the industries EBW power supplies are controlled using linear PI control configurations and the tuning of controller parameters is based on the Linearization of the models of the PSU in a small neighborhood around the stationary operating points. If the process is subjected to larger disturbance or it operates at conditions of higher state sensitivity, the state trajectory can considerably deviate from the aforementioned neighborhood and consequently deteriorates the performance of the controller.

PSU design for operation in the area of its higher state sensitivity and in some cases at the borders of its stability is important, even in the vicinity of an unstable stationary point that might have induced the periodic oscillations. As a result, the nonlinear nature of the PSU acquires more relevance in control systems and creates difficult control problems. If severe nonlinearity is involved in the controlled process, a nonlinear control scheme will be more useful. Nowadays, neural networks have been proved to be a promising approach to solve complex nonlinear control problems.

The use of neural networks in EBW field offers potentially effective means of handling three difficult problems: Complexity, non-linearity and uncertainties. The variety of available neural network architectures permits us to deal with a wide range of process control problems in comparison to other empirical models. Neural networks are relatively less sensitive to noise and incomplete information and deal with higher levels of uncertainty when applied in process control problems. The multilayer feed forward neural networks offer interesting possibilities for modeling any nonlinear process without a priori knowledge. Thus, self-learning ability of neural networks eliminates the use of complex and difficult mathematical analyses, which is dominant in traditional modeling methods.

A DC–DC converter is an integral part of EBW PSU, where low voltage unregulated DC is regulated by a high frequency DC-DC converter. Then regulated DC is converted to quasi square wave AC by means of an inverter and stepped UP to high voltage by a high frequency transformer. Finally,

high frequency high voltage squire wave is rectified and filtered to high voltage regulated DC. Pulse- width modulation (PWM) is often employed to control the DC output voltage by modulating the duty cycle via electronic switching circuits .To improve the power efficiency, many different switching circuit topologies have been proposed [9-14]. In conventional controller design, it is assumed that all the circuit components are ideal with no performance degradation and power loss and the circuit is operated at a stable bias point so that it can be modeled by a set of linear equations. However, in practice, the switching network is highly nonlinear and an accurate mathematical model is very difficult to obtain. In addition, the supply voltage and load current may also fluctuate over a wide range. Thus, real-time adaptive control is necessary to improve the system performance. Recently, artificial neural networks (ANN) have been applied to improve the performance of DC–DC converters to dynamical system changes refer [8, 15]. However, no prior work has yet been reported to control EBW using a neural network approach. Simulation circuit used to test the performance of the EBW is shown in (Fig.1). The block diagram of the EBW with controller is shown in (fig.2).

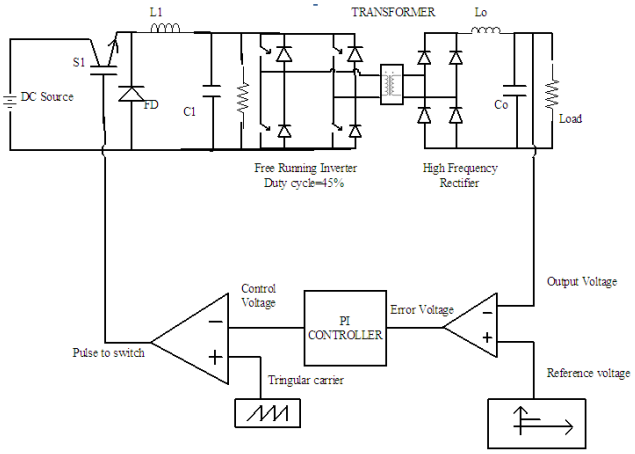


Fig. 1. Simulation circuit, test the performance of the EBW

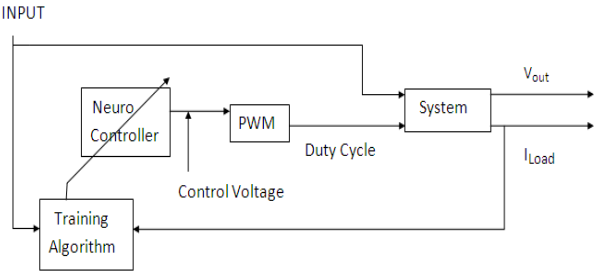


Fig. 2. System block diagram showing Neural controller

Error Back-Propagation Neural Networks and Back-Propagation algorithms

Artificial neural network [5-7] is usually defined as a network composed of a large number of processing units (neurons) that are massively inter connected, operate in parallel and learn from experience (examples). In recent years, ANN has become more popular because of its ease of operation moreover its accuracy in prediction can be improved through a good training process. The robustness of a well-trained network is excellent.

Back-Propagation Neural Networks are a kind of widely used Neural Networks nowadays. It is a multilayer feed-forward neural based on error back propagation algorithm [16, 17]. The model of BP Neural Networks is composed of three parts, input layer, output layer and hidden layer. Sigmoid-type function is usually used in the neurons of hidden layer. Activation function for the output layer depends on the output nature of the concern problem. . Calculation accuracy must be concerned to determine the quantity of hidden layer. The standard BP Neural Networks adopt step-transform based on Widrow-Hoff rule. The following aspects should be considered when designing the BP Neural Networks, the layer-number of the networks, the number of the nerve cell in each layer and activation function, initial value, learning rate.

BP algorithm is composed of two parts, forward transfer of the working signal and reverse transfer of the error signal. The mean square error F(x) is used as the performance function in multi-layer networks BP algorithm. The input of the algorithm is the combining of the sample with correct network behaviour {pl, tl}, {p2, t2}... {pq, tq}. Once a sample is input, network output is compared to target outputs, and the weights of the networks are adjusted by BP algorithm to minimize the mean square error F(x).

$$F(x) = E[e^T e] = E[(t-y)^T (t-y)] \quad (1)$$

In this formula, x is the vector of weights in the networks, and y is the output vector. $\hat{F}(x)$ is used to calculate mean square error approximately, replacing the expected value of the mean square error with the mean square error after k-th iteration.

$$F(x) = [t(k) - y(k)]^T [t(k) - y(k)] = e^T(k)\, e(k) \quad (2)$$

The steepest descent method to approximate mean square error is

$$W_{i,j}^m(k+1) = W_{i,j}^m(k) - \alpha \frac{d\hat{F}}{dW_{i,j}^m(k)} \quad (3)$$

In these formulas, α is the learning rate. $W_{i,j}^m$ is the weight value of the j-th nerve cell is for the i-th nerve cell on the m-th element. Theoretically, the model of the BP Neural Networks precisely depends on the training data provided. So the structure of the Neural Networks should be selected reasonable, and obtaining the model of high precision with few training data.

Neural Network model for the Controller

A three layer neural network (Fig.3) was designed with 10 numbers of hidden neurons. The input layer contains 2 neurons corresponding to two input parameters supply voltage and instantaneous output load current. A single output neuron at the network output corresponding to the control voltage. Uni-polar sigmoid function has been chosen as the activation function for both hidden layer and output layer neurons. The

training and testing data were taken from the simulation model with PI as the controller. Data set were normalized to train the network in the non-linear region of the sigmoid function avoiding the saturation area. An adaptive learning rate was chosen with initial value 0.9 and decremented by a factor of 10% when stuck in the local minima. Initial high value for learning rate was chosen due to the large error at the starting of training. A small momentum value of 0.2 was chosen. Mean absolute error (MAE) was chosen as the performance criteria for the network and training was done with an iteration limited approach.

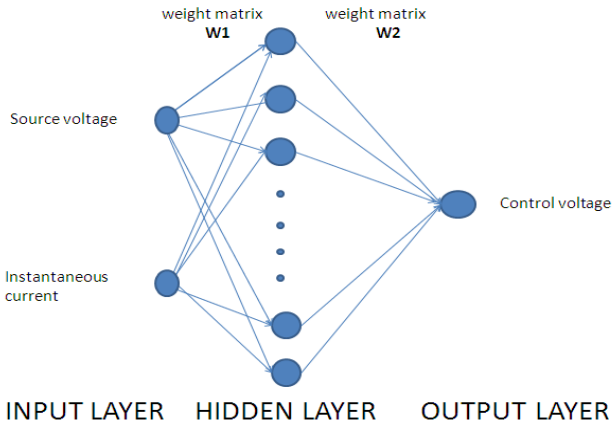


Fig. 3. Neural Network Architecture

Simulation Results

MATLAB Simulink models were developed for both PI controller based on Ziegler Nichols frequency response tuning method and the neural controller. Current and voltage waveforms were traced. Performance of the neural controller is compared with that of PI controller. Other performance parameters have been calculated (Table I) to verify the better characteristics of neural controller in comparison to conventional PI controller. Fig. 4 shows comparison between output voltage for PI and Neural controller. Fig. 5-6 shows output current waveforms for PI and Neural controllers respectively.

Table I: Units for Magnetic Properties

Performance Parameters	PI Controller	Neural Controller
%Over Shoot	15	4.95
%Under Shoot	2.245	1.4
Delay Time (T_d) (10^{-4}s)	1.31	1.4
Rise Time (T_r) (10^{-4}s)	3.25	4.25
Settling Time (T_s) (10^{-3}s)	1.16	1.4
Average Integral Absolute Error (AIAE)	0.0773	0.0758
Integral Squared Error (ISE)	1.5252×10^4	1.119×10^4
Average Integral Squared Error (AISE)	0.1768	0.1138

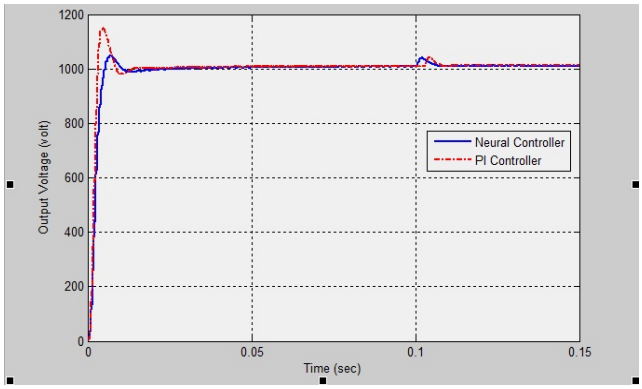


Fig. 4. Output Voltage for both PI and Neural Controller

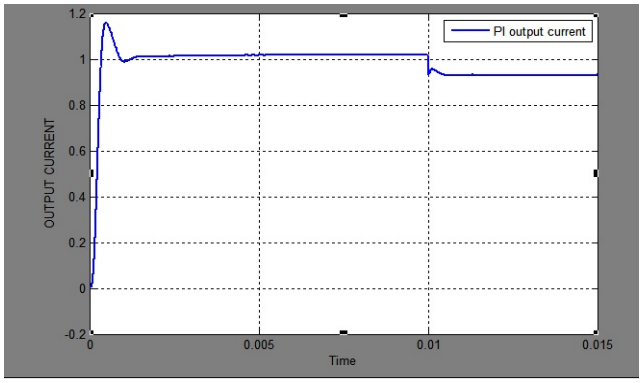


Fig. 5. Output Current for PI controller

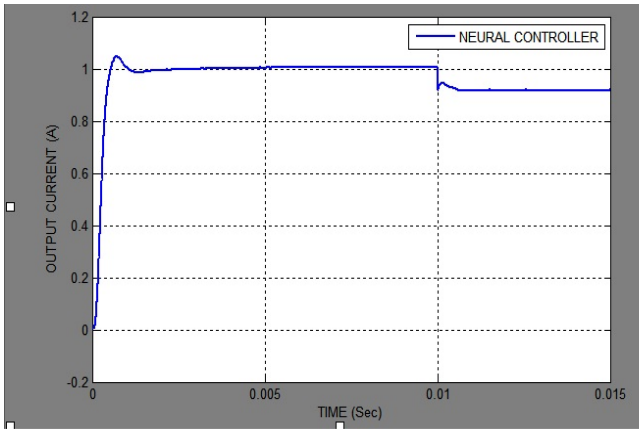


Fig. 6. Output Current for Neural controller

Summary and Conclusions

In this paper the proposed neural network controller gives a better performance in comparison to the PI controller in terms of maximum overshoot and maximum undershoot fig.4. Other factors like AIAE, ISE and AISE also verify the better performance of neural controller over the PI controller.

References

[1] Ziegler JG, Nichols NB. "Optimum settings for automatic controllers" *ASME Trans* 1942; 64:75968.

[2] Ang KH, Chong G, Li Y. "PID control system analysis, design, and technology" *IEEE Trans Control Syst Technol* 2005;13(4):55976

[3] Hang CC, Åström KJ, Ho WK. "Refinements of ZieglerNichols tuning formula". *IEE Proc-D* 1991;138(2):1118.

[4] Mudi RK, Dey C, Lee TT. "An improved auto-tuning scheme for PI controllers". *ISA Trans* 2008;47:4552.

[5] S. Haykin, "Neural Networks: A Comprehensive Foundation", *Prentice-Hall, Englewood Cliffs, NJ,* 1999.

[6] E. Karnin, "A simple procedure for pruning back-propagation trained neural networks", *IEEE Transactions on Neural Networks* 1 (2) (1990) 239242.

[7] P.K. Simpson, "Artificial Neural Systems" *Pergmon Press Elmsford, New York*, 1989.

[8] F. Kamran, R.G. Harley, B. Burton, T.G. Habetler, M.A. Brooke, "A fast on-line neural-network training algorithm for a rectifier regulator", *IEEE Transactions on Power Electronics* 13 (2) (1998) 366–371.

[9] J.M. Carrasco, E. Galva´ n, G.E. Valderrama, R. Ortega, "A. Stankovic, Analysis and experimentation of nonlinear adaptive controllers for the series resonant converter", *IEEE Transactions on Power Electronics* 15 (3) (2000) 536–544.

[10] P.R. Chetty, "Resonant power supplies: their history and status", *IEEE Aerospace and Electronic Systems Magazine* 7 (4) (1992) 23–29.

[11] H.-S. Choi, B.H. Cho, "Novel zero-current-switching (ZCS) PWM switch cell minimizing additional conduction loss", *IEEE Transactions on Industrial Electronics* 49 (1) (2002) 165–172.

[12] H.-S. Choi, J.-W. Kim, B.H. Cho, "Novel zero-voltage and zero-current-switching (ZVZCS) full-bridge PWM converter using coupled output inductor", *IEEE Transactions on Power Electronics* 17 (5) (2002) 641–648.

[13] M.G. Kim, M.J. Youn, "An energy feedback control of series resonant converters", *IEEE Transactions on Power Electronics* 6 (3) (1991) 338–345.

[14] X. Ruan, Y. Yan, "A novel zero-voltage and zero-current-switching PWM full-bridge converter using two diodes in series with the lagging leg", *IEEE Transactions on Industrial Electronics* 48 (4) (2001) 777–785.

[15] J.M. Quero, J.M. Carrasco, L.G. Franquelo, "Implementation of a neural controller for the series resonant converter", *IEEE Transactions on Industrial Electronics* 49 (3) (2002) 628–639.

[16] M. Minsky, & Papert, S., "Perceptrons: An Introduction to Computational Geometry," *The MIT Press*, pp. 3, 26, 31, 33, (1969).

[17] D. B. Parker, "Learning-Logic" (Tech. Rep. Nos. TR {47)." 1985.

Authors Biography

Jagannath Malik is pursuing Integrated Dual Degree (Bachelors/ Masters) in Electronics & Computer Engineering with specialization in Wireless Communication at Indian Institute of Technology (IIT) Roorkee, India. His research interest includes soft computing techniques, artificial neural networks (ANNs), optimization algorithms, image processing, millimeter-wave engineering, metamaterial, microstrip antennas for communications, RF and microwave designs. He has published a number of papers in the fields of ANN and microstrip antennas.

Anil Kumar is currently working as Junior Research Fellow at Biomedical Signal Processing Lab, Indian Institute of Technology Roorkee, India. His research interest includes Machine Learning, Biomedical Signal and Image Processing, Autonomous Robotics, Computer Vision, Computer Aided Instrumentation and Software Application Development. He has successfully completed major projects sponsored by IRDE Defense Research and Development Organization (DRDO India), Rajasthan Electronics and Instruments Ltd. (India) and Ministry of Human Resource and Development (MHRD) India.

Online EEG Experiment using Virtual Labs Architecture

Anil Kumar, Jagannath Malik, Aditya Kotwal and Vinod Kumar

Department of Electrical Engineering, Indian Institute of Technology Roorkee, Roorkee-247667, India
E-mail: anilinfotek@gmail.com
Department of Electronics and Communication Engineering, Indian Institute of Technology, Roorkee-247667, India
E-mail: jags.mallick@gmail.com
Department of Electrical Engineering, Indian Institute of Technology Roorkee, Roorkee-247667, India
E-mail: kotwal13aditya@gmail.com
Department of Electrical Engineering, Indian Institute of Technology Roorkee, Roorkee-247667, India
E-mail: vinodfee@iitr.ernet.in

Abstract
Electroencephalography (EEG) is a non-invasive recording of electrical activity of brain on scalp. Because of its uses in clinical diagnosis and conventional cognitive neuroscience, EEG is an important subject for medical students and biomedical engineers. Availability of a standard acquisition device and data is major bottleneck experimental learning which restrains students from correlating their theoretical knowledge with real world problem. As a solution, this paper presents an online EEG system based on virtual labs architecture for remote experimentation. The presented system provides undergraduate and postgraduate students with standard EEG data and helps to learn practical aspects of electroencephalography i.e. working of an EEG machine, impact of various signal conditioning techniques on acquired signal, and to observe EEG waveform of a patient in various conditions.

Index Terms: Virtual Labs, NI LabVIEW™ 2010, EEG signal analysis.

Introduction

In most of the developing countries, lack of resources necessary for practical education is of the major challenges faced by technical education. In such countries most of educational institutions do not have access to sophisticated industrial instruments; this absence of practical experience hinders the overall techno-intellectual growth of students. This unavailability of sufficient quantity of resources avoids students from learning together and sharing experiences which is not possible through theoretical knowledge available through books alone. Solution to such problem lies in development of systems facilitating hardware sharing and remote access to educational resources through virtual instrumentation. For best utilization of these resources, such virtual systems should allow simultaneous multiple accesses to remote users without disturbing each other's work. One such solution has been proposed in this paper where parallel access to an instrument is possible.

The proposed system is aimed to provide an actual lab like environment where a student will be able to operate a biomedical instrument, record and analyse live EEG data and then make proper report of the observation.

Virtual Labs Architecture

The presented online EEG experimentation system is based on Biomedical Virtual Lab architecture originally proposed by Kumar *et al.* [1]. The proposed system works in two parts named:

- Server Application
- Client Application.

Server Application

This part of system is operated in Biomedical Instrumentation Laboratory at IIT Roorkee. The server system connected to Biopac™ MP150 EEG acquisition system [2] and Grass Technologies 32 channel EEG simulator [3]. This in IIT Roorkee system continuously publishes live EEG multi-channel signal acquired from the EEG simulator at a desired sampling rate. This application provides access of all laboratory resources and central information system. To allow remote users to observe hardware response to the online commands, the system facilitates live video streaming of the lab through IP camera. To avoid any conflict between simultaneous users the system facilitates online time slot booking as per choice cum availability of time. This way only at a time one user (referred as 'active user') having time slot booked can manipulate hardware and acquisition setting, whereas the data is available to all users and hence all users can perform experiments independently. The system hardware (Fig. 1) consists of:

- Biopac™ MP150 20-electrode EEG acquisition system using 10 EEG100C amplifiers.
- Grass technologies 32 channel EEG Simulator
- Dlink™ DCS-5220 IP surveillance Camera.
- Server Workstation (Dual Intel Xeon 2.27 GHz 4 cores processors, 4 GB memory)

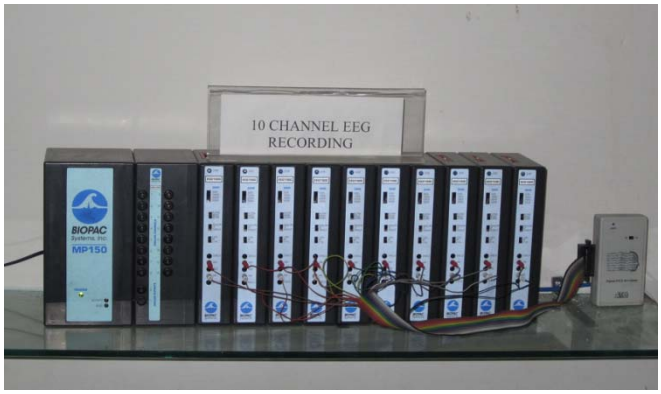


Fig. 1: Biopac™ MP150 biomedical signal acquisition system connected to EEG Simulator.

Client Application

The client application works at remote user end and connects remote users with Server Application. It provides various functionalities like slot booking, data acquisition and control, recording, analysis of EEG signal and report generation of the observation. To ensure proper learning of student, relevant theory and literature regarding EEG provided with experimentation package. After experimentation the users are offered a questionnaire related to the experiment performed to test their learning. Finally after performing the experiment, the system generates experiment report showing the data waveform, experiment analysis and questionnaire results.

System Implementation

Server Operation

Similar to the server system developed by Kumar *et al.* [1], the server for this system also works in 3 states viz.

- Home State
- Admin State
- Data Server State.

Fig. 2 shows the possible state transitions soring operation of server application. The home state facilitates an authenticated server administrator to Add/Edit time slot booking for remote users. Under Admin State, the administrator can exercise administrative rights like:

- Adding/Deleting remote user accounts
- Generating statistical logon report of remote user on virtual lab server
- Communicating remote user by sending email
- Blocking remote users from connecting to server

During the data server state, the server application acquires 20 electrode data out of available 32 electrode ports available in Simulator. The data acquisition on simulator follows internationally 10-20 EEG system and form 10 analog channels in bipolar (differential) acquisition mode. The data publishing and control transfer takes place using NI DataSockets™ based variable sharing [4]. Fig 3 shows front panel of data server state of EEG server.

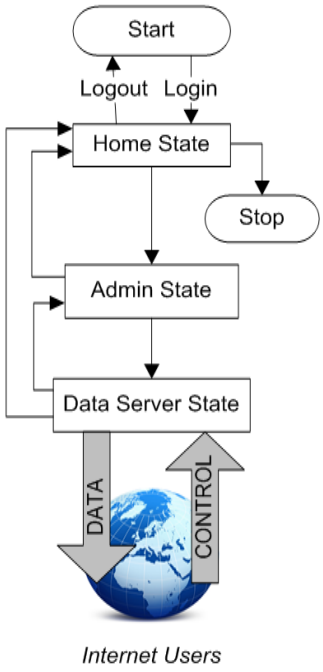


Fig. 2: State transition diagram for server application.

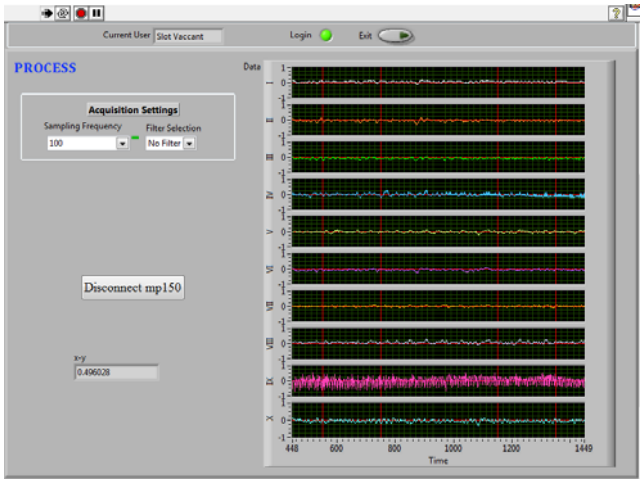


Fig.3: Server Application front panel in Data Server State.

Client Operation

The client application is also based on a state model similar to server application. It works in six working states namely

- Home State
- Data Acquisition State
- Data Review State
- EEG Analysis State.
- Evaluation State
- Report Generation State

Out of these operation states, Home state, Data Review state, Evaluation state and Report Generation state have been adopted from the base architecture [1]. Therefore this paper discusses Data Acquisition State and EEG Analysis in detail. Home state is the first operation state encountered to Client

Application and it automatically tries to contact server application and retrieves login information. It allows remote users to login to virtual labs system and book a time slot in central database to work as active user.

Data Acquisition State provides two modes of EEG data acquisition for experimentation and analysis i.e. online (from virtual lab server) and offline (local system resources). In online mode Active users have access to change acquisition settings in terms hardware sampling frequency. Client application allows users to acquire data at four pre-set sampling frequencies viz. 100Hz, 200Hz, 500Hz and 1000 Hz.

Users can select the acquisition depending on analysis and internet resource availability and record the data for desired length (up to 300 s) for analysis. The system saves the recoded data on local computer in form text file and initiates its playback for user's approval. Fig. 4 shows from client user interface in online data acquisition mode.

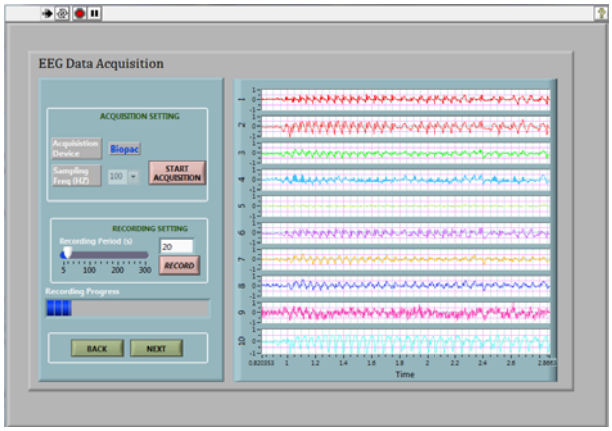


Fig.4: Front panel of EEG data acquisition

The EEG analysis state has been deigned to help students to learn variation in EEG in spectral and statistical domain with different neural activities. During analysis student also get opportunity to verify the processing by testing on standard signals comparing with the results obtained through acquired EEG data. Fig. 5 shows front panel of the EEG analysis mode in client application. Following EEG analysis tools are provided to remote users:

Fig.5: Front panel of EEG analysis

Spectral Analysis

Since different neural activities appear in EEG in different frequency range and at different locations. The users can select any EEG channel and can obtain following standard EEG wave components:

- Alpha wave (8-13 Hz)
- Beta wave (30-100 Hz)
- Gamma wave (13-30 Hz)
- Delta wave(0-4 Hz)
- Theta wave(4-8 Hz)

Apart from these standard wave components, used have freedom to design a filter of their desired specification and use it obtaining custom EEG component wave. The available filter types are

- Butterworth Filters
- Chebyshev Filter
- Elliptic Filter
- Bessel filter

Users can design a filter of desired order and pass band and see the filtered output in available mixed graph.

Statistical Analysis

Apart from spectral features, statistical features also vary with neural activities in brain. The proposed system allows users to apply statistical filters like Standard Deviation, Median, Variance, Local Maximum, Local Minimum and Local Mean on selected EEG channel in a moving window mode. Users can select any window size of their desire and observe the statistical variations in EEG.

All the analysis done by users is automatically forwarded to an inbuilt experiment reporting mechanism. After analysis, users are offered a questionnaire of ten randomly picked multiple answer type questions. Since the system has been designed to cater needs of educationalists and teachers in teaching and training, questionnaire helps in accessing practical learning of students. Finally after experimentation the experimentation report (Fig. 6) is made available to users which can be easily printed as record.

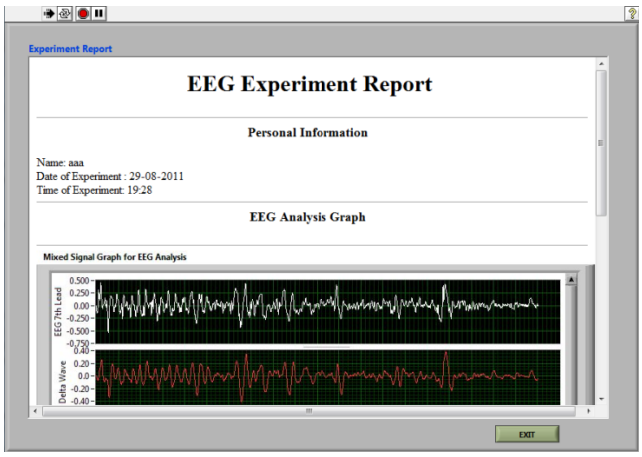


Fig.6: Report generated after performing experiment

System Implementation

Since lab experimentation need high end Graphic User Interface (GUI) and parallel processing capabilities, NI LabVIEW™ has been used to develop this experimentation architecture and virtual instruments. The client application is free available for use can be downloaded from http://210.212.58.232/vlab after registration from the same website. The necessary documentation and technical theory regarding the experiment is provided on website to support students in correlate theoretical knowledge with practical experience. The application works well on Microsoft Windows™ (XP or newer) operating system.

Conclusion

The online experimentation system presented is paper proposes a highly cost effective means for promoting Distance Education. Since only a computer and internet connection is required for perform experiments over virtual labs, the presented system offers a facility to perform experiments anywhere and anytime and hence will revolutionize the technical education system. Attempts are being made to make EEG simulator computer controlled so that students can change the EEG data on their own.

Acknowledgment

Authors are thankful to MHRD and IIT Roorkee for providing financial assistance and laboratory infrastructure to develop the virtual laboratory.

References

[1] A. Kumar, J. Malik, V. Kumar, "Virtual Lab: Real-time Acquisition and Analysis of ECG Signal", in *International Journal of Online Engineering*, Vol 7, No 3 (2011), pp. 19–23.

[2] Grass technology product page. Available Online: http://www.grasstechnologies.com/products/clinsystems/accessories.html

[3] Biopac MP150 product documentation site. Available Online: http://www.biopac.com/data-acquisition-analysis-system-mp150-system-windows

[4] National Instruments Datasockets documentation. Available Online: http://www.ni.com/datasocket/

Authors Biography

Anil Kumar completed his Bachelor's degree in Electrical Engineering from Indian Institute of Technology (IIT) Roorkee and is working as Junior Research Fellow at Biomedical Signal Processing Lab, IIT Roorkee, India. His research interest includes Machine Learning, Biomedical Signal and Image Processing, Autonomous Robotics, Computer Vision, Computer Aided Instrumentation and Software Application Development. He has successfully completed major projects sponsored by IRDE Defense Research and Development Organization (DRDO India), Rajasthan Electronics and Instruments Ltd. (India) and Ministry of Human Resource and Development (MHRD) India.

Jagannath Malik is pursuing Integrated Dual Degree (Bachelors/ Masters) in Electronics & Computer Engineering with specialization in Wireless Communication at IIT Roorkee, India. His research interest includes soft computing techniques, artificial neural networks (ANNs), optimization algorithms, image processing, millimeter-wave engineering, metamaterial, microstrip antennas for communications, RF and microwave designs. He has published a number of papers in the fields of ANN and microstrip antennas.

Adtiya Kotwal is pursuing Bachelor's degree in Electrical Engineering from IIT Roorkee. He has research interests in virtual instrumentation, power electronics and robotics. He has also done research on material characterisation and biological processing using ultrafast femtosecond lasers in the field of nonlinear optics.

Vinod Kumar is Professor and head of Electrical Engineering Department, Indian Institute of Technology Roorkee, India. He is also head of Continuing Education Centre and Quality Improvement Program Centre of IIT Roorkee. He received his both Masters and PhD degree from IIT Roorkee (erstwhile University of Roorkee). He has many academic awards, distinctions and scholarships and more than 150 research papers to his credit. He has 34 years of rich experience of teaching & research. He is a life fellow of IETE and is a senior member of IEEE. His areas of interest are Biomedical Signal and Image Processing, Pattern Recognition, Medical Instrumentation.

Analysis of the Variants of Watershed Algorithm as a Segmentation Technique in Image Processing

Namrata Puri and Sumit Kaushik

[1]Research scholar, [2]Assistant Professor
Ambala College of Engineering & Applied Research, Devsthatli, Ambala Cantt, India
E-mail:namratapuri12@yahoo.in,sumitkaushik24@gmail.com

Abstract
The watershed transform is a popular image segmentation algorithm for grey scale images. It is the method of choice for image segmentation in the field of mathematical morphology. Watershed segmentation is based on sets of neighboring pixels. We present a critical review of several definitions of the watershed transform and the associated sequential algorithms, immersion models and therefore parallel implementation of these immersion and sequential models. In this paper, procedure regarding performance analysis of these three variants is drawn and further the OpenCV tool is used to calculate the final results.

Keywords: Watershed transform, Mathematical morphology, Immersion, OpenCV

Introduction

Image processing is defined as a technique in which the data from an image are digitized and various mathematical operations are applied to the data, generally with a digital computer, in order to create an enhanced image that is more useful or pleasing to a human observer, or to perform some of the interpretation and recognition tasks usually performed by humans. Also known as picture processing[7].

In computer vision, segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels, also known as superpixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image[6].

The watershed transformation considers the gradient magnitude of an image as a topographic surface. Pixels having the highest gradient magnitude intensities (GMIs) correspond to watershed lines, which represent the region boundaries. Water placed on any pixel enclosed by a common watershed line flows downhill to a common local intensity minimum (LIM). Pixels draining to a common minimum form a catch basin, which represents a segment.

Generally spoken, image segmentation is the process of isolating objects in the image from the background, i.e., partitioning the image into disjoint regions, such that each region is homogeneous with respect to some property, such as grey value or texture [1].

The watershed transform can be classified as a region-based segmentation approach. The intuitive idea underlying this method comes from geography: it is that of a landscape or topographic relief which is flooded by water, watersheds being the divide lines of the domains of attraction of rain falling over the region [2].

An alternative approach is to imagine the landscape being immersed in a lake, with holes pierced in local minima. Basins (also called `*catchment basins*') will fill up with water starting at these local minima, and, at points where water coming from different basins would meet, dams are built. When the water level has reached the highest peak in the landscape, the process is stopped. As a result, the landscape is partitioned into regions or basins separated by dams, called *watershed lines* or simply *watersheds*[2].

When simulating this process for image segmentation, two approaches may be used: either one first finds basins, then watersheds by taking a set complement; or one computes a complete partition of the image into basins, and subsequently finds the watersheds by boundary detection.

To be more explicit, we will use the expression `watershed transform' to denote a labelling of the image, such that all points of a given catchment basin have the same unique label, and a special label, distinct from all the labels of the catchment basins, is assigned to all points of the watersheds.

We note in passing that in practice one often does not apply the watershed transform to the original image, but to its (morphological) gradient [3]. This produces watersheds at the points of grey value discontinuity, as is commonly desired in image segmentation.

The most frequently used definition of the watershedding operation follows a geographical analogy (Fig. 1). If a grayscale image is viewed as if high intensity colors were high ground then the image becomes a 3D landscape.

The catchment basin of this minimum is the area, where water falling on the landscape would flow down to the minimum. The watershed of the image is the set of lines (dams) that separate the catchment basins on the image. The "height" in topographic surface may be any measurable property of image pixel: lightness, gradient of lightness,

saturation or other. That makes watershed algorithm useful for color image processing.

The watershed transformation performs very accurate segmentation, which is beneficial in case when objects overlap and their borders are hardly detectable[3].

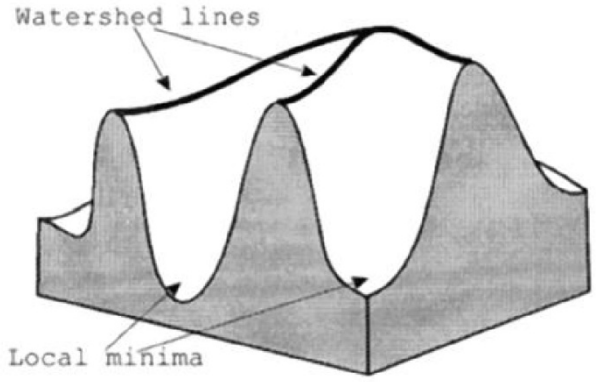


Fig 1. Local minima of a surface

Literature Survey

Generally spoken, existing watershed algorithms either simulate the flooding process, or directly detect the watershed points. In some implementations, one computes basins which touch, i.e., no watershed pixels are generated at all[1].

Watershed algorithms by immersion

Vincent-Soille algorithm

An algorithmic definition of the watershed transform by simulated immersion was given by Vincent and Soille.

The immersion approach [6] is also referred to as the flooding analogy. In the immersion simulation, we first pierce a hole in every local minimum of the topographic surface formed by the gradient magnitude image. Then, we slowly immerse the topographic surface in water. Starting from the minima of lowest altitude, the water will progressively fill up all the different catchment basins. At some point, the rising water in any one specific basin will start to merge with water coming from its neighboring basins. Suppose that this merging can be prevented by constructing dams at the merging sites all the way up to the highest surface altitude (or until the immersion procedure ends). At the end of this immersion simulation, each basin will be completely surrounded by dams and the location of dams corresponds to watershed line.

Order-Invariant Immersion Algorithm

This section presents the details of our order-invariant immersion algorithm for image segmentation. Similar to the one proposed by Vincent and Soille in [6], this algorithm first requires a sorting of the pixels in the increasing order of their gradient magnitude values before running the level-by-level flooding step. It is this sorting step that has made the level-by-level flooding step efficient enough so that the Vincent and Soille algorithm can surpass its predecessors in computational efficiency.

Let G: D -> R^+ be a gradient magnitude image, where D is the indexing domain of the image (e.g., D = Z^2) and max(G) be the minimum value and the maximum value of G, respectively. By sorting the pixels of G in the increasing order of their gradient magnitude values, we can easily decompose D into a finite number of disjoint level sets, each denoted by

$$D_h = \{p \in D \mid G(p) = h\} \quad (1)$$

That is, we have $D = U_h D_h$,$\min(G) \le h \le \max(G)$, and $D_k \cap D_l$ if $k \neq l$. Different sorting techniques can be used here. For better efficiency, our algorithm uses counting sort if the data type of the gradient magnitude is fixed point, while it uses quick sort if the data type is floating point.

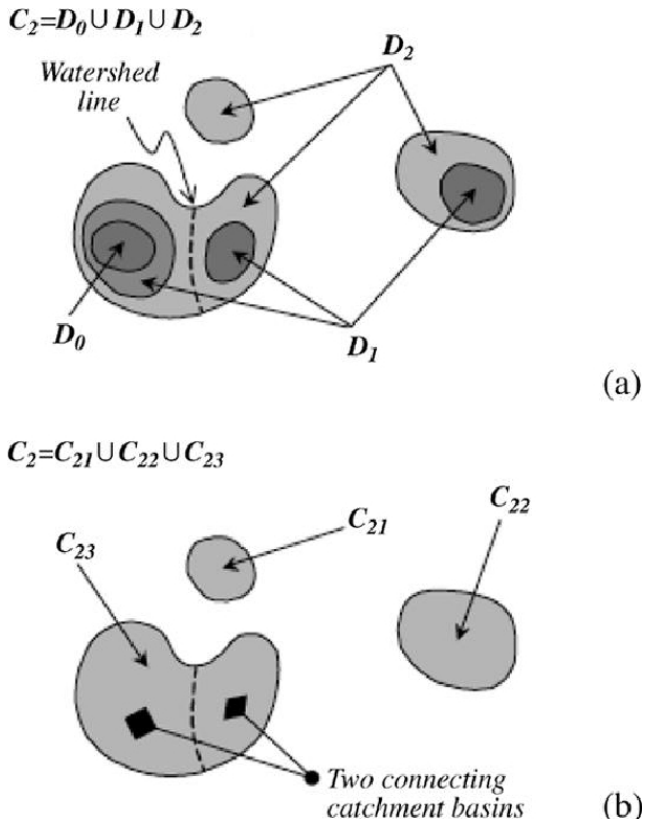


Fig 2. Disjoint level sets

If it is assumed that we have reached to level h-1 after the pixel sorting step during immersion process. C_{h-1} can also be viewed as a union of connected components, i.e. where each connected component contains one or more than one catchment basin. The catchment basin here can be referred to as a pre-h catchment basin because it is formed right before the water level rises up to level-h. For example, in Fig. 2(b) contains three connected components, denoted by C_{21},C_{22} ,and C_{23} . Note that C_{23} contains two pre-2catchment basins separated by a watershed line, while both others contains only one pre-2 catchment basin.

Next, by letting the water level goes up to level h, we have a new level set D_h , which can also be viewed as a union of connected components, i.e., . An example is given in Fig. 3

Here, the three different components are classified.

***Type-2 Component*:** More than one pre-h catchment basin is connected to this type of connected component. For example,

in Fig. 3 D31 is a type-2 component because it connects to three pre-2 catchment basins. Notice that actually contains two pre-2 catchment basins.

Type-1 Component: Exactly one pre-h catchment basin is connected to this type of connected component. For example, in Fig. 3 D32 is a type-1 component because it has only one pre-2 catchment basin, .

***Type-0 Component*:** No pre-h catchment basin is connected to this type of connected component. For example, in Fig. 3 D33 is a type-0 component.

Notice that when the flooding has been completed up to level h-1, every pixel having altitude less than or equal to h-1 will have already been assigned a unique catchment basin label.

To implement the flooding step, we divide the pixels in into the following three classes and label the pixels in each class one by one. It is worth mentioning that the classification of the connected components in is for understanding the algorithm, while the following classification of the individual pixels in is for implementing the algorithm.

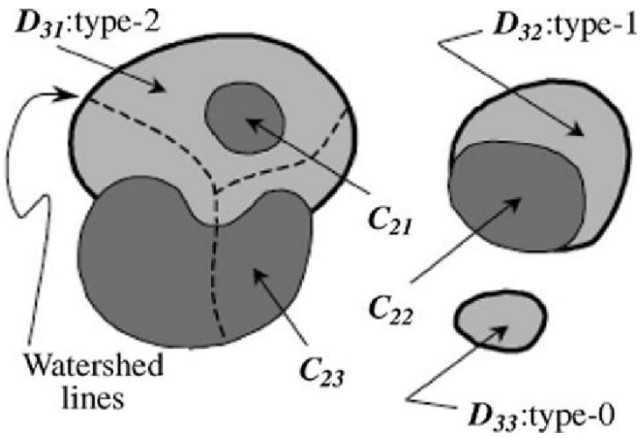


Fig 4. relation between sets and components

***Class-I Pixel*:** A pixel p in Dh is called a class-I pixel if its altitude is strictly greater than the altitude of its lowest neighbor.

Class-II Pixel: This class of pixels can be viewed as interior pixels of nonlocal-minimum plateaus. Fig. 4 shows some examples of class-II pixels.

***Class-III Pixel*:** Pixels in type-0 components of are class-III pixels. All the pixels in one type-0 component will be assigned a new and unique label.

`Algorithm 1. Immersion Approach`

Step 1. Sorting Step: Sort all the pixels in the gradient magnitude image G to obtain level sets D_h in increasing h.

Step 2. Flooding Step:

For each level set Dh , in the increasing order of h. Step 2.1.Simulate flooding for all the class-I pixels in Dh by labeling each class-I pixel with the label of its lowest neighbor . All these class-I pixels are pushed into a FIFO queue for region growing in

Step 2.2.

Step 2.2.Simulate flooding for all class-II pixels in Dh by region growing from class-I pixels using the FIFO queue initialized in Step 2.1.

Step 2.3. Simulate flooding for class-III pixels in Dh by assigning a new and unique label to each of the type-0 components in Dh.

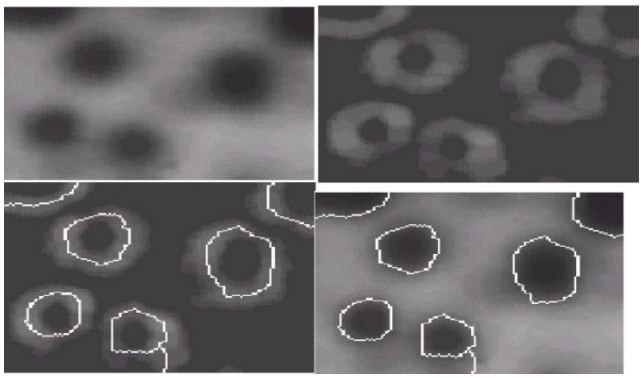


Fig 4. a) image of blobs b) image of gradient c) watershed lines d)watershed lines superimposed on original image.

Sequential Rainfalling Algorithm

The watershed transformation was widely and successfully applied in different domains (e.g., in biomedicine, industry, and, generally in computer vision applications) as a powerful segmentation tool. The idea behind it is to split the morphological gradient of an original image, seen as a topographic surface, into geodesic influence zones . Unlike the other methods the herein introduced algorithm has a higher degree of locality such that it can generate faster parallel implementations.

Description of the Method

The watershed algorithm based on rain falling simulation performs segmentation by labeling connected areas within the gradient of an image. Regarding the morphological gradient of the original image as a topographic surface, the rule of assigning labels can be derived from physics: a particle in free fall on a topographic surface will move due to gravity downward to the deepest neighboring location. On flat areas, the rule is overloaded, such that the motion of the particle is directed toward the nearest brim of a downward slope, or it stops if the particle has reached a regional minimum.

The task performed by the present algorithm is to trace a path for each non-minimum point on the surface (origin) to a minimum (destination), and to mark all pixels along the path with the label of the minimum. This path is a steepest slope line in a lower-complete image. The latter is the transformed gradient image such that any non-minimum pixel has a lower neighboring one. The result is a partition of the image which

has the following properties: regions are connected, they do not overlap, and the partition is complete.

Advantage of the Proposed Method over the Watershed by Immersion

An important advantage of this watershed algorithm is its suitability for parallel implementation. While immersion is a global method (water arising from many sources progressively floods all the surface and interactions between waters coming from different sources are taken into consideration), raining can be denoted as a local method because each droplet follows on its own way regardless of neighboring droplets.

Parallel Watershed algorithm

The computation of the watershed transform of a gray scale image is a relatively time consuming task and therefore usually one of the slow step in this chain. A common solution for such computationally expensive algorithms is to search for implicit parallelism in the algorithm and use this to implement the algorithm on a parallel computer.

As the watershed is sequential it is implemented in parallel by splitting the computation in three stages:

1. In the first stage of algorithm, input image is transformed into a directed components graph.

In the second stage of algorithm, the watershed of this graph is computed by breadth first coloring algorithm.

In the final stage, the flooded graph is transformed back into image domain.

If pixels € watershed nodes, then pixels are colored white and all pixels€ non- watershed nodes are colored black.

Watersheds are "thick" and thinning is done by skeletonization.

Problem Formulation

In order to analyze the different variants of watershed transform, it is necessary to implement the different variants such as immersion and sequential techniques and then compare the results.

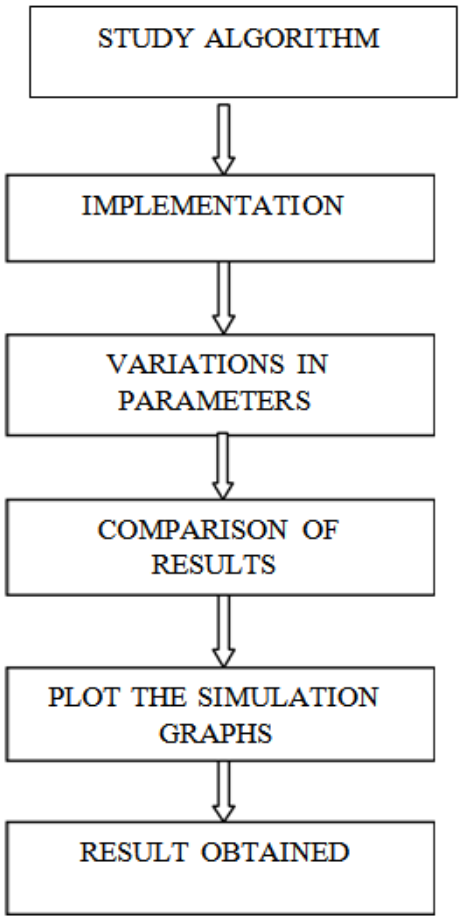
Problem in Mathematical Form

Let us suppose we have set of n watershed algorithms denoted by array W={w1,w2,……wn}

Using size of the image when taken as input as a parameter for comparison, we will find out w_i €W such that 1<i<n in order to find out the effect of increasing the size on w_i is minimum.

Proposed Model

We will do parallel implementation of the given algorithms and input the image of different sizes and then compare the outputs of different variants in order to plot simulation graphs of various outputs and hence it will be identified which algorithm gives the minimized effect of the image as a output.



The tool that will be used for analysis is open CV tool. OpenCV (Open Source Computer Vision Library) is a library of programming functions mainly aimed at real time computer vision.

Conclusion

In this paper, we have presented a definition of the watershed segmentation which is consistent with the behavior of most implementations of the watershed algorithm. All the algorithms described extend to the three dimensional case in a straightforward manner. The watershed algorithm by immersion is hard to parallelize because of its inherently sequential nature. A parallel implementation of this algorithm can be based upon a transformation to a component graph. After analyzing the various algorithms using graphs, we will find that varying the size of image, length of watershed line, intensity value of gradient image as parameters which variant of watershed has given the best performance.

Future Scope

As we have used size of image as a parameter, in future work length of watershed line, intensity value of gradient image, texture can be taken as parameter for performance analysis of variants of watershed transform. It is thus expected to contribute to new insights into the use of watersheds in the field of image analysis. In particular, more experiments are

currently being carried on to evaluate the interest on watershed of graphs with respect to picture segmentation.

References

[1] Meyer, F., and Beucher, S. Morphological segmentation. J. Visual Commun. and Image Repres. 1,(1990).

[2] Serra, J. Image Analysis and Mathematical Morphology. Academic Press,New York, 1982.

[3] S. BEUCHER, C. LANTUEJOUL, Use of watersheds in contour detection. International Workshop on image processing,real-time edge and motion detection/estimation, Rennes, France, Sept. 1979.

[4] Oversegmentation avoidance in watershed-based algorithms for color images Wojciech Bieniecki.

[5] Haralick, R. M., and Shapiro, L. G. Survey : image segmentation techniques. Comp. Vis. Graph

[6] Im. Proc. 29 (1985), 100-132.

[7] http://en.wikipedia.org/wiki/Segmentation_%28image_processing%29

[8] http://www.answers.com/topic/image-processing

[9] http://www.cs.toronto.edu/~jepson/csc2503/segmentation.pdf

[10] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 6, pp. 583–598, Jun. 1991.

[11] A.Meijister and J.B.T.M.Roerdink.proposal for the implememtation of a parallel watershed analysis. In :proceedings of computer analysis of images & Pattern(CAIP‟95),Springer Verlag,1995.

[12] http://en.wikipedia.org/wiki/OpenCV

[13] Digital image Processing 2nd ed.. R. C. Gonzalez & R. E. Woods,2002..

Optimization of Surface Reflectance for Alkaline Textured Monocrystalline Silicon Solar Cell

Charanpreet Sethi[#], Vijay Kumar Anand[#], Kiran Walia[#] and S. C. Sood[#]

#Ambala College of Engineering & Applied Research, Devasthli, Ambala 133101, Haryana, India
E-mail: c.sethi@rediffmail.com, ervijay2222@gmail.com, walia.kiran@gmail.com, soodace@gmail.com

Abstract
Surface texturization is well known as one of the major paths to improve the conversion efficiency of silicon solar cells by increasing the short-circuit current through the enhancement in antireflection property and effective photon trapping. Compared to the antireflection coating, it is more lasting and effective process. The anisotropic texturing of a (100) n-type monocrystalline silicon surface was performed using alkaline etching solution of potassium hydroxide (KOH) including isopropyl alcohol (IPA) additive. The reflection properties of alkaline-etched wafers were investigated using UV-VIS-NIR spectrophotometer and the images of the surface morphology were obtained using a scanning electron microscope (SEM). The influence of KOH/IPA concentration on etched wafers has been studied i.e. process variables considered were KOH & IPA concentration. An optimum value of surface reflectance has been achieved by exploring the better concentration of alkaline solution (KOH, IPA). Minimal reflectance of textured surface achieved was -0.83 % at wavelength of 800nm in the visible region.

Keywords: Monocrystalline silicon; Solar cell; Texturization; Alkaline solution; Surface reflectance

Introduction

Renewable resources affect our environment much lesser than conventional energy resources. These resources are clean and green. They are replenished naturally – which means that they will never run out. For example Solar cells obtain the energy from the sun, a free and inexhaustible source of fuel to produce emission-free electricity. The monocrystalline silicon is the most important material in the solar cell today [1] and it will remain foremost and dominant material over the next 10-30 years, owing to its well recognized properties and its established production technology [2].

With the purpose to increase the light collection and the resultant efficiency of silicon solar cells, the reflection of the front surface needs to be minimized [3]-[5]. Surface texturization of monocrystalline structure of silicon (100) by alkaline etchants is called "random pyramid" texture [6], [7]. The pyramidal structures are formed on the surface of silicon because alkaline solutions etch silicon along crystallographic orientation. The etch rate in the <100> direction is much faster than <111> direction as <111> plane shows a very high atomic packing density. So when the slow etching planes, apparently of (111) orientation are exposed, they intersect at the surface to form square based upright pyramids of random size which are distributed randomly on the surface as shown in Fig. 1 [8]. These pyramidal textures have geometries which allow sunlight to be more easily coupled into the silicon. Thus it allows as much light as possible to be absorbed due to multiple reflections and thus converted to electrical current in the solar cell [9].

In this study, texturization based on alkaline anisotropic etching was investigated by using potassium hydroxide (KOH) as alkaline etchant and isopropyl alcohol (IPA) as surfactant. Etching of silicon in potassium hydroxide solution has the advantages of simplicity, easy handling, its low-cost and its homogeneous etching rate of the (100) crystal plane [10]. The study aims at optimizing the process of surface texturization to obtain as low reflectance as possible. The morphological characteristics of silicon surfaces etched with varying concentration of alkali and alcohol as well as the average reflectance to evaluate the surface texturization effects have been reported.

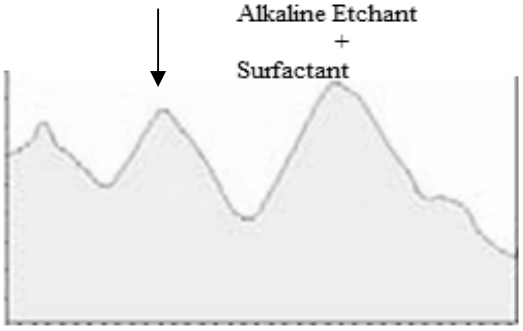


Fig. 1 Formation of pyramids using Alkali/Alcohol solution[8]

Experimental details

Cleaning of Silicon wafers

Experiments were carried out on n-type silicon (100) substrate pieces of size (2.5cm×2.5cm). Wafers were cleaned by standard piranha solution (H_2SO_4:H_2O_2=3:1 by volume) for 10 minutes to remove metal and organic contaminants which cause problem on the surface. After this, the wafers were rinsed thoroughly with de-ionized water. These were then dipped into diluted HF solution (HF: H_2O=1:20) to remove the native oxide [11] and again rinsed in DI water and dried in an oven. The samples were placed in a holder made up of quartz

glass and immersed vertically into etching solution of composition as given in Table I

Table I: the composition of etching solution used for texturization of different samples

Sample No.	KOH (wt. %) at 60°C	IPA (vol. %)	Time (mins)
1.	30	5	10
2.	30	10	10
3.	2	6	10

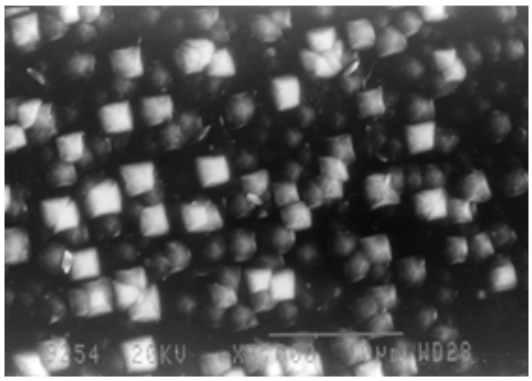


Fig. 2 (a)

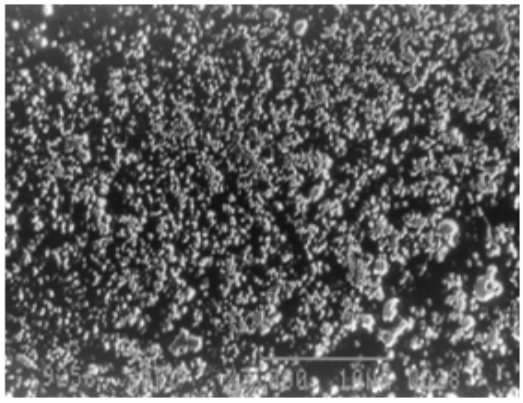


Fig. 2 (b)

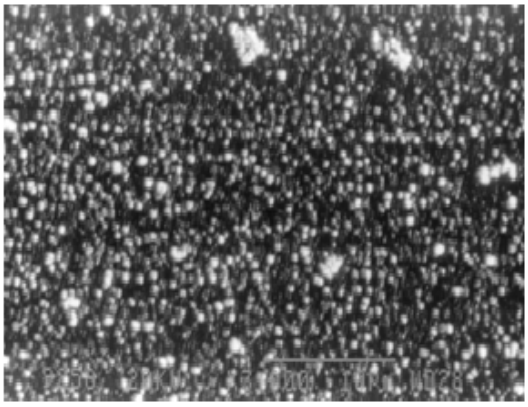


Fig. 2 (c)

Fig. 2 SEM images of textured surfaces using KOH/IPA (a) Sample 1 (b) Sample 2 (c) Sample 3

Characterization technique

Morphology of the texturized samples was observed with SEM (Scanning Electron Microscope) and their reflectivity in the wavelength range from 200nm to 1200 nm was measured with a UV-VIS-NIR spectrophotometer (Perkin Elmer Scan-Lambda 750 double-beam) equipped with an integrating sphere accessory at Punjab University, Chandigarh.

Results and discussions

SEM Results

Fig. 2 presents a perspective view of SEM images of surfaces etched in alkaline solution (KOH/IPA). As a result, random pyramidal structures are formed [12]. The pyramid size for different samples as evaluated from the SEM images is also shown in Table II.

Table II: comparison of pyramid size for different samples varying in concentration of koh & ipa

Sample No.	Pyramid size
1.	2.2 μm
2.	0.58 μm
3.	0.8 μm

Morphology of Fig. 2 suggests uniform etching, dense distribution of pyramids and reasonable pyramid size for sample 1. It shows cracks and non-uniform textures because of different and very small sizes of pyramids in case of sample 2 and exhibits non-uniform etching and small pyramid size for sample 3. Small pyramids, due to their increasing density of pyramid valleys, are not suitable for solar cells. Main problem arising from these valleys are local epitaxial growth and possible capture of wet processes contaminants [13]. Thus shape and coverage of pyramids depend on concentrations of KOH and IPA in the solution. With the concentration of KOH increased, the shape is large and coverage of these pyramids is more uniform. Also low IPA concentration yields less number of pyramids but of bigger size [17]. In general, isopropyl alcohol (IPA) acts as a wetting agent when added to alkaline etchant, resulting in higher uniformity of the pyramid structure. IPA also promotes the formation of pyramids by removing the hydrogen bubbles sticking on the silicon surface. Their masking effect results in a lateral etching action of the solution, which is essential for the formation of the pyramid [14], [15]. Thus concentration of KOH & IPA solution has prominent role in texturing.

Spectrophotometer Results

Fig. 3 uses two-dimensional groove textures to demonstrate how geometrical texturization can reduce the amount of light lost by front surface reflectance for silicon in air, without the use of an antireflection coating. Light, which is reflected away from a groove facet at its first point of incidence may be redirected toward the silicon via a neighboring texture facet, for a second chance of transmission into the silicon thereby lowering reflectance at the front surface. The probability with which light will receive such double-bounce incidence or still higher orders of multiple incidences depends upon the facet

tilt angles of the geometrical textures with respect to the surface of the wafer as in Fig. 3.

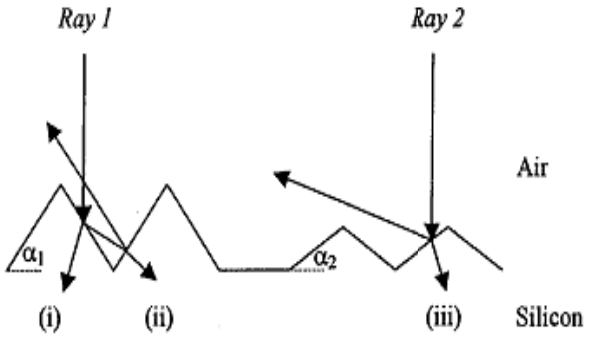


Fig. 3 Possible paths for light incident upon geometrically textured silicon in air [16]

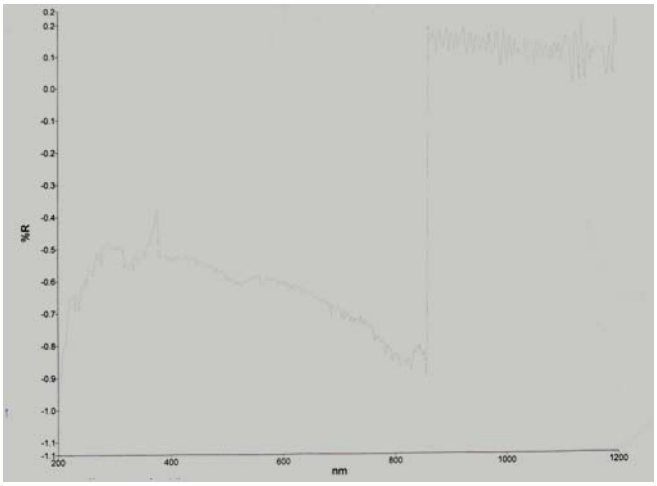


Fig. 4 (a)

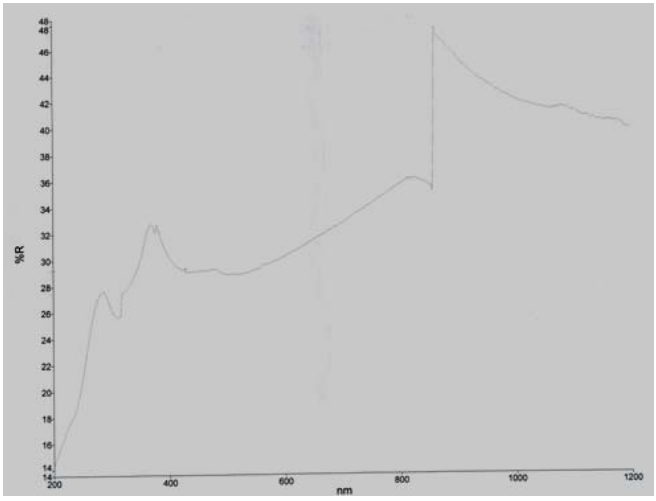


Fig. 4 (b)

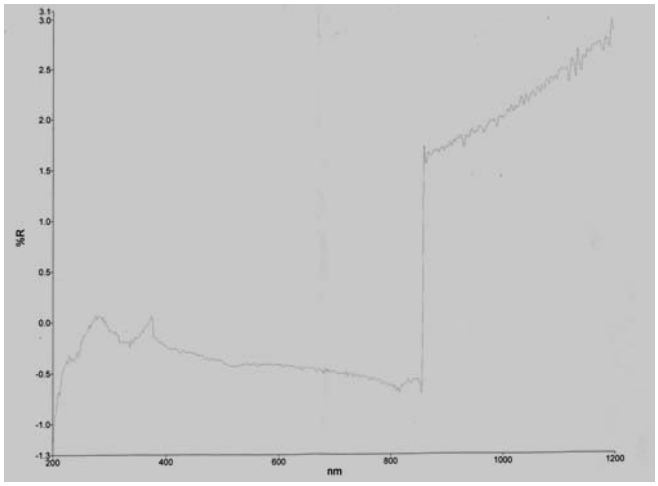


Fig. 4 (c)

Fig. 4 Spectral Reflectance of the texturized silicon (100) surface (a) Sample 1 (b) Sample 2 (c) Sample 3

For normal incident light falling upon periodic textures, ray 1 experiences double bounce incidence for $\alpha_1 > 30°$, whereas for $\alpha_2 < 30°$, ray 2 is reflected directly away without a second chance of incidence. The multiple incidences of light are more possible if angle of pyramid to the surface lies in the range $30° < \alpha < 45°$ (in air). The range $45° < \alpha < 54.7°$ guarantees double bounce incidences of light & the range $\alpha > 60°$ ensures yielding of at least triple bounce reflectances. As a result, texturing of the silicon (100) leads to an absolute reflection reduction of approximately 20% compared to a flat polished wafer in air [16].

Table III: Comparison of reflectance values for different samples.

Sample No.	Average Reflectance (%) (400-800 nm)	Minimum Reflectance (%) (400-800 nm)	Difference of min. & max. Reflectance (%) (400-800 nm)
1	-0.638	-0.83 at 800 nm	0.3
2	31.2	28.8 at 500 nm	6.9
3	-0.414	-0.6 at 800 nm	0.35

Spectral reflectance measurements obtained from textured silicon wafers (using different solutions) are shown in Fig. 4. The reflectance spectra have been studied in the wavelength range from 200 nm to 1200 nm. Their analysis (Table III) shows that for sample 1, the average reflectance and the minimum reflectance are -0.638% and -0.83% respectively which are least amongst all samples for visible range. Average reflectance value achieved for sample 1from Fig. 4(a) is lower than that as obtained in [17]. Also difference between minimum and maximum value of reflectance is about 0.3% showing not much variation in reflectance for visible range in

sample 1. Compared to it, the sample 2 has maximum reflectance amongst all samples in visible range. In infrared region, sample 1 has comparatively constant reflectance (0.2%), sample 3 has increasing reflectance and sample 2 also has high reflectance. This shows that the etchant concentration used for sample 1 is most appropriate for fabrication of silicon solar cells.

Conclusion

Differently composed alkaline texturing solutions have been investigated by varying the concentration of KOH/IPA at 60°C for 10 minutes. By studying the SEM results and the reflectivity of the textured surface, process variables have been identified that give optimized value for both surface reflectance as well as pyramid size. These values can be further utilized for the fabrication of the silicon solar cell.

Acknowledgement

The authors would like to acknowledge the contribution of Dr Jaidev, Chairman, Sh. Nalini Kant, mentor and Dr. J.K.Sharma, Director, ACE by way of their invaluable guidance, kind support, spirited motivation and encouragement, without which this task would not have been accomplished.

References

[1] L.A. DOBRZAŃSKI, A. DRYGAŁA, "Surface texturing of multicrystalline silicon solar cells", *Journal of achievements in materials and manufacturing engineering*, Volume 31 Issue 1, November 2008

[2] M. Lipinski, P. Zieba, A. Kaminski, "Crystalline silicon solar cells in foundation of materials design", *Research Signpost*, 2006, 285-308

[3] D. L. King, M. E. Buck, "Experimental Optimization of an anisotropic etching process for random texturization of silicon solar cells", *in proc. Of 22nd IEEE PVSC,* p.308 (1991)

[4] E. Vazsoni et al., "Improved anisotropic etching process for industrial texturing of silicon solar cells", *Solar energy materials and solar cells,* 57, 179 (1991)

[5] Gim Chen and Ismail Kashkous, "Effect of pre cleaning on texturization of c-Si wafers in a KOH/IPA mixture", 2009, 1-8

[6] Kapila Wijekoon, Timothy Weidman, Steve Paak, Kenneth MacWilliams, "Production ready novel texture etching process for fabrication of single crystalline silicon solar cells", *Applied Materials IEEE*, 003635-003641 (2010)

[7] Ou Weiying et al., "Texturization of mono-crystalline silicon solar cells in TMAH without the addition of surfactant", *Journal of Semiconductors*, Vol. 31, No. 10, October 2010

[8] Solar cell surface characterization.pdf

[9] J. D. Hylton, A. R. Burgers, and W. C. Sinke, "Light trapping in alkaline textured etched crystalline silicon wafers", *ECN solar energy*, 1998

[10] P. Papet, O. Nichiporuk, A. Fave, A. Kaminski, B. Bazer-Bachi, M. Lemiti, "TMAH texturisation and etching of interdigitated back-contact solar cells", *Materials Science-Poland*, Vol. 24, No. 4, 2006

[11] Werner Kern,"The Evolution of Silicon Wafer Cleaning Technology", *J. Electrochem. Soc.,* Vol. 137, No. 6, June 1990

[12] Kazuya Tsujino, Michio Matsumura and Yoichiro Nishimoto, "Texturization Of Multicrystalline Silicon Wafers By Chemical Treatment Using Metallic Catalyst", *3rd World Conference on Photovoltaic Energy Conversion* May 11-18, 2003 Osaka, Japan

[13] L. Fesquet , S. Olibet, J. Damon-Lacoste et.al, "Modification of textured silicon wafer surface morphology for fabrication of heterojunction solar cell with open circuit voltage over 700 mV", *IEEE* 2009

[14] Smail Kashkoush, Akrion Systems, Allentown & David Jimenez, Wright Williams & Kelly," Examining cost of ownership of crystalline-silicon solar-cell wet processing: texturization and cleaning", *Photovoltaics International journal*, 81-90

[15] A.K.Chu, J.S.Wang, Z.Y.Tsai, C.K.Lee, "A simple and cost-effective approach for fabricating pyramids on crystalline silicon wafers", *Solar Energy Materials & Solar Cells*, 93 (2009) 1276–1280

[16] J. D. Hylton, A. R. Burgers, and W. C. Sinke, "Alkaline Etching for Reflectance Reduction in Multicrystalline Silicon Solar Cells", *Journal of the electrochemical society*, 151 (6) G408-G427 (2004)

[17] Ma Xun, Liu Zuming, Liao Hua,Li Jintian, "Surface Texturisation of Monocrystalline Silicon Solar Cells", *IEEE* 2011

Region Based Segmentation for Developing Membering Filters

Gurpreet Kaur and Sumit Kaushik

[1]*M.Tech. Student, Department of Computer Science & Engineering,*
Ambala College of Engineering & Applied Research, Kurukshetra University, India
E-mail: kaurgurpreet78@yahoo.com
[2]*Assistant Professor, Department of Computer Science & Engineering,*
Guru Nanak Institute of Technology, Mullana, Kurukshetra University, India

Abstract
This paper deals with region based watershed segmentation for developing "Membering Filters". Watershed transformation is a common technique for image segmentation. Region based segmentation classify a particular image into a number of regions or classes. Region-based segmentation methods attempt to partition or group regions according to Intensity values from original images, Textures or patterns, Spectral profiles. Membering filters are developed by using region based segmentation in which each pores having same size and by this we can check the quality of product or solution.

Keywords: Watershed segmentation, Image segmentation, Region-based segmentation, Membering filters

Introduction

***Image Processing*:-** Image processing is a physical process used to convert an image signal into a physical image. The image signal can be either digital or analog. The actual output itself can be an actual physical image or the characteristics of an image.

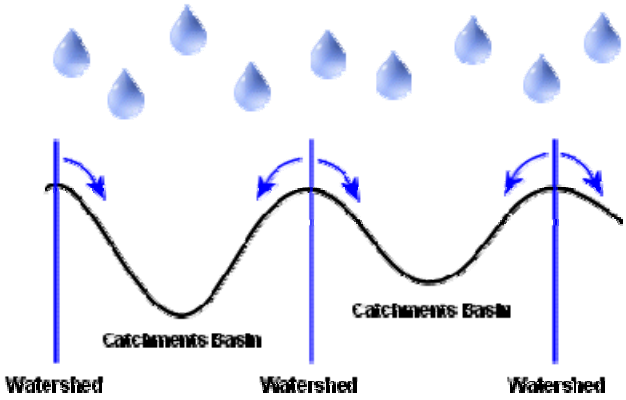
***Image Segmentation*:-** Image segmentation is typically used to locate objects and boundaries in images and should stop when the object of interest in an application have been isolated. Image segmentation is based on three principal concepts

1. Detection of discontinuities
2. Thresholding
3. Region Processing

The goal of image segmentation is to cluster pixels into salient image regions, i.e., regions corresponding to individual surfaces, objects, or natural parts of objects. A segmentation could be used for object recognition, occlusion boundary estimation within motion or stereo systems, image compression, image editing, or image database look-up.

***Watershed Segmentation*:-** Watershed segmentation was simulated based on an immersion process[1]enabling an increase in speed and accuracy. Parallel watershed segmentation was later developed [2], offering clear partitions within images. Watershed transformation has increasingly been recognized as a powerful segmentation process due to its many advantages [3], including simplicity, speed and complete division of the image. Basic watershed segmentation is provided by the Watershed– Packed Features Detect Method. This method can also be explained by a metaphor based on the behavior of water in a landscape.

When it rains, drops of water falling in different regions will follow the landscape downhill. The water will end up at the bottom of valleys. In this case we do not consider the depth of the resulting lakes in the valleys. Each valley is associated with a catchments basin, and each point in the landscape belongs to exactly one unique basin.



Different approches may be employed to use the watershed principle for image segmentation.

1. Local minima of the gradient of the image may be chosen as markers, in this case an over-segmentation is produced.
2. Marker based watershed transformation make use of specific marker positions which have been either explicitly defined by the user or determined automatically. [4]

Advantages of the watershed transform include that it is a fast, simple and intuitive method. More importantly, it is able to produce a complete division of the image in separated regions even if the contrast is poor, thus there is no need to carry out any post processing work, such as contour joining. Its drawbacks will include over-segmentation and sensitivity to noise.

Membrane Filters:-Membrane filters or "membranes" are microporous plastic films with specific pore size ratings. Also known as screen, sieve or microporous filters, membranes retain particles or microorganisms larger than their pore size primarily by surface capture. Some particles smaller than the stated pore size may be retained by other mechanisms. Advantec membranes are produced by three different processes. Mixed Cellulose Esters, Cellulose Acetate, and Nylon are reverse phase solvent cast membranes, where controlled evaporation or removal of the complex solvent system forms the porous structure.

Quick Way to Selecting Membrane Filters

1. Determine what liquid or gas will be filtered.
2. Check which membranes are chemically compatible.
3. Determine the maximum pore size required to achieve the results we want.
4. Check the membrane specifications for any unusual process conditions that might otherwise limit your choice of membrane (e.g. temperature).

Litreture Survey

Litreture survey reveals various applications of watershed segmentation that have been developed yet.

Medical confocal image analysis(Cancer Detection)

Cancer is a serious global healthcare problem in which image was imported into MATLAB and converted to grayscale for faster processing using inbuilt functions. It was used to eliminate incomplete nuclei and only account for complete and visible ones.The gradient magnitudes were calculated and the threshold was increased to outline cell membranes. Although widely considered a disease of the developed world, 60% of cancers occur in developing countries, where low per-capita healthcare expenditure, unreliable infrastructure and facilities render advanced cancer screening technologies inaccessible.

We have developed a low-cost, handheld, microelectromechanical systems (MEMS)-based in vivo confocal microscope for sub-cellular-resolution imaging of tissue towards early detection of epithelial pre-cancers from which 85% of cancers originate. Endoscopic procedures are performed for biopsy samples and images are manually segmented for initial testing of pre-cancer.[6]However, this results in long turnaround time, high costs, discrepancies among different segmentation methods, and inconvenience for patients. An advanced algorithm is needed to provide fast, low-cost, standardized results which are essential for in vivo pre-cancer detection.

3D face recognition

To understand watershed-based segmentation we consider shape of face as a topographic surface. We flood this surface from its minima and, if we prevent the merging of the waters coming from different sources, we partition the image into two different sets: the catchment basins and the watershed lines.[7]The main problem of this approach is the over segmentation due to the small variations which exist in the surface of the face. A method of 3D face recognition based on facial curve matching, where, each of 3D facial range data was aligned using various heuristics via five feature points, then, three curves in aligned and unified coordinate system were extracted and thus formed a feature vector.[8] Two faces could be considered to be matched when the distance between the two feature vectors is small.

Weed detection in cereals field

In this the development of near-ground image capture and processing techniques in order to detect broad leaf weeds in cereal crops, under actual field conditions. The proposed methods use both colour and shape analysis techniques for discriminating crop, weeds and soil. The performance of algorithms was assessed by comparing the results with a human classification, providing a good success rate & the potential of using image processing techniques to generate weed maps.[9]

Weed detection using image processing techniques have shown a good potential to estimate weed distribution even the difficulties due to the similarity in spectral reflectance between weed and crop plants, and because of the variability of natural scenes.

It seems convenient that this approach be complemented by other sources of information in order to generate weed maps that are sufficiently comprehensive to use in a patch spraying system.

In order to reduce the error rate, different approaches are being undertaken. Concerning acquisition, B/W video cameras equipped with NIR filters, seems to provide enhanced images which facilitates initial segmentation.

On the other hand, illuminant modelling algorithms try to overcome the drawbacks own to the presence of highlights and shadows. Finally, more resolution images should be obtained to reduce the error rate of shape analysis steps.

Crack Detection on pavement surface images

Crack detection on road surface images using the wavelet transform combined with grey level morphology and segmentation.[10] Pavement crack detection is not a "simple" edge detection problem due to the various pavement textures that can be encountered on "road image". A way to reduce the texture effect is to use low spatial resolution images. Pavement crack detection is not a "simple" edge detection problem due to the various pavement textures that can be encountered on "road image". A way to reduce the texture effect is to use low spatial resolution images. But low resolution tends to erase thin crack signatures. So, they won't be detected by image segmentation. Consequently, we have chosen to work with images whose spatial resolution is between 1 and 2 mm per pixel. If we look forward to the final on road operational system, such spatial resolution seems to be realistic, due to available technologies on the market.

Echocardiography

Echocardiography is a valuable diagnostic imaging modality for patients with heart diseases. With increasing computational power, automatic detection of the left verticle (LV), and particularly the endocardial boundary from echocardiographic images, is a very useful step for clinical diagnosis and is done

by ROI segmentation and LV boundary determination. [11]

The steps involved are described as follows:

Preprocessing

To reduce the influence of speckle noise, the composite image is smoothed by the adaptive neighborhood smoothing method. Since the interior region of the ventricle has intensities consistently below a threshold gray level τ, a large smoothing kernel is used for pixels below τ, and a small smoothing kernel is used for pixels above τ. For an image of m × n pixels, the sizes of the large kernel and the small kernel are Kl × √m × n and Ks × √m × n, respectively. It is repeated twice to ensure that the low-intensity regions are smoothed more thoroughly than the high-intensity regions.

Binary Image Processing

The preprocessed image is then converted to a binary image by the threshold τ. The binary image is subjected to a morphological close operation of radius C × √m × n to separate the overlapping chambers. At the completion of this step, four regions corresponding to the cardiac chambers are visually separated.[12]

Watershed Segmentation

The Euclidean distance transform of the image after close is computed. To accomplish binary image segmentation, the watershed immersion algorithm [13] is applied to the Euclidean distance map to produce the label image. The binary image after close is then masked by multiplying with the label image. Every object in the masked binary image is labeled with different integer value for distinct identification. In this particular case, LV is the object on the top-right of the image. In the watershed label map, the catchment basin to which this object belongs is considered as the region corresponding to the LV. This region defines the ROI, and its center is the LV center point

Problem Formulation

We will be using region –based segmentation for developing membering filters to check quality of solution or product.

Let we have X number of applications

{X1, X2, …….Xn}

And Y number of segmentation techniques

{Y1, Y2, …….Ym}

We will apply different techniques at different applications.

Suppose we develop one application “membering filters”for that size of pore is an important parameter.

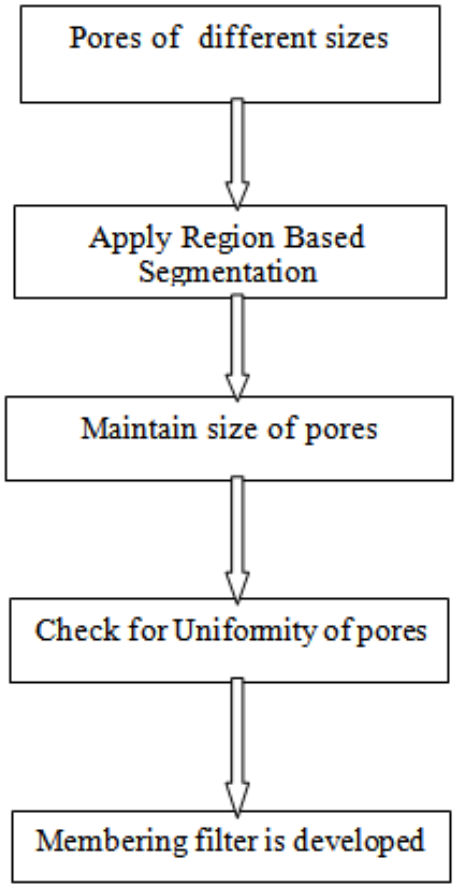
{S1, S2……Sp}

Proposed Model

We will try to see how watershed technique is useful in different domains.

We will take pores of different sizes and by applying region based segmentation technique we maintain the size of pores and make the uniformity in the pores.Pores must be of equal size so that quality of solution is maintained.

We will implement this by using OPEN CV tool.



Conclusion & Future Scope

We are concerning with size of pores. If the pores are not uniform then the product is not purified so we produce membering filter to improve quality of a solution. In future work, performance of system can be further enhanced by using various techniques. Size of the pores is an important parameter for developing membering filters by using region based segmentation.

References

[1] L. Vincent and P. Soille, "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations", IEEE Transactions on PAMI, Vol. 13, No. 6, June 2006, pp. 583-598.

[2] A.N. Moga, M. Gabbouj, Parallel image component labelling with watershed transformation, IEEE transactions on pattern analysis and machine intelligence. 19 (1997) 441–450.

[3] S. Beucher, F. Meyer, The Morphological approach to segmentation: the watershed transform, in: E.R.Dougherty (Ed.), Mathematical Morphology in Image Processing, Marcel Dekker, New York, 1993,

pp. 433– 481.

[4] Rafael C.Gonzalez, Richard E.Woods"Digital Image Processing"3rd edition.

[5] Woebbecke D. M., G.E. Meyer, K. Von Bargen and D. Mortensen (1995). Shape features for identifying young weeds using image analysis. Transactions of the A.S.A.E. 38 (1): 271-281.

[6] K. Kumar, K. Hoshino, H.J. Shin, R. Richards-Kortum and X.J. Zhang, "High-reflectivity Two-Axis Vertical Combdrive Microscanners for Sub-cellular Scale Confocal Imaging Applications", Proceedings of International Conference on Optical MEMS and their Applications (Optical MEMS '06), August 21-24, Montana, USA, 2006.

[7] B. Ben Amor, M. Ardabilian, L. Chen, "3D Face Modeling Based on Structured-light Assisted Stereo Sensor". Proceeding of ICIAP 2005, Cagliari, Italia, 6-8 September 2005.

[8] A.J. Pérez, F. López, J.V. Benlloch, S. Christensen "Colour and Shape Analysis Techniques For Weed Detection In Cereal Fields" University of Politecnica of Valencia Departament of Ingenieria of Sistemas, Computadores y Automática

[9] B. Schmidt, Automated Pavement Cracking Assessment Equipment : State of the Art, Routes-Roads, World Road Association (PIARC), N°320, October 2003, pp. 35-44.

[10] J. Park and J. M. Keller, "Snakes on the watershed, " IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 10, pp. 1201–1205, Oct. 2001.

[11] A. Krivanek and M. Sonka, "Ovarian ultrasound image analysis: Follicle segmentation, " IEEE Trans. Med. Imag., vol. 17, no. 6, pp. 935–944, Dec. 1998.

[12] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulation s, " IEEE Trans. Pattern Anal. Mach. Intell., vol. 13, no. 6, pp. 583–598, Jun. 1991.

[13] http://www.google.co.in/#hl=en&source=hp&biw=1280&bih=607&q=what+is+image+segmentation&aq=o&aqi=&aql=&oq=&fp=c27588bd413fbfe3

Analysis of Different Clustering Algorithms on Image Databases

Stuti Mehla and Ashok Kajal

[1]M.Tech Research Scholar, [2]Assistant Professor
Ambala College of Engineering & Applied Research, Devsthali, India
E-mail:stuti21mehla@gmail.com

Abstract
When we apply Image Retrieval techniques to large image Databases .It provides restriction of search space to provide adequate response time. This restriction can be done minimized by using Clustering technique to partition the image dataset into subspaces of similar elements .In this article we will apply different clustering algorithms on large image database and then evaluate and analyse the performance of these algorithms to determine which algorithm is best for image retrieval.

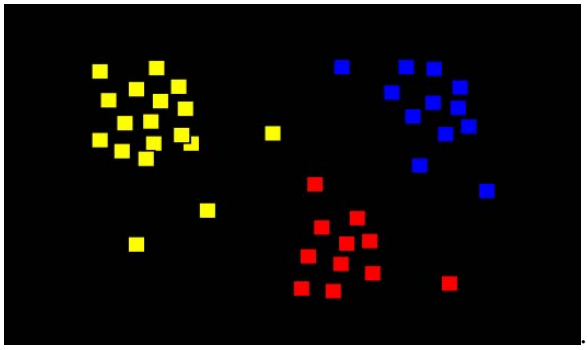
Keywords: Clustering, Image Retrieval, Database.

Introduction

What is Clustering?

Clustering is the process in which set of observations is divided into subsets called clusters, so that observations in same cluster are similar in some sense. Clustering is method of unsupervised learning, used in many fields, including machine learning, data mining, pattern recognition, image analysis, and binformatics.

Cluster analysis itself is not an algorithm but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the data space and multivariate normal distributions. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery that knowledge that fires.



Result of cluster analysis shown in three different squares.

The Goals of Clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.

For instance, we could be interested in finding representatives for homogeneous groups (*data reduction*), in finding "natural clusters" and describe their unknown properties (*"natural" data types*), in finding useful and suitable groupings (*"useful" data classes*) or in finding unusual data objects (*outlier detection*).

Image Clustering

The goal of image clustering is to find out a mapping of query image into classes called clusters such that these set of clusters provide same information about query image as the entire set collection. Because searching large databases of images is a challenging task. We calculate the similarity between the query image and all the images in the database and rank the images by sorting their similarities. One problem with this exhaustive search approach is that it does not scale up for large databases. The retrieval time for exhaustive search is the sum of two terms: Tsim and Tsort. Tsim is the time to calculate the similarity between the query and every image in the database, and Tsort is the time to rank all the images in the database according to their similarity to the query.

Texhaustive= n T1sim+ O (nlogn)

Where n is the number of images in the database, T1sim is the time to calculate the similarity between two images, and O (nlogn) is the time to sort n elements. When the images in the database are clustered, the retrieval time is the sum of three terms, the time to calculate the similarity between the query and the cluster centers, the time to calculate the similarity between the query and the images in the nearest clusters and the time to rank these images. Therefore the total search time is:

Tcluster = k T1sim+ l T1sim+O(l log l)

Here k is the number of clusters, l is the number of images in the clusters nearest to the query. Since k<<n and

l<<n, Tcluster <<Texhaustive.

Image clustering is important for efficient search and retrieval in large image databases. [1]

Possible applications

Clustering algorithms can be applied in many fields, for instance:

***Marketing*:** finding groups of customers with similar behaviour given a large database of customer data containing their properties and past buying records;

***Biology*:** classification of plants and animals given their features.

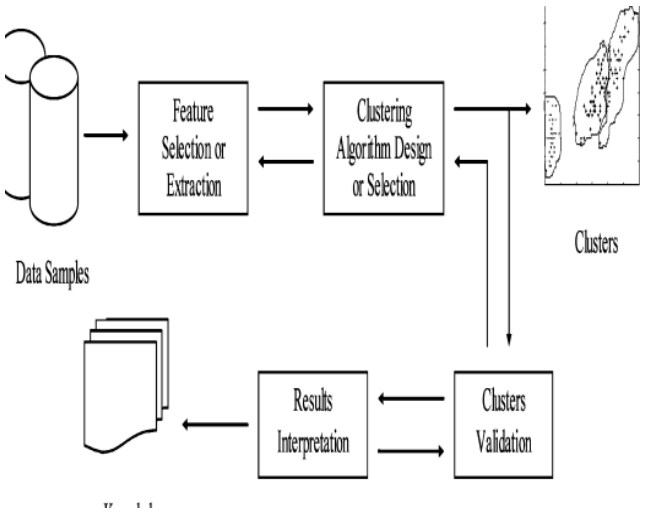
***Libraries*:** book ordering;

***Insurance*:** identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;

***City-planning*:** identifying groups of houses according to their house type, value and geographical location;

***Earthquake studies*:** clustering observed earthquake epicentres to identify dangerous zones;

***WWW*:** document classification; clustering weblog data to discover groups of similar access patterns.



Steps in clustering:

***Feature Extraction*:**

Feature selection chooses distinguishing features from a set of candidates, while feature extraction utilizes some transformations to generate useful and novel features from original ones. Both are very crucial to effectiveness of clustering applications. [9]

***Clustering algorithm design or selection*:** The step is usually combined with the selection of a corresponding proximity measure and the construction of a Criterion function .Patterns are grouped according to whether they resemble each other. Obviously, the proximity measure directly affects the formation of resulting clusters. Almost all clustering algorithms are explicitly or implicitly connected to some definition of proximity measure [9].

***Cluster validation*:** Given a data set, each clustering algorithm can always generate a division, no matter whether the structure exists or not. Moreover, different approaches usually lead to different clusters, and even for the same algorithm, parameter identification or presentation order of input patterns may affect the final result. Therefore effective evaluation standards and criteria are important to provide users with a degree of confidence for the clustering results derived from the used algorithms [9].

***Results interpretation*.** The ultimate goal of clustering is to provide users with meaningful insights from the original data, so that they can effectively solve the problems encountered [9]

Literature Survey

Different approaches have been used in literature for clustering of database.

***Hierarichal Clustering algorithm*:** In this paper I have studied the hierarchical clustering algorithm which is used for image retrieval .In Hieratical clustering we create a hierarchy of clusters which may be represented in a tree structure called a dendrogram. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations. Algorithms for hierarchical clustering are generally either agglomerative starts at the leaves and successively merges clusters together; or divisive in which one starts at the root and recursively splits the clusters [1].

K-Means Clustering: In this paper I have Studied k-means algorithm and a modified version have also discussed. In this paper distance matrices are improved. In k-means procedure follows a simple and easy way to classify a given data set through a certain number of clusters is to define k centroids, one for each cluster. These centroids should be placed in a way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bar centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. [3]

In this paper I have studied comparative analysis of Hierarchial, K-means, PAM, CLARA algorithm. Basics about PAM and CLARA are described below PAM: The pam is a clustering algorithm related to the means algorithm and the medoidshift algorithm. Both the k-means and PAM algorithms are partitioned (breaking the dataset up into groups).In contrast to the k-means algorithm, k-medics chooses data points as centres. PAM is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters known a priori. A useful tool for determining k is the

silhouette. It is more robust to noise and outliers as compared to k-means. A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal. CLARA :(Clustering large applications) improves time complexity of PAM by using samples of dataset. The basic idea is that it applies PAM to a sample of underlying database and then uses medics found as the medoids for the complete clustering. Each item from complete database is then assigned to the cluster with medoid to which it is closest.Because, of sampling CLARA is more efficient then PAM. [6]

In this paper improved fuzzy c-means algo is discussed. Fuzzy c-means algorithm is developed because in Hard clustering every data element is related to only one cluster. If we want that a data element should related to more than one cluster than we use fuzzy logic. [7]

In this paper Image clustering and compression technique is applied. Color base Image clustering is done. [8]

Problem Formulation

We will analyze the performance of clustering algorithms described above.

{I1, I2, I3…….In} is an array of images in image database.

{A1, A2, A3…….Am} is an array of clustering algorithms.

After applying clustering algorithms a€A on

{i1…..ik}€I we will analyse the performance of clustering algorithms by applying them on image databases.

Proposed Model

In the proposed model we will apply clustering algorithms on image database which are discussed above and then graph is plotted with the help of MATLAB and after this performance is evaluated. The steps are pictorially represented

Conclusion & Future Scope

We are concerning with performance analysis of clustering algorithms on image database.

In future work we can analyze the algorithms on relational database, image and pattern recognition.

We can also improve efficiency of any clustering algorithm.

References

[1] "Hierarchical clustering algorithm for fast image retrieval" Santhana Krishnamachari Mohamed Abdel-Mottaleb Philips Research 345 Scarborough Road Briarcliff Manor, NY 10510 {Sgk, msa}@philabs.research.philips.com2)

[2] R. Yager, "Intelligent control of the hierarchical agglomerative clustering Process," IEEE Trans. Syst., Man, Cybern., vol. 30, no. 6, pp 835–845, 2000

[3] M. Su and C. Chou, "A modified version of the K-means algorithm with A distance based on cluster symmetry," IEEE Trans. Pattern Anal. Mach. Intel., vol. 23, no. 6, pp. 674–680, Jun. 2001

[4] C. Ordonez and E. Omiecinski, "Efficient disk-based K-means clustering For relational databases," IEEE Trans. Know. Data Eng., vol. 16, no. 8, pp. 909–921, Aug. 2004

[5] "Data clustering: 50 years beyond K-means "Anil K. Jain * Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan 48824, USA Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seoul, 136-713, Korea 0167-8655/$ - see front matter _ 2009

[6] "Comparing Clustering Methods for Database Categorization in Image Retrieval" Thomas Kaster, Volker Wendt, and Gerhard Sagerer, Springer-Verlag Berlin Heidelberg, 2003

[7] J. Zhang and Y. Leung, "Improved possibilistic C-means clustering algorithms," IEEE Trans. Fuzzy Syst., vol. 12, no. 2, pp. 209–217, Apr. 2004.

[8] An Algorithm for Image Clustering and Compression, Mertin KAYA, Turk Demirdokum Fabrika A.S Bozuyuk Bielecik-TURKEY, VOL13, NO-1, 2005

[9] Survey of Clustering Algorithms Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005

A Robust Algorithm for Iris Segmentation and Normalization using Hough Transform

[1]Sunil Chawla and [2]Aashish Oberoi

[1]Department of Information Technology, MMEC, M.M. University, Mullana, India
[2]Department of Computer Science & Engineering, MMEC, M.M. University, Mullana, India
E-mail: true.umang@gmail.com , a_oberoi01@yahoo.co.in

Abstract
There have been several implementations of security systems using biometric, especially for identification and verification cases. An example of pattern used in biometric is the iris pattern in human eye. The iris pattern has been proved unique for each person. The use of iris pattern poses problems in encoding the human iris. The iris recognition system consists of an automatic segmentation system that is based on the Hough transform, and is able to localize the circular iris and pupil region, occluding eyelids and eyelashes, and reflections. The extracted iris region was then normalized into a rectangular block with constant dimensions to account for imaging inconsistencies. Finally, the phase data from 1D Log-Gabor filters were extracted and quantized to four levels to encode the unique pattern of the iris into a bit-wise biometric template using Daugman's rubber-sheet model. The Hamming distance was employed for classification of iris templates, and two templates were found to match if a test of statistical independence was failed.

Keywords: Segmentation, Normalization, Histogram equalization, Canny, Hough Transform, 1D Log Gabor filter.

Introduction

The increasing advancements in the field of Information Technology & World Wide Web, causing human beings frequently confront with various kinds of unauthorized access. Also with the enlargement of mankind's activity range, the importance for person's status identity is becoming more and more important. So many different techniques for person's status identity have been proposed for this practical task. Conventional methods for status identity check like password and identification card are not always reliable, because these methods can be easily forgotten, stolen or forged. A wide variety of biometrics have been developed for this challenge, examples include automatic retinal vasculature scan, iris recognition, fingerprints matching, hand shape identification, handwritten signature verification, and voice recognition systems. Since 1987, when L. Flom and A. Safir [12] concluded about the stability of iris morphology and estimated the probability for the existence of two similar irises at 1 in 10^{72}, the use of iris based biometric systems has been increasing. Iris is commonly recognized as one of the most reliable, unique and noninvasive biometric measures: it has a random morphogenesis and, apparently, no genetic penetrance. A biometric system provides automatic identification of a human being based on some unique physical or behavioral feature of the individual. Iris recognition is regarded as the most reliable and accurate biometric identification system being used in modern era. Most commercial iris recognition systems use patented algorithms developed by Daugman [1, 2], and these algorithms are able to produce perfect recognition rates. However, published results have usually been produced under favourable conditions, and there have been no independent trials of the technology.

Overview

The system, as shown in Figure 1, is implemented in MATLAB.

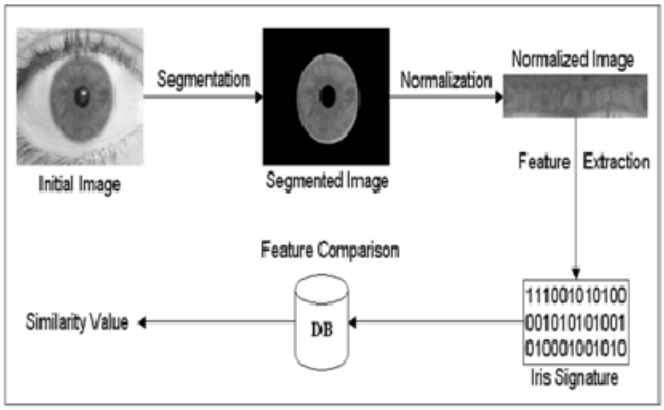


Fig.1. Typical stages of iris recognition.

A general iris recognition system is composed of four steps. Firstly an image containing the eye is captured then the original image containing iris is preprocessed to extract the iris. Thirdly iris features are extracted from the segmented image and is encoded in the form of an iris template and finally decision regarding acceptance or rejection of the subject is made by means of matching.

This paper is divided into five sections. The Section 1 introduces what is the position of iris technology in personal authentication. In the Section 2, we sum up the state of the art approaches in iris recognition. The most widely documented

in open literature and well known iris recognition system developed by J. Daugman [2] is taken as reference for comparison. The Section 3 presents the proposed approach in details, and discusses the different issues we chose. The Section 4 provides test results and illustration of typical iris signature. At last a conclusion is made in Section 5, which talks about the future considerations for the improvement of the proposed solution as well.

Earlier Works

Daugman [1, 2] proposed an integro-differential operator for localizing iris regions along with removing the possible eyelid noises. From the publications, we cannot judge whether pupil and eyelash noises are considered in his method. Wildes [4] processed iris segmentation through simple filtering and histogram operations. Eyelid edges were detected when edge detectors were processed with horizontal and then modeled as parabolas. No direction preferences lead to the pupil boundary. Eyelash and pupil noises were not considered in his method. Boles and Boashah [5], Lim et al. [6] and Noh et al. [7] mainly focused on the iris image representation and feature matching, and did not introduce the information about segmentation. Tisse et al. [8] proposed a segmentation method based on integro-differential operators with a Hough Transform. This reduced the computation time and excluded potential centers outside of the eye image. Eyelashes and pupil noises were also not considered in his method. Ma et. al. [9] processed iris segmentation by simple filtering, edge detection and Hough Transform. In Masek's segmentation algorithm [11], the two circular boundaries of the iris are localized in the same way. The Canny edge detector is used to generate the edge map. Then after doing a circular Hough transform, the maximum value in the Hough space corresponds to the center and the radius of the circle.

Proposed Approach

For every iris recognition system, accuracy of the system is highly dependent on accurate iris segmentation. Better the iris is localized, better will be the performance of the system. Our basic experimentation of the *Daugman's* mathematical algorithms for iris processing, is derived from the information found in the open literature, led us to suggest a few possible improvements. For justification of these concepts, we implemented in *MATLAB*. Afterwards we tested individually the performances of the different processing blocks previously identified as follows: (1) locating iris in the image, (2) Cartesian to polar reference transform, (3) local features extraction, and encoding (4) matching.

Segmentation

Segmentation is a process of finding the most useful portion of the iris image for further processing. It is done by localizing pupil and iris boundaries, eyelashes and eyelids. In case there is no proper segmentation, next stages of iris recognition will suffer and false data will be generated as template which in turn affects the recognition rates. To speed iris segmentation, the iris has been roughly localized by a simple combination of Gaussian filtering, canny edge detection, and Hough transform. Hough Transform is used to deduce the radius and center of the pupil and iris circles. Canny edge detection operator [15] is used to detect the edges in the iris image which is the best edge detection operator available in MATLAB.

Normalization

The localized iris is then normalized to a rectangular block with a fixed size radius being in correspondence to the width of the block and angular displacement θ being in correspondence with the length of the block as shown in fig. 2.

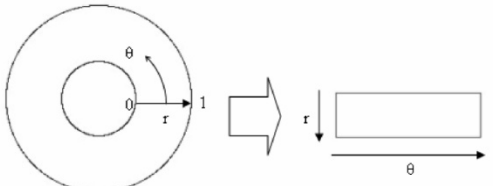


Fig. 2: Daugman's rubber sheet model.

Formally, the rubber sheet is a linear model that assigns to each pixel of the iris, regardless its size and pupillary dilation, a pair of real coordinates *(r, θ)*, where r is on the unit interval [0, 1] and θ is an angle in range [0, 2π]. The remapping of the iris image *I(x, y)* from raw Cartesian coordinates *(x, y)* to the dimensionless non concentric polar coordinate system *(r, θ)* can be represented as:

$$I(x(r, \theta), y(r, \theta)) \rightarrow I(r, \theta) \quad (1)$$

where x(r, θ) and y(r, θ) are defined as linear combinations of both the set of pupillary boundary points ($x_p(\theta)$, $y_p(\theta)$) and the set of limbus boundary points along the outer perimeter of the iris ($x_s(\theta)$, $y_s(\theta)$) bordering the sclera:

$$\begin{aligned} x(r,\theta) &= (1-r) * x_p(\theta) + r * x_s(\theta) \\ y(r,\theta) &= (1-r) * y_p(\theta) + r * y_s(\theta) \end{aligned} \quad (2)$$

where *I(x, y)* is the iris region image, *(x, y)* are the original cartesian coordinates, *(r, θ)* are the corresponding normalized polar coordinates, and (x_p, y_p) and (x_i,y_i) are the coordinates of the pupil and iris boundaries along the θ direction.

Feature Extraction and Encoding

Feature Extraction is a process to extract the information from the iris image. These features can not be used for reconstruction of images. But these values are used in classification. Gabor filters are used for the purpose. Gabor filters give rotation – invariant system for feature extraction. In our experiments, we employed a Gabor filter with isotropic 2D Gaussian for rotation invariant classification. Gabor filter's frequency domain equation is as follows:

$$\begin{aligned} G(x,y) &= g(x,y)\exp(-2\pi j(u.x + v.y)) \\ g(x,y) &= -\exp(x^2 + y^2 / 2\sigma^2).j = \sqrt{-1} \end{aligned} \quad (3)$$

The complex function G(x, y) can be spilt into two parts, even and odd filters. $G_e(x, y)$ and $G_o(x, y)$, which are also known as symmetric and anti-symmetric filters respectively. The spatial Gabor filter is given in eq. 4.

$$G_e(x,y) = g(x,y)\cos(2\pi f(x\cos\theta + y\sin\theta))$$
$$G_o(x,y) = g(x,y)\sin(2\pi f(x\cos\theta + y\sin\theta)) \quad (4)$$

where G(x,y) is Gabor filter's kernel and g(x,y) is an isotropic 2D Gaussian function.

Matching

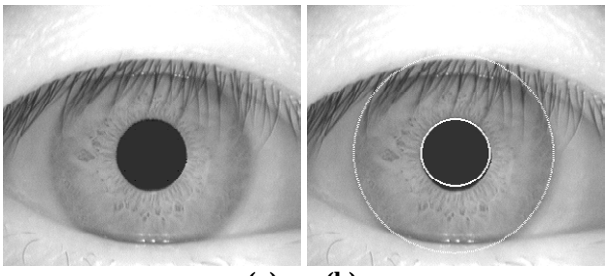
Matching is performed using Hamming distance measure defined in eq. 5.

$$HD = \frac{1}{N}\sum_{j=1}^{N} X_j \oplus Y_j \quad (5)$$

The result is the no. of bits that are different between the binary codes X_j *and* Y_j. If the hamming distance between two images is 0, provided that there have not been any noise patterns in the image while it was segmented and normalized, the two images are from same subject and same eye. However, it is the ideal case and even in most perfect conditions, it is not the case. Hamming distance in practical conditions, i.e. considering some amount of noise is also available while acquiring the image, varies in the range (0,1]. It is measured against a pre defined threshold value which says if the calculated hamming distance is greater than the threshold this means the two images are not from the same subject else the images are from the same subject.

Experiments and results

In the experiments, we implemented the method described by Daugman [13] which is composed of four main stages. In the segmentation we implemented the integro-differential operator that searches for both iris and pupil borders. Feature extraction was accomplished through the use of two dimensional Gabor filters followed by a binarization process. Finally, feature comparison was made through the Hamming distance. A data set of grayscale iris digital images provided by he Chinese Academy of Sciences (CASIA) is used for testing. The CASIA [14] version 1 database consists of 756 gray scale images coming out of 108 distinct classes and 7 images of each eye. We carried out our experiments in MATLAB 7.8 (R2009a). Elapsed time for the iris preprocessing (segmentation and normalization) followed by feature extraction and encoding and hamming distance matching is 133.7 seconds with an average hamming distance 0.3486. Results of segmentation and normalization process are shown in fig. 3.



(a) **(b)**

(c)

Fig. 3. (a) Original Iris Image, (b) Segmented Iris Image, (c) Normalized Iris Image

Results in the form of graph between frequency and Hamming distance are shown for both genuine and imposters separately as well as combined. Curve between FAR and FRR is also shown. Example results are here shown which are performed on a subset of 112 images of CASIA database from 16 different subjects

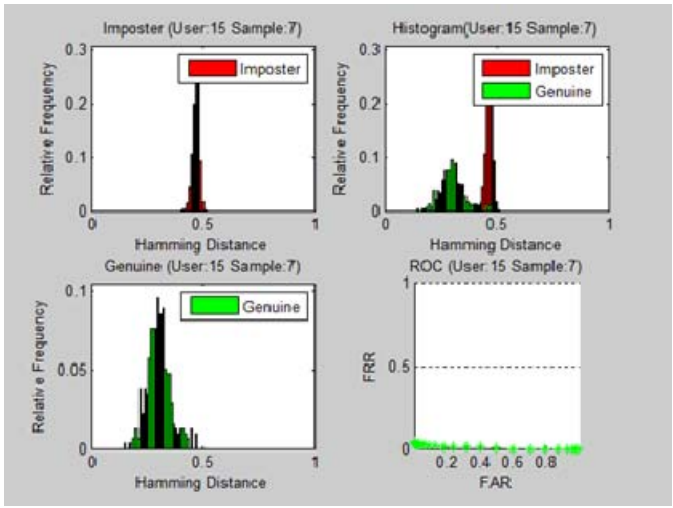


Fig. 4. (a) Hamming Distance vs. Relative Frequency for imposter, for genuine, for both imposter and genuine combined, and FAR vs. FRR

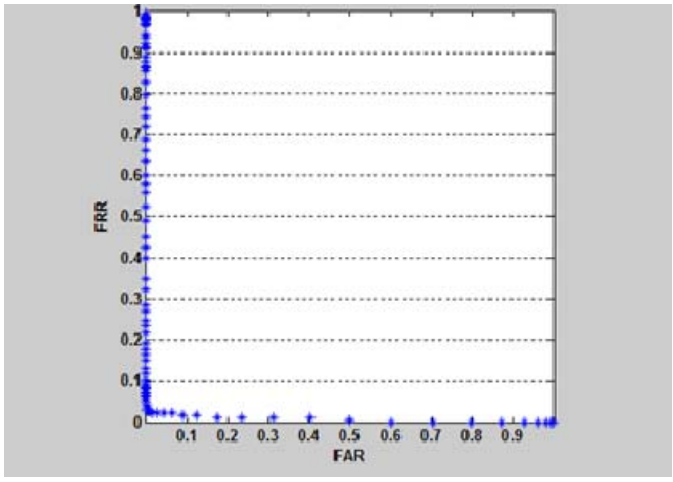


Fig. 4. (b) FAR vs. FRR (shown separately)

Conclusion

Accuracy of the results depends upon how effective segmentation and normalization is done in preprocessing stage of iris recognition. Various researchers have contributed

significant amount of research for developing a constraint free iris recognition system and a lots of research is currently going on this direction. Basic need of iris recognition is valid input iris image which can be preprocessed accurately and efficiently so that normalization and other later stage functionalities discussed above can be implemented and handled with effectiveness. Eyelashes removal and other noise diminishing methods are not considered which shows there is some scope of development available in proposed approach too. Angular deflection also affects the recognition performance of the system which is not considered in the proposed work and can be taken as a future scope of the work.

References

[1] J. G. Daugman, "High confidence visual recognition of person by a test of statistical independence," IEEE Trans, PAMI 15(11), 1148-1161 (1993).

[2] J. G. Daugman, "The importance of being random: statistical principles of iris recognition," Pattern Recognition, 36(2), 279-291 (2003).

[3] Yong Zhu, Tieniu Tan and Yunhong Wang, "Biometric Personal Identification Based on Iris Patterns," National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, P. R. China (2003).

[4] R. Wildes, "Iris recognition: An emerging biometric technology," Proceedings of the IEEE, 85(9):1348–1363 (1997).

[5] W. Boles and B. Boashash, "A human identification technique using images of the iris and wavelet transform," IEEE Trans. Signal Processing, 46(4):1185–1188 (1998).

[6] S. Lim, K. Lee, O. Byeon, and T. Kim, "Efficient iris recognition through improvement of feature vector and classifier," ETRI Journal, 23(2):61–70 (2001).

[7] S. Noh, K. Pae, C. Lee, and J. Kim, "Multiresolution independent component analysis for iris identification," In Proceedings of ITC CSCC'02, 1674–1678, (2002).

[8] C. Tisse, L.Martin, L. Torres, and M. Robert, "Person identification technique using human iris recognition," In Proceedings of ICVI'02, 294–299, (2002).

[9] L. Ma, Y. Wang, and T. Tan, "Iris recognition using circular symmetric filters," Proceedings of the 25th International Conference on Pattern Recognition, vol. 2, pp. 414–417, (2002).

[10] X. Yuan and P. Shi. A non-linear normalization model for iris recognition. In Proceedings of the International Workshop on Biometric Recognition Systems IWBRS 2005, pages 135–142, China, 2005.

[11] Libor Masek and Peter Kovesi, MATLAB Source Code for a Biometric Identification System Based on Iris Patterns, The School of Computer Science and Software Engineering, The University of Western Australia, 2003, http://www.csse.uwa.edu.au/pk/studentprojects/libor/sourcecode.html.

[12] Flom, L. and Safir, A., "Iris Recognition System", US Patent 4,641,349, 3 Feb. 1987.

[13] J. G. Daugman, "How iris recognition works," Pattern Recognition, 36(2), 279-291 (2003).

[14] Institute of Automation, Chinese Academy of Sciences. CASIA iris image database, 2004. http://www.sinobiometrics.com

[15] J.Canny, "A computational approach to edge detection," IEEE Transactions on Pattern analysis and Machine Intelligence, 8: 679-698, November 1986.

Multi-Variant Spatial Outlier Approach to Detect Less Developed Sites in Given Region

Ankita Sharma[1] and Arvind Sejwal[2]

[1]M.Tech, Research Scholar, [2]Assistant Professor & Head
Department of Computer Science & Engineering
Ambala College of Engineering & Applied Research, Devsthali, Ambala, India
E-mail: ankitasharma788@yahoo.com, arvind_ace11@rediffmail.com

Abstract
The term "outlier" can generally be defined as an observation that is significantly different from the other values in a data set. The outliers may be instances of error or indicate events. The task of outlier detection aims at identifying such outliers in order to improve the analysis of data and further discover interesting and useful knowledge about unusual events within numerous applications domains. A Spatial Outlier (SOutlier) is an object whose non-spatial attribute value is significantly different from the values of its spatial neighbors Spatial Outlier detection techniques may be used in many real life applications like geographical information system (GIS), climate prediction, fire detection and etc .In this paper we have discussed how the spatial outlier detection technique may be used to detect less developed sites in given region. We have used multiple non-spatial attributes of many spatially distributed sites. We have applied two very popular mean and median based spatial outlier detection technique on a real data set of twenty one sites in the state of Haryana. Results of these techniques may be used by development planners to make effective and efficient decisions. Strategic planners will have an idea regarding less developed spatially distributed sites, where they may put special attention for development.

Introduction
Outliers can be defined as observations which appear to be inconsistent with the remainder of the dataset. They deviate too much from other observations. Outlier detection is a data mining technique like classification, clustering, and association rules. Recently, a few studies have been conducted on spatial outlier detection for large datasets. [4].

This paper focuses on the question how SOutlier can be detected. There are many known algorithms for detecting outliers, but most of them are not fast enough when the underlying probability distribution is unknown, the size of the data set is large, and the number of dimensions in the space is high. There are, however, applications that need tools for fast detection of outliers in exactly such situations. Planners are concerned about development; it provides a decision support for development process.

Literature Survey
Spatial Data Mining Techniques has been used to reveal valuable information from large spatial data sets in many real applications. Spatial objects cannot be simply abstracted as isolated points. Such techniques have been used in many

Real life applications like Geographical information system (GIS), Climate prediction, fire detection and etc. These techniques may also be used in real life applications such as to detect less developed region based upon parameters like size, population density, sex ratio, literacy rate and etc.

Outlier Detection Approaches
The existing approaches to outlier detection can be classified into five categories:

- Distribution based
- Depth-based
- Clustering-based
- Distance based
- Density-based

Clustering-based approaches detect outliers as by-products [5]. Some clustering algorithms such as CLARANS, DBSCAN [2] [3], CURE [4] have the capability of handling exceptions. However, since the main objective of the clustering algorithms is to discover clusters, they are not developed to optimize outlier detection.

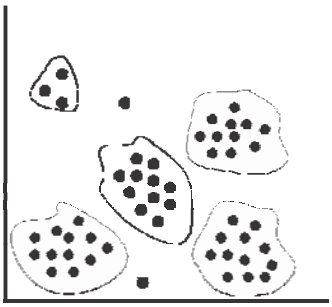


Fig. 1. An example of clusters of points.

Problem Formulation
Taking a real data set, as there are several states in Haryana, where variation in growth and development is noticed. Taking the inspiration from that, we have collected a dataset of non

spatial and spatial attributes for each district. Further SOutlier detection algorithm is applied on the dataset to detect the region which requires more development to be made.

Suppose we have a dataset of n districts

D= (d1, d2,.....dn)

d_i= districts in the state of Haryana or sites which are spatially distributed.

In our problem n=21,

Where d is district with r spatial attributes and m non spatial attributes as given below.

$d_i = (S_1, S_2....S_r, A_1, A_2 .... A_m)$

Where S_1,... S_r is spatial attributes and A_1,.....Am is non spatial attributes.

We want to find out a set of j sites called as spatial outlier sites say SO_j such that SO_j Є D and j<n.

Table 1: Typical non-spatial parameters.

State	Total Population				
District	Male	Female	0<6	Growth Sex	Sex Ratio
HARYANA	25353081	13505130	11847951	19.9	877
Ambala	1136784	604044	532740	12.1	882
Bhiwani	1629109	864616	764493	14.3	884
Faridabad	1798954	961532	837422	31.7	871
Fatehabad	941522	494834	446688	16.8	903
Gurgaon	1514085	817274	696811	73.9	853
Hisar	1742815	931535	811280	13.4	871
Jhajjar	956907	514303	442604	8.7	861
Jind	1332042	712254	619788	12.0	870
Kaithal	1072861	570595	502266	13.4	880
Karnal	1506323	798840	707483	18.2	886
Kurukshetra	964231	510370	453861	16.8	889
Mahendragarh	921680	486553	435127	13.4	894
Mewat	1089406	571480	517926	37.9	906
Palwal	1040493	553704	486789	25.5	879
Panchkula	558890	298919	259971	19.3	870
Panipat	1202811	646324	556487	24.3	861
Rewari	896129	472254	423875	17.1	898
Rohtak	1058683	566708	491975	12.6	868
Sirsa	1295114	683242	611872	16.0	896
Sonipat	1480080	798948	681132	15.7	853
Yamunanagar	1214162	646801	567361	16.6	877

TABLE 2 Typical non-spatial parameters (cont.)

State/district	LITERACY RATE				
	%age 0-6	Sex ratio	Total	Male	Female
HARYANA	13.0	830	76.6	85.4	66.8
Ambala	10.9	807	82.9	88.5	76.6
Bhiwani	12.6	831	76.7	87.4	64.8
Faridabad	13.2	842	83.0	89.9	75.2
Fatehabad	12.6	845	69.1	78.1	59.3
Gurgaon	13.1	826	84.4	90.3	77.6
Hisar	12.1	849	73.2	82.8	62.3
Jhajjar	12.1	774	80.8	89.4	71.0
Jind	12.4	835	72.7	82.5	61.6
Kaithal	12.6	821	70.6	79.3	60.7
Karnal	12.9	820	76.4	83.7	68.3
Kurukshetra	12.0	817	76.7	83.5	69.2
Mahendragarh	11.9	778	78.9	91.3	65.3
Mewat	22.3	903	56.1	73.0	37.6
Palwal	16.5	862	70.3	82.6	56.4
Panchkula	11.7	850	83.4	88.6	77.5
Panipat	13.7	833	77.5	85.4	68.2
Rewari	12.5	784	82.2	92.9	70.5
Rohtak	11.9	807	80.4	88.4	71.2
Sirsa	11.9	852	70.4	78.6	61.2
Sonipat	12.7	790	80.8	89.4	70.9
Yamunanagar	11.8	825	78.9	85.1	72.0

TABLE 3 Typical Spatial Parameters

District name	Total area	Longitude	Latitude

Proposed System

The proposed methodology is discussed step by step in figure3 given below.

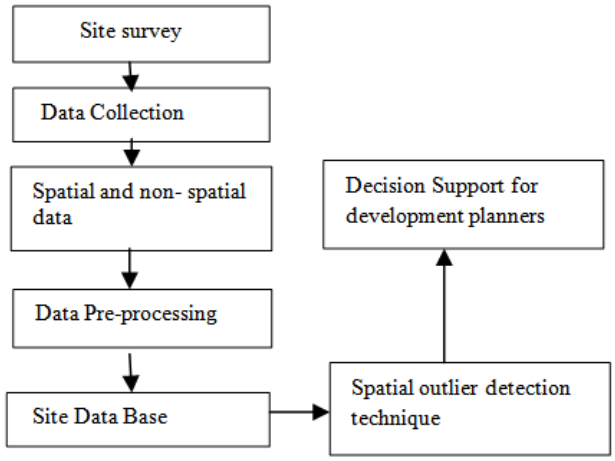


Figure 3. Methodology for spatial outlier detection

Now we are in position to apply a suitable spatial outlier detection algorithm. Many outlier detection algorithms are available. We discuss following two such important algorithms.

Mean Algorithm

1. Given the spatial data set X = {x_1, x_2, x_n}, predefined threshold θ, attribute function f, and the number k of nearest neighbours
2. for each fixed j ($1 \le j \le q$); standardize the attribute function fj, i.e., fj (xi) ← fj (x_i) –μfj

Σfj for i = 1, 2, n.

3. For each spatial point xi, compute the k nearest

neighbour set NNk (x_i)

4. For each spatial point x_i, compute the neighbourhood function g such that gj (x_i) = average of the data set {fj(x): x →NNk (xi)}, and the comparison function h (x_i) = f (x_i) − g(x_i).

5. Compute d2 (xi) = (h (xi) − μs) TΣ−1s (h (x_i) − μs).

If d2 (x_i) > θ, x_i is a spatial outlier w.r.t. A.

Median Algorithm

1. Given the spatial data set X = {x_1, x_2, x_n}, predefined threshold θ, attribute function f, and the number of nearest neigh
2. for each fixed j (1 ≤ j ≤ q); standardize the attribute function fj, i.e., fj (x_i) ← fj (x_i) −μfj.

Σfj for i = 1, 2, . . . , n.

3. For each spatial point xi, compute the k nearest neighbor set NNk (xi) based on its spatial location.
4. For each spatial point xi, compute the neighborhood function g such that gj (xi) = median of the data set {fj(x): x → NNk (x_i)}, and the comparison function h (xi) = f (x_i) − g (x_i).
5. Compute d2 (x_i) = (h (x_i) − as) TΣ−1s (h (x_i) − μs).

If d2 (xi) > θ, xi is a spatial outlier w.r.t. A.

Conclusion & Future Directions

In this paper we have discussed SOutlier detection techniques to solve a real life problem. These SOutlier sites will be very useful for the strategic planners to make efficient decisions regarding development work in a region. The efficiency of the proposed system may also be improved using some other better outlier detection techniques. In future we will implement these techniques to detect outlier sites using real data set.

Acknowledgement

This work would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this work. First and foremost, our utmost gratitude to Dr. JK Sharma (Director, ACE) whose sincerity and encouragement we will never forget.

References

[1] M. M. BREUNIG, H-P. KRIEGEL, R. NG, J. SANDER, LOF: Identifying Density-Based Local Outliers, ACM SIGMOD Int. Conf. on Management of Data, Dallas, TX; 2000, pg. 93–104

[2] M. ESTER, H-P. KRIEGEL, J. SANDER, X. XU, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, In: Proceedings of the 2nd Int. Conference on Knowledge Discovery and Data Mining, Portland, OR; 1996

[3] M. ESTER, H-P. KRIEGEL, J. SANDER, X. XU, Clustering for Mining in Large Spatial Databases, KI Journal (Artificial Intelligence), Special Issue on Data Mining 1998; 12 (1), pg. 18–24.

[4] S. GUHA, R. RASTOGI, K. SHIM, CURE: An Efficient Clustering Algorithm for Large Databases, In: Proc. ACMSIGMOD Int. Conf. on Management of Data, Seattle, WA; 1998. pp. 73–84.

[5] A. JAIN, M. MURTY, P. FLYNN, Data Clustering: A Review, ACM Computing Surveys, 1999, 31(3), pp. 264–323.

[6] T. JOHNSON, I. KWOK, R. NG, Fast Computation of 2-Dimensional Depth Contours, In: Proc. 4th. Int.Conf. on KDD, New York, NY, 1998, pp. 224–228.

[7] E. M. KNORR, R. T. NG, Algorithms for Mining Distance-Based Outliers in Large Datasets, In: Proc. 24th Int. Conf. Very Large Data Bases, New York, NY; 1998, pp. 392–403.

[8] E. M. KNORR, R. T. NG, V. TUCAKOV, Distance-Based Outliers: Algorithms and Applications, Journal: Very Large Data Bases, 2000, 8 (3-4), pp. 237–25

[9] Arvind Sejwal, Application of Spatial Outlier Detection Technique to Detect Ambiguous Sites to Establish an Industry, Department Of Computer Science & Engineering, Ambala College of Engg & Applied Research, Devsthali, Ambala

[10] www.censusindia.gov.in

[11] Karmakers A, Syed M. Rahman,"Outlier Detection in Spatial Databases Using Clustering Data Mining", Sixth International Conference on Information Technology: New Generations-2009, DOI 10.1109/ITNG.2009.198, 2009 IEEE explore.pp 1657-1658.

[12] Ma Yiming et al. "Toward Managing Uncertain Spatial Information for Situational Awareness Applications" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 10, OCTOBER 2008, pp 1408-1423.

[13] S. Sotoodeh, "Hierarchical Clustered Outlier Detection in Laser Scanner Point Clouds", IAPRS Volume XXXVI, Part 3 / W52, 2007, pp 383-387.

[14] Thomas Binu and G Raju , "A Novel Fuzzy Clustering Method for Outlier Detection in Data Mining" International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009, pp 161-165.

Static Data Mining Algorithm with Progressive Approach for Mining Knowledge

Shilpa[#1] and Sunita Parashar [*2]

[#]Student, Department of Computer Science & Engineering
Haryana College of Technology & Management, Kaithal, Haryana, India
E-mail: [1]shilpa.goel12@gmail.com
[*]Associate Professor, Department of Information Technology,
Haryana College of Technology & Management, Kaithal, Haryana, India
E-mail: [2]sunita.tu@gmail.com

Abstract
Frequent itemsets generation is an important area of data mining. This paper is concerned with applying progressive approach to extract interesting information from a static database using dynamic approach. This provides an intelligent environment to discover frequent itemsets while reading a particular set of transaction from static database. We performed extensive experiments and calculate the execution time to generate frequent itemsets on the basis of support and number of transaction read at a time.

Keywords: Static data mining, Dynamic data mining, Support, Number of transactions read at a time, Execution time.

Introduction

With the rapid growth in size and number of available databases in commercial, industrial, administrative and other applications, it is necessary and interesting to examine how to extract knowledge automatically from huge amount of data [1]. Knowledge discovery in databases (KDD), or Data Mining, is the effort to understand, analyze, and eventually make use of huge volume of data available**.** Data mining is the discovery of hidden information found in databases and can be viewed as a step in overall process of Knowledge Discovery in databases (KDD) [2][3]. It is the integration of various techniques from multiple disciplines such as statistics, machine learning, pattern recognition, neural networks, image processing, and database management system and so on[4]. It makes use of various algorithms to perform a variety of tasks. These algorithms examine the sample data of a problem and determine a model that fits close to solving the problem. The models that we determine to solve a problem are classified as predictive and descriptive [5][6]. Predictive mining tasks perform inference on current data in order to make predictions. The data mining task that forms the part of predictive model are Classification, Regression, and Time series analysis. Descriptive mining tasks characterize the general properties of the data in the database. This enables us to determine the patterns and relationships in a sample data. A data mining task that forms the part of descriptive model are Clustering, Summarization, Association rules, Sequence discovery. Classification derives a function or model that describes and distinguishes data classes or concepts, which determines the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data. The training data includes data objects whose class label is known. Regression is to forecast future data values based on present and past data values by means of mathematical formula. Time series analysis is to predict future values for current set of values that are time dependent. Clustering identifies the classes also called clusters or groups for the set of objects whose classes are unknown. The objects are so clustered that the intraclass similarities are maximized and the interclass similarities are minimized. This is done based on the criteria defined on the attributes of the objects [1][6]. Summarization is the abstraction or generalization of data. This results in a smaller set, which gives a general overview of data, usually with aggregated information. Summarization is used to summarize huge amount of data containing in a web page or document. The summarization can go to different abstraction levels and can be viewed from different angles. It is also known as characterization or generalization. Association rule mining is to generate correlation between large unclassified data items based on certain attributes and characteristics and association rule. Association rules are used to identify relationship among a set of items in database of transactions on the basis of large itemsets [7]. Sequence discovery is to determine the sequential patterns that exist in data by using the time factor.

In data mining, with the increasing amount of data stored in real application system, the discovery of association relationship (Association Rule mining) attracts more and more attention. Mining for association rules can help in business, decision making, and the development of customized marketing programs and strategies. Thus goal of data mining is to turn “data into knowledge” [8].Therefore, mining association rules from large database has been a focused topic in recent research into knowledge discovery in databases [9].

Database can be static and dynamic. Static databases are those databases that do not change with time while in dynamic databases, new transactions append as time advances. This may introduce new frequent itemsets and some existing

frequent itemsets may become invalid. Thus, the maintenance of large itemsets for dynamic databases is very costly if re-run of previous mining algorithm on updated database is applied because it repeats much of work done in previous computations. Furthermore, there is not enough space to store all the data for its processing. So instead of finding large itemsets again some heuristics are used for mining dynamic databases [10].

This paper is organized as follows. . In Section 2, related work to the new algorithm is discussed In Section 3, Static Data Mining algorithms are discussed. In Section 4, Dynamic Data Mining algorithms are discussed. In Section 5, progressive approach for mining is discussed. In Section 6, results related to current work are discussed. In Section 6, the paper is concluded.

Related Work

Static data mining algorithms like Apriori, Fp-Growth, Fast Algorithm, Partition Based Algorithms apply only on original database. If there is a need to modify or delete some or all the existing set of data during the process of data mining then repetition of whole procedure is required, which is time-consuming in addition to its lack of efficiency. So incremental update methods like Fast Update, Probability based & Promising based algorithms are used to extract interesting information from dynamic databases. On the basis of this, new approach (PAPRIORI) can be used that takes original database progressively i.e. read a particular set of transactions at a time while we know the size of original database. PAPRIORI is static data mining algorithm that uses dynamic approach. Since execution time to generate frequent itemsets remains a great challenge, so the goal is to calculate the execution time of proposed approach at varying value of number of transactions read at a time (K).

Static Data Mining

Data Mining that uses static database for mining is known as static data mining. There are different static data mining algorithms like Apriori, Fp-Tree, Fast algorithm, Partition based algorithm etc.

Apriori Algorithm

Apriori is the most widely accepted static data mining algorithm [7][9]. This is described as a “fast algorithm for mining association rules”. Apriori algorithm is driven by market-basket data. It efficiently generates large itemsets along with generation of candidate itemsets by repeatedly scanning the database. Apriori algorithm is based upon candidate set generation and test method. The problem that always appears during mining frequent relations is multiple scans of original database, huge number of candidate generation and tedious workload of support counting for candidates. So there is need to reduce passes of transaction database scans, to shrink number of candidates and to facilitate support counting of candidates.

FP-Growth Algorithm

FP-Tree is an order of magnitude faster than the Apriori algorithm. This is used for mining static databases. In this, the frequent patterns generation process includes two sub processes: constructing the Fp-Tree, and generating frequent patterns from the FP tree. This uses divide-and-conquer method and takes 2 scans of database [11]. Candidate itemsets generation does not occur in this.

Fast Algorithm

Most time consuming operation in the discovery of association rules from the database is the computation of the frequency of the occurrences of interesting subset of items called candidates. So there is need to develop a method that avoids or reduces candidate generation and test and utilizes some novel data structures to reduce the cost in frequent pattern mining. Fast algorithm uses TreeMap which is a structure in java that store key / value pair[12]. Moreover Arraylist technique that greatly reduces the need to traverse the database is also used. This reduces usage of memory.

Partition Based Algorithm

Partition based algorithm divides the database into partitions that reduces the number of database scans to two. This algorithm reduces both CPU and I/O overheads [13]. This algorithm is especially suitable for very large size databases. During first scan, divide database into partitions and generate frequent itemsets in different partitions separately by scanning the database once in each partition. During second scan, counters for each of these itemsets are set up and their actual support is measured to determine if they are large across entire database. If the items are uniformly distributed across partitions then a large fraction of itemsets will be large.

Dynamic Data Mining

Data Mining that uses dynamic databases that take into considerations all updates (insert, update, and delete problems) into account is known as dynamic data mining. There are different dynamic data mining algorithms like Fast Update (FUp), incremental method like promising based algorithm and probability based algorithm.

Fast Update Algorithm

An incremental updating technique FUp (Fast Update) algorithm is used for efficient maintenance of discovered association rules when new transactional data are added to a transaction database [14]. In this, we seperate winners (those that remain large in updated database) from losers (that are not large in updated database) among large items in original database and find new winners that are large in original database (DB) and incremental database (db) i.e. (DB U db). This algorithm is 2 to 16 times faster than Apriori.

Promising Based Incremental Approach

Promising frequent itemset algorithm, an incremental method,

is proposed for dynamic data mining [15]. This algorithm uses maximum support count of 1-itemsets obtained from previous mining to estimate infrequent itemsets, called promising itemsets, of an original database. These itemsets are capable of being frequent itemsets when new transactions are inserted into the original database. Thus, the algorithm reduces a number of times to scan the original database. As a result, the algorithm has execution time faster than that of previous methods like FUP (Fast Update).

Probability Based Incremental Approach

Probability-based incremental association rule discovery algorithm is used to extract interesting information from dynamic databases [16]. This uses principle of Bernoulli trial to find expected frequent itemsets that reduces number of scans to original database. This proposes a new updating and pruning algorithm that guarantee to find all frequent itemsets of an updated database efficiently. The results show that this algorithm has better performance than that of FUp (Fast Update).

New Static Data Mining Algorithm(PAPRIORI)

PAPRIORI algorithm generates frequent itemsets progressively in static database by means of reading K transactions at a time. It is based upon basic data mining algorithm(Apriori). For first K transactions m large itemsets will be generated then for next K transactions m, m+1 large itemsets will be generated progressively and so on. This is based on the following considerations.

- The itemsets that are counted initially or does not satisfy minimum support are Estimated Infrequent (EI) itemsets.
- The itemsets that satisfy minimum support threshold are Estimated Frequent (EF) itemsets.
- CF (Confirmed Frequent) itemsets are those that have been counted throughout whole database once and satisfy minimum support.
- CI (Confirmed Infrequent) itemsets are those that have been counted throughout whole database once and do not satisfy minimum support.
- Following are the algorithmic steps:

Step 1: Set all 1-itemsets as Estimated Infrequent (EI) itemsets.

Step2: Read database with K transactions at a time (until transactions read is less than total number of transactions in database).

- For each transaction, increase counter for the itemset.
- For each itemset that belongs to EI if value of counter satisfies minimum support then set itemset as EF.
- If itemsets belong to EF or CF then their immediate superset is set as EI.
- For each itemsets that belongs to EF if it is read throughout the whole database once move that into CF.
- On the other hand if itemsets belongs to EI, if it is read throughout the whole database once move it into CI.

This is repeated until Estimated Frequent (EF) and Estimated Infrequent (EI) itemsets are present.

Experimental Setup

To evaluate the performance of PAPRIORI algorithm, the algorithm is implemented and tested on a workstation with Pentium(R) Dual-Core CPU, 2.19 GHz and 2.93GB main memory. The experiments are conducted on a Synthetic dataset and Zoo dataset. The Synthetic dataset comprises 1,000 transactions over 10 items. The Zoo dataset comprises 101 transactions over 15 items. Proposed algorithm is used to find frequent itemsets from static database consisting of transactions. Set fixed value of support for both datasets and vary number of transactions read at a time (K) to calculate execution time.

Results for Synthetic Dataset

On the basis of K and execution time the following graphs with fixed value of support (50%, 45%) can be drawn for analysing the results.

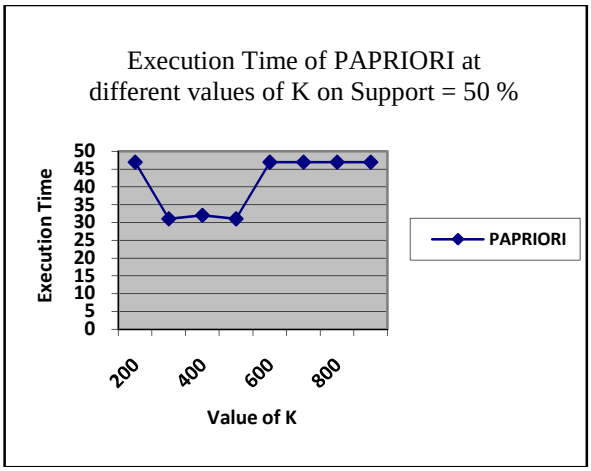


Figure 1 Execution Time with Support = 50%

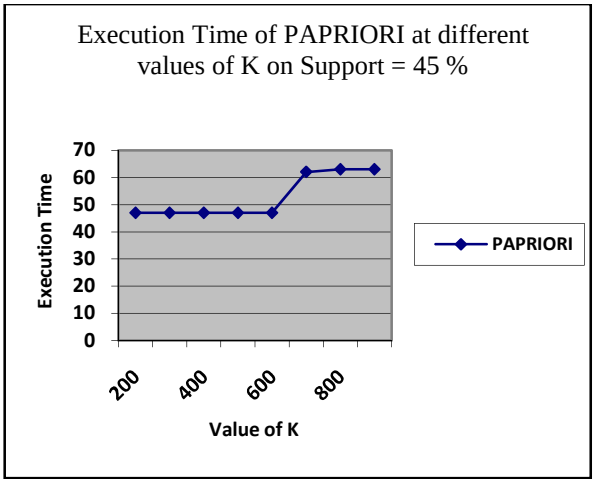


Figure 2 Execution Time with Support = 45%

Results for Zoo dataset

On the basis of K and execution time the following graphs with fixed value of support (50%, 55%) can be drawn for analysing the results.

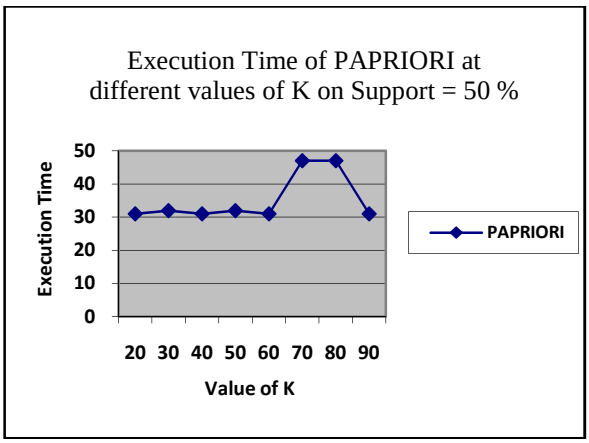


Figure 3 Execution Time with Support = 50%

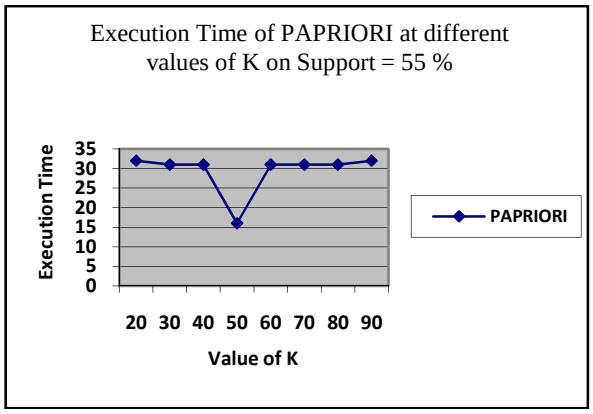


Figure 4 Execution Time with Support = 55%

It is obtained from the Figure 1, Figure 2, Figure 3 and Figure 4 that at intermediate value of K, execution time of PAPRIORI algorithm is less. So selection of right value of K is required. If value of K is very less, no frequent itemsets can be obtained easily and execution time will increase. On the other hand, if value of K is very large then again execution time increases and it behaves like Apriori Algorithm.

Conclusion

Mining knowledge from database is both practical and desirable. We have proposed static data mining algorithm that generates itemsets progressively with less execution time at intermediate number of transactions read. In the future, further researches and experiments on the proposed algorithm will be presented.

References

[1] M. Dunham. "Data Mining – Introductory and Advanced Topics". Pg 185-186. Section 6.7.2. Pearson Education. 2003.

[2] B.N. Lakshmi , G.H. Raghunandhan," A Conceptual Overview of Data Mining", Proceedings of the National Conference on Innovations in Emerging Technology, pp.27-32, February 2011.

[3] Qi Luo, "Knowledge Discovery and Data Mining," in Proc. Workshop on Knowledge Discovery and Data Mining, Adelaide, SA , 2008, pp 3-5,IEEE.

[4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence Magazine, pp. 37-54, 1996.

[5] V.Umarani, Dr.M.Punithavalli, "A Study on Effective Mining of Association Rules From Huge Databases", IJCSR International Journal of Computer Science and Research, Vol. 1 Issue 1, 2010, pp 30-34.

[6] Jiawei Han and Micheline Kamber, "Data Mining: Concept and Techniques," N. Harcourt India Private Limited ISBN: 81-7867-023-2,2nd Edition, 2001.

[7] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases". In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pages 207-216, Washington, DC, May 26-28,1993.

[8] Tian Lan, Runtong Zhang and Hong Dai, "A New Frame of Knowledge Discovery," in Proc. 1st International Workshop on Knowledge Discovery and Data Mining, WKDD 2008, Jan. 2008, pp 607 – 611.

[9] Rakesh Agrawal & Ramakrishan Srikant," Fast algorithm for mining Association rules**",** IBM Almaden Research Center, 650 Harry road, San Jose, CA 95120: In proceedings of the 20th VLDB conference Santiago, Chile, pp 487-499,1994.

[10] Hebah H. O. Nasereddin, "Stream Data Mining", International Journal of Web Applications, Volume 1, No. 4, December 2009, pp183-190.

[11] J. Han, J. Pei, and Y. Yin." Mining frequent patterns without candidate generation", in W.Chen, J. Naughton, and P. A.Bernstein, editors, 2000 ACM SIGMOD Intl. Conference on Management of Data, Vol. 29, No.2 pp 1-12.

[12] M.H.Margahny and A.A.Mitwaly," Fast Algorithm for Mining Association Rules", AIML 05 Conference, pp 19-21, December 2005, CICC, Cairo, Egypt.

[13] Ashok Savasere, Edward Omiecinski, Shamkant Navathe," An Efficient Algorithm for Mining Association Rules in Large Databases", in proceedings of 21st VLDB Conference , Zurich , Switzerland, pp432-444, 1995.

[14] David W. Cheung, Jiawei Han, Vincent T. Ngt C.Y. Wongj," Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique", in proceedings of the 12th ICDE, New Orleans, Louisiania (IEEE) , pp 106-114,February 1996.

[15] Ratchadaporn Amornchewin, Worapoj Kreesuradej," Incremental Association Rule Mining Using Promising Frequent Itemset Algorithm", 6th International Conference on Information, Communications & Signal Processing (ICICS), 2007, IEEE, pp1-5.

[16] Ratchadaporn Amornchewin, Worapoj Kreesuradej," Mining Dynamic Databases using Probability-Based Incremental Association Rule Discovery Algorithm, Journal of Universal Computer Science, pp 2409-2428,Vol. 15, No.12, 28 June 2009.

A Critical Review of Data Warehouse

Sachin Chaudhary[1], Devendra Prasad Murala[2] and V. K. Srivastav[3]

Department of Master of Science, Asia Pacific Institute of Information Technology, SD, India
Panipat-132103 [Haryana] India.
E-mail: [1]dynamic.chaudhary@gmail.com, [2]murala7@gmail.com, [3]virendra@apiit.edu.in

Abstract
Data warehousing and OLAP have become the most important aid for the decision makers of any industry. Basically Data warehousing refers to collecting and storing historical data into single repository, which is known as Data warehouse and using that warehouse to produce Analytical results. Being the helping hand for the top level professional, it is continuously under the focus of Database industry and posing new challenges to the database industry day by day. In this paper we present the critical review of the Data warehousing along with different kind of architectures and the data modelling of the data warehouse. We described some of the current tools and techniques available at present for data warehousing in terms of the front end and backend tools. We further analysed problems and issues and identified some of the research areas in the field of data warehousing.

Keywords: Data Warehouse, Online Analytical Processing (OLAP).

Introduction

Data warehouse is a Data repository containing historical data from heterogeneous sources. It is designed for query and analysis rather than for transaction processing. In addition to this Data warehousing concept consists of the tools and techniques available for Extraction, Transformation and loading, an OLAP engine, client analysis tools and other applications that are used to manage and process the data to provide decision support to the knowledge workers or decision makers. (Managers, analyst etc.)

According to William H.Inmon, a well known Data warehouse architect, "A Data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process. [6]. This definition separates the Data warehouse from the other Data repository system for example relational database system, Transaction system and File system.

Data Warehouse is a step towards making the computer system able to analyse the trends and help in critical decision making in any organization. Sometimes we get very interesting and useful trend from the historical data that we can use for the future planning. The normal operational databases were meant to provide a help in the clerical operations of the organization but data warehouse and OLAP technologies are meant to provide help to the decisions makers(e.g. Managers, Analyst etc.) of any organization. Therefore new challenges are arising everyday in the field of data warehousing and OLAP to satisfy the demands of the higher professionals.

From last two decades the field of data warehousing has gone through lots of research and changes. From offline operational database to integrated data ware house, it was a long journey, but we still have a long distance to cover. At present we have several areas to improve some of them are identified in this paper. The failure rate of data warehousing projects is still high and if successful the time it is taking is usually more than expected. Therefore we still have to work a lot to achieve a highly efficient data warehousing and OLAP technologies.

In this paper we present a critical review of the data warehousing technology. We described different kind of architectures and the data modelling of the data warehouse. We further analysed the tools and techniques available at present for data warehousing. Some of the major research issues are also identified.

Foundation of Data Warehousing

Data warehousing came into picture as a distinct type of computer database during the late 1980 and early 1990s. The concept of Data warehousing arises to fulfil the demand of the higher management to get analytical results which normal operational database was not providing efficiently. With the improvement in technologies and higher demand from the user the concept of Data warehousing has gone through several fundamental stages namely

- Offline operational Database
- Offline Data warehouse
- Real time Data warehouse
- Integrated Data warehouse.

Architecture of Data Warehousing:

The architecture of Data warehouse depends on the Business process of any organization taking into the account Data consolidation across the organization with security, the level of query requirement management of the Meta, Data modelling and organization, warehouse staging area planning for optimum bandwidth utilization and full technology implementation.

The warehouse architecture may include:
[18]

- Process Architecture
- Data Model architecture
- Technology Architecture
- Information Architecture
- Resource Architecture

***Process architecture*:**
It refers to the process or steps followed in converting row Data into information. It mainly include three sub process which are commonly referred as "ETL" process

Extract: Extracting Data from different sources with proper compression and encryption technique.

***Transform*:** Conversion of the extracted Data from different sources into similar format.

***Load*:** The stages include loading the transformed data into the data warehouse.

***Data model architecture*:**
It is Dimensional Data model, According Georgia University, there are five Data modelling styles for warehouses:

- Independent Data Mart
- Data mart bus architecture with conformed dimensions
- Hub and spoke
- Centralized
- Federated

Technology Architecture
It refers to technological structure of data warehouse. It includes Data base connectivity protocols (ODBC, JDBC, OLE DB etc.), implementation standards in data base management, middleware (based on ORB, RMI, CCOM/DOM etc), network protocols (DNS, LDAO etc), and related technologies.

Information Architecture
It is the structure for step-by-step conversion of the information from one form to another to manage the storage, retrieval, modification and deletion of data in the Data warehouse.

Resource Architecture
It refers to the various resources available for example software resources to maintain and manage data warehouse. The quality of the resource architecture is directly proportional to the performance of the data warehouse system.

Typical model of Architecture of Data warehouse
Above mentioned classification gives an overview of the different kind of attribute that we should keep in our mind to build architecture of a data warehouse. But if we talk about the overall architecture of data warehouse, it is usually multi-tiered architecture. A typical three tier architecture is represented in the following image.

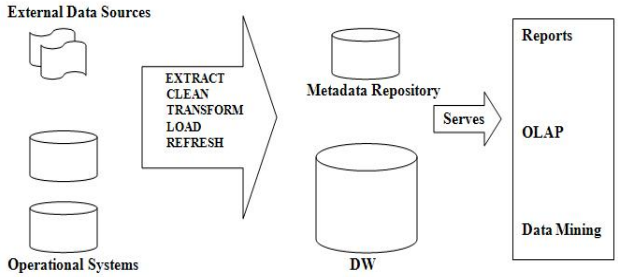


Figure 1: Architecture of Data warehouse

The bottom tier usually consist of several database systems usually relational databases, back end tools and utilities to extract, clean, transform and feed data to the bottom tier from different sources of databases.

The middle tier is an OLAP server, it may either ROLAP, MOLAP or HOLAP server [7.2].

Top tier contains reporting tool, analysis tool, data mining tool.

Multidimensional Data Model
We are very much aware with entity relationship modeling for normal operational Databases but we use different approach known as dimensional modeling for representing the Data warehouse, using the concept of fact and dimension. Basically dimensional modeling is a technique for logical designing of data in a standard, intuitive framework for high performance access composed of one table with a multi-part key, called fact table, and a set of smaller tables called dimension tables.

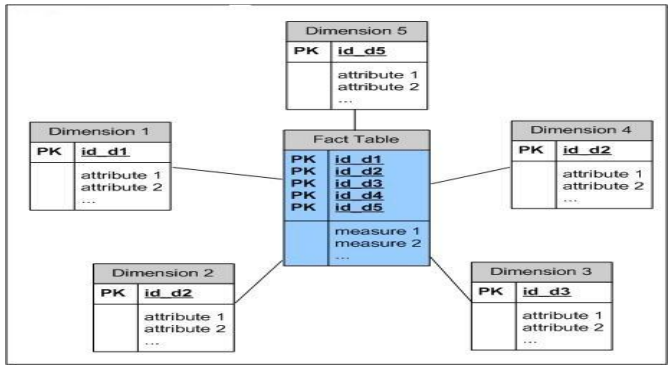


Figure 2: Multidimensional Data model

Fact table has two types of columns one containing fact and other containing foreign key. Facts are numeric measures.

Dimension table is known as looked up reference table. It is the table containing the detail of perspective or entities with respect to which an organization wants to keep record.

Combining the facts and dimensions we get a multidimensional view of the data which is known as data cube. But this cube is n- dimensional not restricted to 3-D like the geometric cube. The multidimensional data modelling has several advantages compare to the conventional relational data

modelling technique using ER diagrams. The figure shows the example of a data cube considering the sales volume as a function of product, month and region. [9]

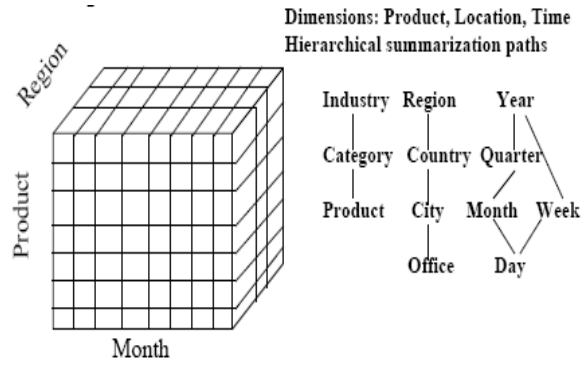


Figure 3: Data cube multidimensional model [9]

Schemas of Multidimensional Model
The multidimensional model can exit in the three schemas

***Star schema*:** According to this schema, the data warehouse contains (a).Large central table (Fact table) containing bulk of data with no redundancy. (b). some called dimension table one for each dimension. When represented on the graph of schema represents as star in which dimension tables are radially arranged around the fact table.[20].

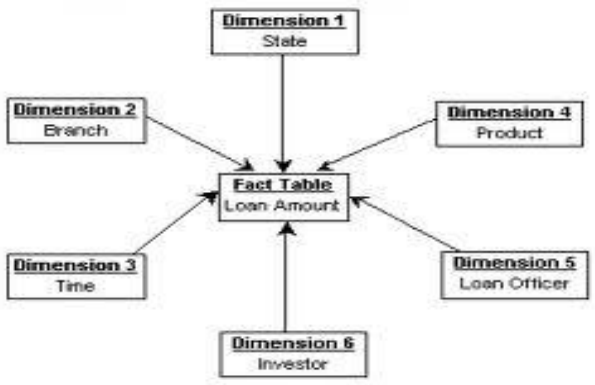


Figure 4: Example of Star Schema

Snowflake schema: It is also like Star schema but the main difference is that in snowflake schema we can normalize the dimension table to reduce the redundancy. This is easy to maintain and save storage space but can reduce the effectiveness of browsing, since more joins will be needed to execute the query.

Fact constellation (Galaxy schema): It is a more complex structure having multiple Fact tables which can share the common dimension table.

Meta Data
Meta data is Data about data. In terms of Data warehouse it defines the objects of warehouse. Meta data is created to explain the following. [2]

- Description of the Data warehouse structure
- Operational metadata
- The algorithm used for summarization
- Mapping from the operational environment to the data warehouse
- Data related to system performance
- Business metadata.

Data Warehouse Models
Enterprise warehouse: It is a warehouse containing data about subject spanning the entire organization. It is usually a huge data warehouse and requires detailed business modelling. It is a data warehouse containing the data of all the subjects related to the entire organization.[15].

Data mart: It is the subset of the enterprise data warehouse containing the data about specific subject that of value to the specific group of users. They contain information about specific subject only.

Virtual warehouse: It is built over the operational databases as a set of views. It is basically the set of views over operational database.

Tools and Techniques:
Data Warehousing Tools can be divided into the following categories.

Back End Tools and Utilities: These tools are also Generally Known has ETL (Extraction, Transform, Load) tool, these tools are used to perform the following operations:

- Data extraction
- Data cleaning
- Data Transformation
- Load
- Refresh

Some of the most important tools used in the market are Oracle warehouse Builder (OWB), Microsoft Integration Services (SSIS), Telnet Open Studio, IBM Information Server, IBM Cognos Manager, Open Text Integration Centre, Information Builders, ETL Solutions (ETI) etc. [19]

Conceptual Model and Front End Tools: Front end tool are also known as OLAP tool, there are mainly three types Multidimensional OLAP (MOLAP) and Relational OLAP (ROLAP), Hybrid OLAP (HOLAP). [20].

MOLAP: A cube is aggregated from relational data source. It is faster in generating report as data is pre-aggregated within the cube.

ROLAP: Unlike MOLAP there is no pre-aggregation of Data into the cube. The ROLAP engine may be consider as small SQL generator.

HOLAP: It is the Hybrid of both MOLAP and ROLAP. Some of the Tools available are Business objects, Cognos, Microsoft, Analysis service, micro Strategy, Palo OLAP server.

Problems and Issues

In spite of going through huge amount research during the last decade Data warehouse still have several areas to research and improve. Some of the major issues to be tackled are as follows

1. Data extraction and cleaning are the first step to build a data warehouse. For any kind of database the quality of data is the most important aspect to get the desired output efficiently. Today we have number of tools available for Data extraction and Cleaning but they are not providing the desired efficiency. For getting the quality result it is obvious that we should have the quality data therefore extraction and cleaning of the data to get the quality data is one of keen research area for data warehouse.
2. Data transformation and integration is another area to be researched further as data warehouse is build up using data from heterogeneous sources therefore we should have efficient tools then available at present. This is one of the most important tasks in data warehousing as different databases have different schemas and format and it's a prerequisite to convert them to similar format before loading into the data warehouse. The transformation of data with least error and least loss of information is still to go miles ahead.
3. Maintenance of a data warehouse is another aspect in which we have lot of chances to improve. We should look for some better maintenance technologies along with the software and better hardware to efficiently manage the increasing size of the data warehouse. Management of Meta data should also be researched further.
4. Efficient retrieval of the result is the main aim of any system. In data warehouse we have several technologies available for efficient query processing but still they have to be improved a lot to achieve the required efficiency. Query processing needs to be researched further.

Conclusion

Data warehousing is the basis of automated decision support system. It has been researched a lot in the past decade but still there are many issues to be tackled in future. Performance and management are among the top research issues at present. We have identified some of the latest tools available for data warehousing and classified the tools in logical manner. The architecture of the data warehouse is also divided logically as well as a typical model of the architecture is also given. We further analysed some of the major research areas like data cleaning, data transformation, maintenance and efficient query processing. We identified major research areas in the data warehousing and the things to be done in future to achieve the best out of our data warehousing.

References

[1] Stolba, N., Banek, M. and Tjoa, A.M. (2006): The Security Issue of Federated Data Warehouses in the Area of Evidence- Based Medicine. Proc. of the First International Conference on Availability, Reliability and Security (ARES'06, IEEE), 20-22 April, 2006.

[2] Inmon, W. (2002): Building the Data Warehouse, 3rd edition, Wiley-New York.

[3] SAS© (2002): Building a Data Warehouse Using SAS/Warehouse Administrator®, Software Course Notes (Book code58787). SAS Institute Inc., Cary, NC 27513, USA.

[4] Sen, A. and Sinha, A. P. (2005): A Comparison of Datawarehousing Methodologies, Communication of the ACM, 48(3), 79-84.

[5] Stephen R. (1998) .Building the Data Warehouse.,

[6] Communications of the ACM, 41(9), 52-60 (September 1998).

[7] Inmon, W.H., "What is a Data warehouse?" Prisum solution, Inc, http://www.cait.wustl.edu/cait/papers/prisum/voll_nol/ ,1995

[8] Greenfield, L.,"The Case Against Data Warehouseing" LGI Systems, Inc, http://www.dwinfocenter.org/gotchas.html, June 2001

[9] Greenfield, L.," Data Warehouseing Gotchas" LGI Systems, Inc, , http://www.dwinfocenter.org/gotchas.html, June 2001

[10] Harinarayan V., Rajaraman A., Ullman J.D. " Implementing Data Cubes Efficiently" Proc. of SIGMOD Conf., 1996.

[11] Roussopoulos, N., et al., "The Maryland ADMS Project: Views R Us." Data Eng. Bulletin, Vol. 18, No.2, June 1995.

[12] O'Neil P., Quass D. "Improved Query Performance with Variant Indices", To appear in Proc. of SIGMOD Conf., 1997.

[13] Gupta, A., I.S. Mumick, "Maintenance of Materialized Views: Problems, Techniques, and Applications." Data Eng. Bulletin, Vol. 18, No. 2, June 1995.

[14] Codd, E.F., S.B. Codd, C.T. Salley, "Providing OLAP (On-Line Analytical Processing) to User Analyst: An IT Mandate." Available from Arbor Software's web sitehttp://www.arborsoft.com/OLAP.html.

[15] Inmon, W.H., Building the Data Warehouse. John Wiley, 1992.

[16] J. Hammer, H. Garcia-Molina, J. Widom, W. Labio, and Y. Zhuge. The Stanford Data Warehousing Project. IEEE Data Engineering Bulletin, Special Issue on Materialized Views and Data Ware housing, 18(2):41{48, June 1995.

[17] W.H. Inmon and C. Kelley. Rdb/VMS: Developing the Data Warehouse. QED Publishing Group, Boston,

Massachussetts, 1993.
[18] A. Gupta and I.S. Mumick. Maintenance of materialized views: Problems, techniques, and applications. IEEE Data Engineering Bulletin, Special Issue on Materialized Views and Data Warehousing, 18(2):3{18, June 1995}.
[19] Providing Architecture of the Data warehouse.http://it.toolbox.com/wiki/index.php/Data_ warehouse_Architecture. Providing ETL Back end tools. [http://etltool.com]
[20] Jiawei Han, Micheline kamber, Jian pei “Data mining Concepts and Techniques” Third edition.

Green Database

Sachin Chaudhary[1], Devendra Prasad Murala[2] and V.K. Shrivastava[3]

Department of Master of Science, Asia pacific Institute of Information Technology SD India
Panipat-132103 [Haryana] India.
E-mail: [1]dynamic.chaudhary@gmail.com, [2]murala7@gmail.com, [3]virendra@apiit.edu.in

Abstract
Green is the word of greatest concern these days. Green as clear to almost everyone actually refers to the environment friendliness. We are developing new technologies and new methods for the betterment of our life but we usually forget to take into consideration the adverse effects of those on our environment. With the continuously increasing computing and storage capabilities, we are easing out our day to day life a lot but we never thought of the adverse effects it is posing to our environment. Each one of us is heavily dependent on the search engines, mail servers, social networking sites and other kind of databases in their day to day life, but we never thought of the undesirable effects they are creating to the environment. In this paper, we attracted the focus in the direction of the increasing consumption of power by the data centres with their ever increasing size and analysed the adverse impact on the environment. We have also found out and added some new dimensions to the concept of green computing.

Keywords: Green, Data Centres, Green Database.

Introduction

Every one of us is well aware of the search engines, mail servers, social networking sites and other kind of databases available over the network. We have a severe dependence on the databases now a day. In the last decade, we emphasize on the reduction of paper usage to maintain the records because to produce paper, we need to cut trees which is a threat to the environment. But shifting all our data to storage memories, we have not thought that it could also harm the environment in some or the other ways.

Now, the time has come to realize this thing that by reducing the use of paper, we cannot achieve the Green i.e. we cannot save our environment from the threats posed by the maintenance of data. The storage that we are using these days is also harming the environment in some or the other ways.

In this paper, we explain how data management using large storage devices is posing a threat to the environment and what can be the possible solution to the problem. The size of data present all over the world is increasing with every smallest transaction done anywhere in the world. To maintain huge amount of data, we manage very large data centres across the world. The size of these databases or data centres is continuously increasing. These centres consume a huge amount of power and as the size is increasing their consumption is also increasing. Therefore it's a high time to think about the future of the data centres and the storage devices.

Data Centres

A Data centre is a place, room or building where servers and storages of an enterprise are located, managed and operated. Sometimes data centres are also referred to as server farms as it is consisting of large number of servers. It consists of mainly:

***White space*:** This is a usable area measured in square feet inside the data centres. Usually it is a raised floor environment. The size usually ranges from few hundred to few thousand square feet.

***Support infrastructure*:** This refers to the supporting or complementary infrastructure required to manage data centre.

***Centre operations*:** It comprise of the uninterruptible power source (UPS), generators, computer room air conditioners (CRACs), power transformers, remote transmission units (RTUs), chillers, air distribution systems, etc. In a high density, Tier 3 class data centre (i.e. a concurrently maintainable facility), this support infrastructure can consume 4-6 times more space than the white space and must be accounted for in data centre planning.

***IT equipment*:** This includes the racks, cabling, servers, storage, management systems and network gear required to deliver computing services to the organization.

***Operations*:** The operations staffs assure that the systems (both IT and infrastructure) are properly operated, maintained, upgraded and repaired when necessary. In most companies, there is a division of Responsibility between the Technical Operations group in IT and the staff responsible for the facilities support systems.

Some of the largest data centres in the world

Some the largest data centres in the world are shown in Table 1.With the data shown in Table 1. We get the picture of the mightiness of the data centres and we can easily understand the huge consumption of power in cooling these mighty system. Therefore now it is very clear that we have to think about our databases. Shifting from paper to storage devices did not solved the purpose of being environment friendly completely but now we have to think different way.

Table 1: Largest Data centres in the world [2].

S.No	Name	Location	Size (sq.fet)	Owner	Power consumption in MW
1.	The Lake side Technology centre	Chicago	1.1 million	Digital Reality Trust	100
2.	Metro Technology centre	Atlanta	990,000	Quality Technology services	80
3.	The NAP of the Americas	Miami	750,000	Terre mark	N/A
4.	Next Generation Data Europe	Wales	750,000	BT & Logica	N/A
5.	Microsoft Chicago Data centre	North Lake	700,000	Microsoft	N/A

Effect of Data Centres on Environment

Till date, we are dependent mostly on the non renewable, pollution causing sources for the production of energy. The data centres when operated produces heat but the hardware will give best performance in the specified temperature range only. Therefore, we need a strong cooling system and of course air conditioning as we know that it consumes huge amount of energy. The energy where ever produced using non renewable sources will produce toxic gases and in turn will lead to global warming. Figure 1. shows the cyclic effect of the data centres on environment. The effect of data centre on environment is multi fold. The data centre itself is producing lot of heat, the power consumption by cooling system in also indirectly effecting the environment. Therefore it's a high time look into the problem and find out a better solution for better future.

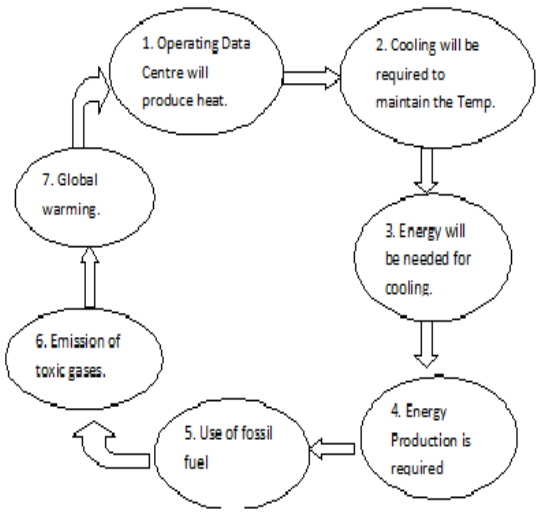


Figure 1: Effects of data centres on Environments.

Solution to the Problem "Green Data Database"

A green database refers to the repository of data with minimal effect to the environment. It is an environment friendly database unlike the present day data centre as explained above. But the question that arises in the mind is how to develop green database. As the effect of data centres on the environment is multi fold, we need to have multi level solution for problem. To achieve green i.e. environment friendliness we need a proper planning and strategy. To look out for the solution we should start from the problem first as follows.

***Heat production during operation*:** The problem starts with the production of heat during the operation of data centre. When the processing of the data will be more, the heat produced will also be more. We can directly relate this problem to the efficiency of the hardware and the software used. Therefore first thing to achieve green is to increase the performance and efficiency of the existing system. This approach is further divided in three parts

- To make new high performance hardware this produces less heat: But this approach is costlier as it's almost impossible to change the total hardware of any data centre. We can use this approach while developing new data centre.
- To utilize the existing resources in such manner so that less heat will be produced: This can be achieved by using the several techniques already available and which will be produced in future. Some of the include Virtualization, de-duplication, Cloud computing. We all are well aware of all these techniques which are used to increase the performance of the computing resources.
- To locate the data centres at a place where temperature is less so that energy spent in cooling will be less.
- Designing of the data centre to consume less power.

***Energy Production for the data centres*:** The data centres consume a huge amount of energy. In a research done by Microsoft daily power consumption of a typical data centre equals the monthly power consumption of thousands of homes. With such a huge consumption of energy they are posing challenges to the energy production technologies at present. We can increase the efficiency of the data centres we can locate them at colder places but still they will need electricity to operate. Therefore it's a high time to think about the power generation strategy for the data centres. Some of the options include:

- Using dedicated nuclear power plants to feed electric supply to the data centres. But nuclear energy has its own drawbacks.
- Another approach is to use non-conventional sources of energy to feed electricity to data centres. This is a costly approach and still not that much efficient to feed all the data centres in the world. But still we can use to feed at some part of our energy consumption.
- The other alternative technologies may include photovoltaic cells, heat pumps, evaporative cooling.

The other things that we can do to achieve "Green" includes waste recycling, use of hybrid company vehicle, the use of low-emission building material, carpets and paints.

Steps to Build a Green Database

Building a green database is strategic process in which you have to plan your steps to make the conversion into green database cost effective and efficient as well. It usually includes the following steps:

Understanding: You should first understand the energy requirement of your data centre.

Design: Based on the understanding you have to develop a plan to build or upgrade the data warehouse to make it energy efficient.

Optimize: Then you have to optimize and load balance the server rooms.

Virtualization: Then to increase performance you have to follow virtualization of the server.

Measure, manage and report: the final steps include the analysis of the developed system and report to the authorities.

Energy Saving Options & Roi

Some of the energy saving options for data centres are analysed and there result are shown in the table below:

Energy Saving Options & ROI:

Table 2: Energy Saving Options & ROI [5]

Energy Saving Action	Saving Independent of Other Actions		Energy Saving With the Cascade Effects			ROI
	Saving (KW)	Savings (%)	Savings (KW)	Savings (%)	Cumulative Savings (KW)	
Lower Power Processors	111	10%	111	10%	111	12-18 months
High-efficiency power supplies	141	12%	124	11%	235	5 to 7 months
Power Management features	125	11%	86	8%	321	Immediate
Blade Servers	8	1%	7	1%	328	TCO reduced 38%
Server virtualization	156	14%	86	8%	414	TCO reduced 63%
415v AC power distribution	34	3%	20	2%	434	2 to 3 months
Cooling best practices	24	2%	15	1%	449	4 to 6 months
Variable capacity cooling: Variable speed fan drives	79	7%	49	4%	498	4 to 10 months
Supplemental cooling	200	18%	72	6%	570	10 to 12 months
Monitoring & optimization: Cooling units synchronized	25	2%	15	1%	585	3 to 6 months

Current Situation

At present the data centres are a great matter of concern and there is a global understanding about the future of the data centres. Although data centres are developing at a very fast pace but still the efforts are going on to standardise the things and reduce the environmental effects of the data centres.

Some of the big IT companies like Google, yahoo, IBM etc. have gone a far ahead to achieve green data centres. Some companies like HCL, IBM etc are providing packages to help the organizations develop green data centres in a cost effective manner. The larger player of the IT industry all well aware of the effects of the data centres on the environment and they are trying their best but still we need to increase the awareness into small organizations to achieve Green.

Conclusion

To maintain the environment in which we live is our responsibility. With the advance of technology we are posing new challenges to the environment but now we have to think about the environment friendly technologies. This paper is in accordance with the responsibility to think about nature while developing new technologies. The data centres are posing new environmental threats these days by producing heat and consuming a vast amount of electricity. Therefore we have to rethink about the uncontrolled increase in the number of data centres. In paper we have analysed and discuss about the suggested green database as the solution to the situation. We further suggested some of the solution and steps to build green database to make the databases and data centres environment friendly.

References

[1] Michael Bullock. (2009, August) http://www.cio.com. [Online]. http://www.cio.com/article/499671/Data_Center_Definition_and_Solutions

[2] Rich Miller. (2010, April) http://www.datacenterknowledge.com. [Online]. http://www.datacenterknowledge.com/special-report-the-worlds-largest-data-centers/largest-data-centers-ngd-terremark-qts/#napota

[3] (2011, september) http://www.datacentres.com. [Online]. http://www.datacentres.com/

[4] (2011, September) http://gearenergysys.com/. [Online]. http://gearenergysys.com/

[5] Rick Baur. (2011, September) http://net.educause.edu. [Online]. http://net.educause.edu/ir/library/pdf/bauer.pdf

A Critical Review on Concept of Green Databases

Krishan Bansal, Himanshu Goel and Dr. V.K. Shrivastava

Department of Computer Science and Engineering, APIIT SD India
Panipat- 132102, Haryana, India
E-mail: krishan.bansal22@gmail.com, Himanshu3335@gmail.com, virendra@apiit.edu.in

Abstract
Energy Consumption in data centre is becoming one of the most important issues in today's world. The hardware designers are making efforts to produce the hardware that minimize the power Consumption. But despite of best efforts hardware still consumes more energy even at idle state. Data Centers impact the environment in two ways: (1) the direct generation of heat used to operate IT and cooling equipment. (2) The on-going process of power generation, supply and consumption. Data Centers are configured to work at maximum capacity but the average workload is very less compared to their configuration. Green House Gases (GHG) emissions are becoming major issues in the Information and Communication Society (ICS). Renewable energy sources (e.g. solar, wind, tide, etc.) are emerging as promising solution both to achieve drastically reduction in GHG emissions and to cope with the growing power requirements of data centers. Database management systems offer a great opportunity to reduce energy consumption DBMS have been designed to prioritize performance but are not focused on the power-related issues till now if they are focused on this they can help a lot to a "greener" DBMS.

Keywords: GHG (Green House Gases), DBMS (Database Management System), CO_2 (Carbon Dioxide) SLA (Service Level Agreement), TPC (Transaction process performance council)

Introduction

The Original objective of computing system design has been optimizing the execution time. No concern was made for the energy consumption of the system. Now the power management has become a critical issue in the system design due large electricity bill and the government all over the world has started imposing the taxes on the CO_2 emission and the main objective is now to reduce the emission and help to make the computing system green i.e. emit less CO_2 and GHG. In typical Data Centre electricity consumed by the Servers and cooling systems (needed to remove the heat generated by servers). The contribution of energy in total ownership cost is very high and is always increasing. While the above costs are calculated directly from energy consumption, power (i.e., energy consumption per unit time) savings are of more practical importance than energy savings in system design. Power consumption should be controlled to avoid system failures caused due to overload or overheating due to high server density. Therefore, considering power consumption reduction in the database design so that the total power consumption of an entire system can be kept below a given power budget is important.

Fig 1 http://www.scribd.com/doc/46601972/Green-Databases

In the past few years, the research community has shifted much interest to power-aware applications [5, 6, 8].

In this paper, we will try to find the power consumption patterns and identify power-saving opportunities in database management systems (DBMS). Our main focus should be on characterizing the DBMS regarding energy consumption keeping the SLA and exploit the power consumption of the CPU. Memory should also be used in the energy efficient way. Our ultimate goal is to build power-aware DBMSs or Green DBMS (GDBMSs). Following Observation can be used to realize the potential of DBMSs.

First, Power reduction in DBMS is economical. DBMS is the important part in the software development in three tier architecture mostly used by computing business people. In data center the maximum resources are deployed to the database servers, making DBMS the largest consumer of power in all software application deployed on the server. The processing capacity of the database servers decides the speed of result in the front-end application. Most database servers are configured to work at the maximum load. Thus, this implies great opportunities for power saving.

Fig2 (Source EEDC SEMINAR 2011 Barcelona UPC) http://www.jorditorres.org/news/wp-content/uploads/2011/05/GreenDatabases.pdf

Second, DBMS has the features that offer a great opportunity to reduce power consumption. Specifically a DBMS 1) they are designed to maximize the performance and throughput i.e. they maintain statistics about database states for the purpose of efficient query processing. This information can be used to derive power profile of the database workload and help power- related decision making; 2) query optimizer can help in the number of execution plans for the same query i.e. combination of performance and power saving; 3) can work like OS by requesting the resources (e.g. Memory Buffer, disk space) when needed. This provides opportunities to optimize resource management towards high power-efficiency inside the DBMS.

Motivation

DBMS can be complimentary to the OS- level power management solutions and thus yield greater power savings. Our primary focus is on the Power aware or Green DBMS (PDBMS). We are looking for answers for questions like:

1. Can we really save the power with PDBMS?
2. Memory is also a high energy consuming device. A) Memory should be used in a better way in terms of energy efficiency. B) Considerations should be made on memory consumed in each operation and the energy consumed?
3. Does reducing the amount of memory assigned to an operation reduce the overall energy consumption?

In most case the queries need not be executed immediately thus queries could be delayed and avoid Peak on workloads, queries can be delayed according to Service Level Agreement. The power consumption in databases has started drawing attention from the research community. The Claremont report on database research [4] explicitly states the importance of "designing power-aware DBMSs or Green DBMSs that limit energy costs without sacrificing scalability..."

We are focused on the power consumption patterns of individual queries according to TPC benchmarks by both modeling-based and experimental approaches. TPC Energy: It is a new TPC specification [3] which augments the existing TPC Benchmarks with Energy Metrics developed by the TPC. According to TPC Energy Less energy also have reduced cooling requirements. Redesign the query optimizer to take the power cost of query plans into consideration. [2]. We assume that there exist the power saving capability in the current DBMSs, we can take this as a opportunity by designing a query optimizer that focus on both power consumption and performance. This assumption is made on the fact that the current DBMSs are focused on the performance i.e. the query should be executed fastly with no concern over the energy consumption. And also the argument supports that query processing time is a measure of some "load" to be finished by the DBMS, and more "load" the system faces, more energy will be consumed.

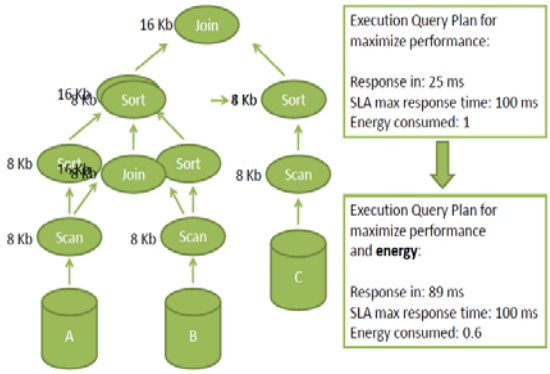


Fig 3 (Source EEDC SEMINAR 2011 Barcelona UPC) http://www.jorditorres.org/news/wp-content/uploads/2011/05/GreenDatabases.pdf

The concept of green database is emerging high and gaining more and more popularity and everyone wants to be on top of it and they have started researches on the concept of energy saving in the DBMSs. The companies like IBM; Oracle that develops their own DBMSs wants to decrease the energy consumption. And they want to on top of TPC-Energy. Hardware providers especially Ram providers are also focusing on the opportunities to develop new energy-aware or green Hardware. Cloud Computing is also emerging as one of the best alternatives for the power saving DBMSs and it offers a huge potential for saving the power consumption.

Conclusion and Future Opportunities

The Green or the Power-aware DBMSs offers a great potential. It is still in the baby phase and lot of work has to be

done in this field. This will not only help companies to reduce their power consumption needs and correspondingly reduce in their power bill hence make running data centers economical, it will also help them to be helpful to the environment by emitting less green houses gases (GHG) and making the environment cleaner. Government is also very much concerned with the amount of carbon being emitted by the companies and on some companies the government has imposed carbon tax which is emitting the large amount of carbon. As a result of this the big companies like IBM, Oracle are concerned for the green database i.e. less power consumption. The organizations like TPC are emerging that are interested in this field of computing. ICT should make every aspect of their business model as energy-efficient as possible. There is lot of room for exploring the future opportunities in the databases to make them greener in terms of energy saving. The research community has started the work on this to explore the potential of PDBMSs. The coordination among the processing capability and execution time can also be checked. How much optimization of energy consumption can be achieved and many more areas the researchers are working to make the database more energy efficient.

References

[1] http://www.jorditorres.org/news/wp-content/uploads/2011/05/GreenDatabases.pdf

[2] Zichen Xu, Yi-Cheng Tu, and Xiaorui Wang. Exploring Power-Performance Tradeoffs in Database Systems. In ICDE, 2010.

[3] R. Agrawal, A. Ailamaki, P. A. Bernstein, E. A. Brewer, M. J. Carey, S. Chaudhuri, A. Doan, D. Florescu, M. J. Franklin, H. Garcia-Molina, J. Gehrke, L. Gruenwald, L. M. Haas, A. Y. Halevy, J. M. Hellerstein, Y. E. Ioannidis, H. F. Korth, D. Kossmann, S. Madden, R. Magoulas, B. C. Ooi, T. O'Reilly, R. Ramakrishnan, S. Sarawagi, M. Stonebraker, A. S. Szalay, and G.Weikum. The claremont report on database research. Commun. ACM, 52(6):56–65, 2009.

[4] [TPCEnergy] TPC-Energy. Transaction Processing Performance Council. http://www.tpc.org/tpc_energy

[5] Y. Chen, T. Wang, J. M. Hellserstein, and R. H. Katz. Energy Efficiency of Map Reduce, http://www.eecs. berkeley.edu/Research/Projects/Data/105613.html, 2008.

[6] C. Xian, Y.-H. Lu, and Z. Li. A Programming Environment with Runtime Energy Characterization for Energy-Aware Applications. In Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED), pages 141–146, 2007.

[7] IDC. Solutions for Data Centers' Thermal Challenges, http://www.blade.org/docs/wp/idc cool blue whitepaper.pdf, January 2007.

[8] P. Bohrer, E. Elnozahy, T. Keller, M. Kistler, C. Lefurgy, C. McDowell, and R. Rajamony. The case for power management in web servers. In R. Graybill and R. Melhem, editors, Power-Aware Computing, pages 261–287. Klewer Academic/Plenum Publishers, 2002.

[9] http://www.scribd.com/doc/46601972/Green-Databases

Ranking of Software Reliability Growth Models using Greedy Approach

Neha Miglani[1] and Poonam Rana[2]

[1]M Tech. Research Scholar, [2]Senior Lecturer
Department of Computer Science & Engineering
Ambala College of Engineering & Applied Research, Devsthali (near Mithapur), Ambala, India
E-mail- neha.miglani27@gmail.com, rana.poonam1@gmail.com

Abstract
A large number of software reliability growth models (SRGMs) have been proposed during the past thirty years to estimate software reliability measures such as the number of residual faults, software failure rate, and software reliability. Selection of optimal SRGM for use in a particular case has been an area of interest for researchers in the field of software reliability. Tools and techniques for software reliability model selection found in the literature cannot provide high level of confidence as they use a limited number of model selection criteria. There is therefore a need for evolving more efficient techniques. An effort has been made in this paper to review some of the well known techniques of this area and the possibility of developing a more efficient technique.

Introduction
In recent years software systems such as operating systems, control programs, and application programs have become more complex and larger than ever. It is quite natural to produce reliable software systems efficiently since the breakdown of the computer system, which is caused by software errors, results in a tremendous loss and damage for social life. Then, software reliability is one of the key issues in modern software product development. Although advances have been made towards the production of defect free software, any software required to operate reliably must still undergo extensive testing and debugging. This can be a costly and time consuming process, and managers require accurate information about how much software reliability is achieved as a result of a particular process in order to effectively manage their budgets and projects. A process, by which it is hoped that software can be made more reliable may be modeled using Software Reliability Growth Models(for e.g., Generalized Goel Model, Goel-Okumoto Model, Gomperts Model, Inflection S-Shaped Model, Logistic Growth Model and so on).

Applying the SRGM's to the observed software error data, the important software reliability measures, such as the number of errors remaining in the system and the software reliability function, can be estimated. These models enable software reliability practitioners to make predictions about the expected future reliability of software under development. Such techniques allow managers to accurately allocate time, money and human resources to a project, and assess when a piece of software has reached a point where it can be released with some level of confidence in its reliability [3]. An error made by a human being and results in a fault in the project. The manifestation of a fault, which means departure from what the software is supposed to do, is referred to as a failure (IEEE standard 782[9]). There is difference between reliability and fault content. A product may have a number of Faults, but these may be locked in paths that are seldom executed; then this product is considered to be reliable. Faults considered by reliability models are those that effect reliability under prevalent conditions, and not necessarily the total faults contents of the software.

Techniques and tools are needed for keeping track of the fault content and the reliability, as long as fault free software cannot be guaranteed. The customer, who buys the software system, need to know if the product fulfils the quality constraints put on it. The tools available for this are mainly software reliability models.

Literature Review
Today the number of existing models exceeds hundred with more models developed every year. Still there does not exist any model that can be applied in all cases. Models that are good in general are not always the best choice for a particular data set, and it is not possible to know in advance what model should be used in any particular case [6]. Over the past thirty years, many SRGMs have been proposed for estimation of reliability growth of products during software development process [1], [5], [7], [8], [10].

Many researchers like Musa et al. [4] have shown that some families of models have, in general, certain characteristics that are considered better than others.

Goel[2] and others[11],[12] started describing processes for which each model would be tested to see how well the model fits the data and predicts the future events. The assertion was that different models predict well only on certain data sets.

The power of several of these statistical tests has been evaluated for a variety of reliability models including those based on a non homogeneous Poisson process, and the Moranda model. Power of these tests has also been compared later.

Proposed Approach

In the present study we are considering the effectiveness of greedy search approach in ranking reliability models. A **greedy algorithm** is any algorithm that follows the problem solving heuristic of making the locally optimal choice at each stage with the hope of finding the global optimum.

In general, greedy algorithms have five components:

1. A *candidate set*, from which a solution is created
2. A *selection function*, which chooses the best candidate to be added to the solution
3. A *feasibility function*, that is used to determine if a candidate can be used to contribute to a solution
4. An *objective function*, which assigns a value to a solution, or a partial solution, and
5. A *solution function*, which will indicate when we have discovered a complete solution

Greedy algorithms are characterized as being *'short sighted'*, and as *'non-recoverable'*. They are ideal for problems which have *'optimal substructure'*. Despite this, greedy algorithms are best suited for simple problems.

Designing of greedy algorithm is based on finding out the shortest path by using suitable algorithms and calculating its weight or distance from the origin, naming it as OPTIMUM and then comparing the values of different models with the optimum value.

Assume x is a candidate model and its objective value is Alt(x). Our aim is to find x which minimizes value of Alt(x).

In Mathematical terms, it can be represented as:

Minimize s |OPT-Alt(x)|

Subject to x.

Where Alt(x) represents SRGM alternative and s represent distance from optimum value 'OPT', OPT is desired optimal value.

Figure.1 describes how ranking of the models would occur by calculating the distances of different models from the optimum value OPT. All the models would lie in a feasible region named ACTIVE. By calculating the distance of these models from OPT, we will attain the objective values for different models and hence, models can be ranked on this basis.

The flowchart of the proposed technique is shown in figure2.

Example

We present here for illustrations Sharma et.al [3] .It targeted testing the suitability of the developed DBA method so that a comprehensive ranking of the alternative SRGMs could be made combining various attributes relevant to SRGMs for a data set. The paper included NHPP SRGMs namely,

- Generalized Goel Model
- Goel-Okumoto Model
- Gomperts Model
- Inflection S-Shaped Model
- Logistic Growth Model
- Modified Duane Model
- Musa-Okumoto Model
- Yamada imperfect debugging Model
- Yamada Rayleigh Model
- Delayed S-Shaped
- Yamada imperfect debugging Model2
- Yamada exponential Model
- P-N-Z Model
- P-Z Model
- Pham Zhang IFD Model
- Zhang-Teng-Pham Model

A dataset was taken from the open literature for evaluation, optimal selection and ranking of the NHPP SRGMs based on criteria named Bias, MSE, and MAE and so on. The dataset was collected from a subset of products for four separate software releases at Tandem Computers Company as shown in Table1.

The value of the comparison criteria are calculated using Least Square Estimation. Then, estimated and optimal values are used to compare the rankings of all the models based on values of comparison criteria.

Work in hand

Currently we are trying to rework ranking of these models using greedy approach and compare the results with those of Sharma et.al [3] and others. The work is still in progress and we hope to present the results in the conference.

Figures and Tables

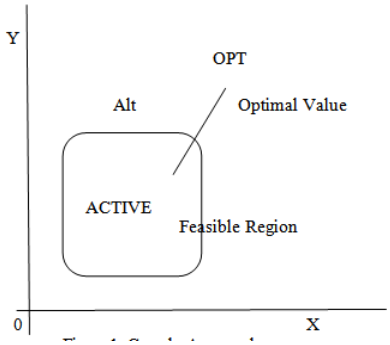


Figure 1. Greedy Approach

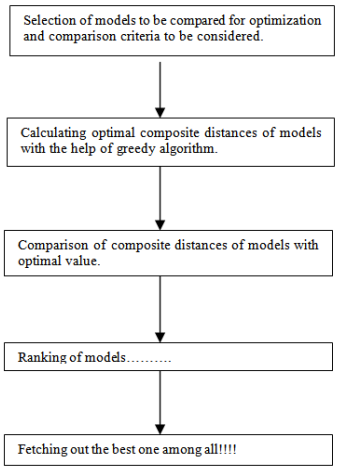


Figure 2. Flowchart representing proposed technique.

Table1: Tandems Computers Software Failure [3].

Weeks	CPU hours	Defects found	Weeks	CPU hours	Defect found
1	519	16	11	6539	81
2	968	24	12	7083	86
3	1430	27	13	7487	90
4	1893	33	14	7846	93
5	2490	41	15	8205	96
6	3058	49	16	8564	98
7	3625	54	17	8923	99
8	4422	58	18	9282	100
9	5218	69	19	9641	100
10	5823	75	20	10000	100

Conclusions

The objective of this work is to develop a ranking technique which is based on greedy approach to rank different types of software reliability models and compare its performance with the currently available algorithms.

We also intend to apply this approach to some specific case studies of softwares.

References

[1] M. Xie, Software Reliability Modeling, World Scientific Publishing Co. Ltd., 1991.

[2] L.Goel,"Software Reliability Models: assumptions, limitations, and applicability." IEEE Trans. On Softw. Engineering, December 1985, pp.1411-1423.

[3] Kapil Sharma, Rakesh Garg, C. K. Nagpal, and R. K. Garg, "Selection of Optimal Software Reliability Growth Model using Distance Based Approach" TR2009-063.

[4] J. D. Musa, and K. Okumoto, Software Reliability Measurement, Prediction, Application, McGraw Hill, 1987.

[5] M.R. Lyu, Handbook of Software Reliability Engineering, McGraw-Hill, 1996.

[6] A. D. Denton, "Accurate Software Reliability Estimation," Master of Science Thesis, Colorado State University, Fort Collins, Colorado, Fall 1999.

[7] Q. P. Hu,M. Xie,S.H. Ng,and G. Levitin,"Robust recurrent neutral network modeling for software fault detection and correction prediction,"Reliabilty Engineering and System Safety,vol 92 no.3,2007,pp. 332-340.

[8] D. R. Jeske, and X. Zhang,"Some successful approaches to software reliability modeling in industry,"J. Syst.Softw., vol. 74, no. 1, 2005, pp.85-99.

[9] Measures for Reliable Software, IEEE Standard 782, 1986.

[10] C. Y. Huang, and C. T. Lin, "Software reliability analysis by considering fault dependency and debugging time lag," *IEEE Trans. Reliability*, vol.55, no. 3, 2006, pp. 436–450.

[11] J. D. Musa, "A theory of software reliability and its application," IEEE Trans. Software Eng., vol. SE-1, pp. 312-327, Sept. 1975.

[12] S. Yamada, M. Ohba, and S. Osaki, "S-shaped reliability growth modeling for software error detection," IEEE Trans. Rel., vol. R-32, pp. 475-478, 484, Dec. 1983.

Optimum Software Reliability: A Literature Review

Gunjan Sethi[1] and Poonam Rana[2]

[1]M.Tech. Research Scholar and [2]Senior Lecturer, Department of Computer Science and Engineering, Ambala College of Engineering and Applied Research, Devsthali, Ambala, Haryana, India
E-mail: gunjansethi87@gmail.com, rana.poonam1@gmail.com

Abstract
A number of analytical models have been proposed during the past 15 years for assessing the reliability of a software system. In This Paper we proposed a model to get the optimum software reliability & optimum cost subject to time .Now a days industry needs a software reliability product with optimum cost.

Keywords: Software reliability, Two dimensional software reliability growth model, Goodness of fit, optimal release problem.

Introduction

A metric which reflects the degree of program correctness and which can be used in planning and controlling additional resources needed for enhancing software quality .One such quantifiable metric of quality that is commonly used in software engineering practice is software reliability. A commonly used approach for measuring software reliability is via an analytical model whose parameters are generally estimated from available data on software failures. Reliability and other relevant measures are then computed from the fitted model. Software reliability is a probabilistic measure and can be defined as the probability that software faults do not cause a failure during a specified exposure period in a specified use environment [1].Assessing software reliability in a testing phase of a software development process is one of the important issue to develop a highly reliable software system. Software reliability models are used for the prediction and estimation of software reliability [2].A software reliability growth model is known as one of the fundamental technologies for quantitative software reliability assessment and playing an important role in software project management for producing a highly reliable software system [3]. Nonhomogeneous process (abbreviated as NHPP) model is to determine an appropriate mean value function to denote the expected no of failures experienced up to a certain time. NHPP can be classified by the fault –detection time distribution [4].Software developing managers have a great interest in how to develop a reliable software product economically and when to release the software to the customers.

Literature Review

Software reliability models are used for the prediction and estimation of software reliability [2]. Jintao Zeng Jinzhong Li, Xiahoui Zeng, Wenlang Luo. In A Prototype System of Software Reliability Prediction and Estimation have proposed an approach for software reliability model selection based on experiences from history software projects [2]. In this paper it is considered that experiences from model selection of history projects which can be used to serve model selection of the current project. By using this approach two questions be solved how to measure the similarity of software projects and how to make full use of history project experiences. Shinji Inoue and Shigeru Yamada in Two-Dimensional Technologies describes a software reliability growth process depending on two-types of software reliability growth factor : Testing –time and Testing–effort factors[3].Two-dimensional software reliability measurement technologies enable us to conduct more feasible software reliability assessment than the one-dimensional (conventional) software reliability measurement approach . In this approach software reliability growth process depends only on testing-time. Two-dimensional software reliability growth modeling approaches for feasible software reliability assessment, and conduct goodness-of –fitness comparisons of our models with one-dimensional software reliability growth models. Shinji Inoue in Generalized Discrete Software Reliability Modeling With Effect of Program Size have proposed an approach for cost-optimal and cost-reliability-optimal software release policies[4],[5],[6].Cost-optimal software release policies based on generalized discrete binomial process model. Optimal software release problems which take both total software cost and reliability criteria into consideration simultaneously.

Proposed Approach

Optimization of software reliability is considered as objective with respect to cost as constraint. The expected software cost E (T) and the software reliability R(x/t), we have to determine the optimum release time that minimizes the expected software cost subject to attaining a desired reliability level.

Minimize E (T)
Subject to R (x|T)$\geq$R0

The Software reliability R(x/T) and the expected software cost E (T).The Optimization problem can be formulated as

{Maximize R (x|T)
{Subject to E (T) < C

E (T) can be calculated as

$E(T) = {}_0\int^{T} [C_3 t + \sum^{3}_{i=1} C_{i1} m_i(t)] g(t)dt + {}_T\int^{\infty} [C_3 T + \sum^{3}_{i=1} C_{i1} m_i(T) + \sum^{3}_{i=1} C_{i2}(m_i(t) - m_i(T))] g(t)dt$

The Function E(T) represents testing costs per unit time and of fixing errors during testing incurred if the determination of the software life cycle is less than or equal to the software release time. On the other hand if the determination of the software lifecycle is greater than the software release time, then an additional cost factor should be involved i.e. the cost of fixing errors during the operation phase.

Mathematical Model

Let E (T) be the expected software cost.

R (X|T) is the software reliability.

Optimize the software reliability taking cost as a constraint.

So, the problem is formulated as:

Optimize (E (T),

Subject to R (X|T)

Example

To optimize the software reliability taking cost as a constraint let us assume that C_3, C_R, C_{i1} and C_{i2} for $i=1,2,3$ be the cost and the optimal value of T say $T_{rel,}$ that maximizes R(X|T) subject to the cost constraint is determined from

If $E(T^*) > C_R$ Then there is No solution

Else $T_{rel} = \{T \geq T^*: T = E^{-1}(C_R)\}$; End if.

This shows that if E (T*)> CR, the minimum software system cost required to develop and debug the program exceeds the maximum amount allowable. Therefore, it is impossible to produce the software under these conditions. Similarly , if E(T*) ≤ CR, and as the reliability of software continually improves with testing and debugging time, then the program should be debugged until the cost constraint is binding, implying that additional debugging will violate the constraint.

The software reliability function can be formulated as

$R(x|t) = e^{-[\Sigma^{3}_{i=1}(ap_{i/(1-\beta i)})}(e^{-(1-\beta_i)b_i t)}[1- e^{-(1-\beta_i)b_i x}]$

In this equation:

An expected number of software errors to be eventually detected

b_i error detection rate per type i error i = 1, 2, 3; $0 < b_1 < b_2 < b_3 < 1$

p_i content proportion of type i errors

β_i type i error introduction rate that satisfies, $0 \leq \beta_i < 1$

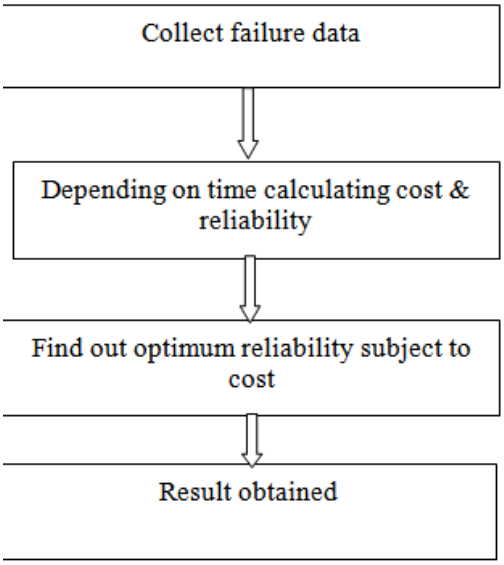


Figure1. Flowchart representing proposed technique.

Table1. Failure Data

Test		Failures		
Week	Hours	Critical	Major	Minor
1	62.5	0	6	9
2	44	0	2	4
3	40	0	1	7
4	68	1	1	6
5	62	0	3	5
6	66	0	1	3
7	73	0	2	2
8	73.5	0	3	5
9	92	0	2	4
10	71.4	0	0	2
11	64.5	0	3	4
12	64.7	0	1	7
13	36	0	3	0
14	54	0	0	5
15	39.5	0	2	3
16	68	0	5	3
17	61	0	5	3
18	62.6	0	2	4
19	98.7	0	2	10
20	25	0	2	3

Conclusions

In today's life we are dependent on some of the products that are highly critical in nature so such kind of products need a high level of reliability. To achieve high level of reliability a high cost is required which is not always possible. So, the objective of the work is to maximize the reliability with respect to software cost using heuristic based approach such as genetic algorithm.

References

[1] Amrit L. Goel "Software Reliability Models: Assumptions, Limitations and Applicability," IEEE Trans. on Software Engineering, Vol, SE-11, No. 12 December 1985.

[2] Jintao zeng, Jinzhong Li, Xiaohui Zeng, Wenlang Luo "A Prototype System of Software Reliability Prediction and Estimation". IITSI 2010

[3] Shinji Inoue and Shigeru Yamada "Two –Dimensional Software Reliability Measurement Technologies'' IEEE 2009.

[4] Shinji Inoue "Generalized Discrete Software Reliability Modeling with Effect of Program Size" IEEE Trans on Systems, man, and cybernetics-Part A: Systems and humans, Vol, 37, No. 2, March 2007.

[5] M. Xie, Software Reliability Modeling, World Scientific Publishing Co. Ltd., 1991.

[6] D. R. Jeske, and X. Zhang,"Some successful approaches to software reliability modeling in industry,''J. Syst.Softw., vol. 74, no. 1, 2005, pp.85-99.

[7] H.Pham, Software Reliability. Singapore: Springer-Verlag, 2000.

[8] Ying Zhou, Joseph Davis, "Open source software reliability model: an empirical approach,'' Proceedings of the 5th WOSSE, pp. 1-6, 2005

[9] Yoshinobu Tamura, Shigeru Yamada, "Comparison of Software reliability Assessment Methods for open source software,'' Proceedings. 11th ICPDS 2005 Vol 2, 20-22 july 2005 pp:488-492 Vol. 2

[10] C. Smidts, R. W. Stoddard , M. Stutzke, "Software reliability models : an approach to early reliability prediction, '' ISSRE, p. 132, October 30-November 02, 1996

AcceptSoftware: A Tool for Executable Acceptance Test Driven Development

[1]Durgesh Samadhiya and [2]Ashish Ranjan

[1]Advanced Institute of Business Management, Faridabad, India
E-mail: samadhiya.durgesh@gmail.com
[2]Capital Business School, Delhi, India
E-mail: ashish23@gmail.com

Abstract
This paper introduces AcceptSoftware which is a tool to easily create and run client readable acceptance tests, and describes how it can be used to allow a simple but powerful acceptance-test driven software development. We then describe our AcceptSoftware tool that extends EasyAccept by maintaining a history of acceptance test results. Based on the history, AcceptSoftware is able to generate reports that show when an acceptance test is suddenly failing again.

Keywords: software testing, acceptance test, ATDD, test-driven development

Introduction

Acceptance testing is an important aspect of software development. Acceptance tests (sometimes referred as story tests in agile teams) are high level tests of business operations. They are not meant to test internals or technical elements of the code, but rather are used to ensure that software meets business goals. Executable (i.e. automated) acceptance tests can be used as a measure of project progress.

As the software system becomes more complex, analysts spend more time on requirements specifications. A solution is to repeat the development cycle in small incremental iterations, as recommended by agile methods [5]. One of the biggest contributions of agile methodologies is the concept of test-driven development (TDD). In TDD, the tests are written before writing the actual code. The tests can be used to evaluate the development progress by measuring the number of passing or failing tests and to perform continuous regression testing, which can help maintain high software quality by notifying the developers of software defects as soon as the code is changed.

Automated acceptance tests [6] are used in TDD which is called Executable Acceptance Test Driven Development (EATDD). It is also known as Story Test Driven Development or Customer Test Driven Development. Acceptance tests for a feature should be written first by the customer with the help of the development team, before the application code is implemented. The tests represent system requirements and specifications. Then, the development team will work on implementation with guidance of the acceptance tests. The implementation is completed when all the corresponding acceptance tests are passed.

While TDD focuses on unit tests to ensure the system is performing correctly from a developer's perspective, EATDD starts from business-facing tests to help developers better understand the requirements, to ensure that the system meets those requirements, and to express development progress in a language that is understandable to the customers [2].

There is often a substantial delay between defining an acceptance test and its first successful pass [3]. Therefore, it becomes important for teams to easily be able to distinguish between tasks that were never tackled before and tasks that were already completed but whose tests are now failing again. This is achieved by using AcceptSoftware.

This paper introduces AcceptSoftware which is a tool to easily create and run client readable acceptance tests, and describes how it can be used to allow a simple but powerful acceptance-test driven software development. We then describe our AcceptSoftware tool that extends EasyAccept [8,11] by maintaining a history of acceptance test results. Based on the history, AcceptSoftware is able to generate reports that show when an acceptance test is suddenly failing again.

The rest of this paper is organized as follows. We review the related work in Section II. We then review EasyAccept and present our motivation to improve it in Section III, and also we introduce AcceptSoftware. We discuss its implementation in Section IV. Finally, Section V concludes the paper.

Related Work

In this section, we review existing researches and tools related to EATDD. We divide them into three categories as the following sub-sections.

Table-based frameworks

There are several open-source frameworks and tools that support EATDD. Table-driven tests are best suited to express business rule examples in input-output pairs that can be linked to the business logic algorithmically. On the other hand, sequential command-driven tests are suited to express the business logic workflow. It is well suited to testing from a business perspective, using tables to represent tests and automatically reporting the results of those tests.

Examples of tools in this category include Fit [4], FitNesse [7], and Selenium [9]. The most widely known tool

for acceptance testing is FiT (Framework for Integrated Testing). FiT requires developers to design individual fixture classes with hookup code for every type of table used in the tests and cope with data being referenced across tables.

Text-based frameworks

Although table-based frameworks might be the mainstream right now, they are not the only class of frameworks suitable for acceptance testing. Not everyone likes authoring tests as tables. The text written into the cells of test tables is often close to written English, but the table structure brings with it a degree of syntax.

Text-based tests are written as simple texts using a text editor. These kinds of tests are useful to represent work flows [10]. Examples of tools in this category include Exactor [12], TextTest [13], EasyAccept, jaccept [14]. Exactor uses textual scripts, JAccept is based on a graphical editor and XML test files, and TextTest tests programs with command-line textual input and output. They are suited to express the business logic workflow.

Scripting language-based frameworks

There is another category of acceptance-testing tools that can offer a great deal of power through flexibility and friendliness of a scripting language. A good example of this category of tools is Systir [15] which makes use of the Ruby scripting language's syntax for building reasonably-good custom domain-specific languages.

Easyaccept and Motivation to Improve

AcceptSoftware tool extends EasyAccept by maintaining a history of acceptance test results. EasyAccept is an open-source tool that can be found at [11]. It takes acceptance tests enclosing business rules and a Façade to access the software under development, and checks if the outputs of the software's execution match expected results from the tests. Driven by EasyAccept runs, software can be constructed with focus, control and correctness, since the acceptance tests also serve as automated regression tests.

In short, EasyAccept is a script interpreter and runner. It takes tests enclosed in one or more text files and a Façade to the program that will be tested. Accessing the program through Façade methods that match user-created script commands, EasyAccept runs the entire suite of tests and evaluates actual and expected outputs or behaviors of the program under test. In a test report, the tool shows divergences between actual and expected results, or a single message indicating all tests were run correctly.

The acceptance tests are written in text files with user-created commands close to natural language. EasyAccept provides some built-in commands which are combined with such customized user-created commands specific for each application to create the tests.

The overhead of getting started with EasyAccept is practically zero, and it requires minimal additional work on the part of the developers. They only need to provide a Façade to the program to be tested containing methods whose signatures match the user-created commands. A single Façade that exposes the program's business logic helps separate business and user interface concerns, and may even already exist in programs not created with an ATDD approach, since this separation is an advocated architectural best practice. Other textual testing tools use various approaches, none of which involves the use of a single Façade.

Motivation to improve

A major difference between UTDD and EATDD is the timeframe between the definition of a test and its first successful pass. In UTDD, the expectation is that all unit tests pass all the time and that it only takes a few minutes between defining a new test and making it pass [1]. As a result, any failed test is considered as a problem that needs to be resolved immediately. Unit tests cover very fine grained details which make this expectation reasonable in a TDD context.

Acceptance tests, on the other hand, cover larger pieces of system functionality. Therefore, we expected that it takes the developers several hours or days, sometimes even more than one iteration, to make them pass. Due to the substantial delay between the definition and the first successful pass of an acceptance test, a development team can not expect that all acceptance tests pass all the time. A failing acceptance test can actually mean the followings.

- Non-implemented Feature: The development team has not yet finished working on the story with the failing acceptance test (including the developer has not even started working on it).
- . Regression Failure: The test has passed in the past and is suddenly failing – i.e. a change to the system has triggered undesired side effects and the team has lost some of the existing functionalities.

Keeping history of number of passed and failed acceptance tests of a project helps the development team understand the development progress. From such statistics, the development team can grasp the speed of their development and where they are in the development process.

AcceptSoftware has the functionality of showing the test result history. Test result history is kept in the database. To show the test result history, a chart showing the test running date and result details are provided.

Changes are often made to acceptance tests. Most people make changes to acceptance tests many times a day when they come up with new ideas. Acceptance tests which were changed before might need to be reversed back to a previous version. However, only keeping the version information is not sufficient enough. Sometimes the developers or tests make improper changes and keep adding changes to the tests for a period of time. Afterwards, when people discover the mistake, provided only a version number and a date, it is very hard for them to decide which version of the test is useful. It will be very helpful if the test result information can be kept with the corresponding versions of the test. By viewing the test results, people can easily identify the test that is performing as expected. AcceptSoftware achieves this goal by keeping test result record after each test run. In addition, identifying the regression failure of acceptance tests requires keeping history of the tests to identify the last version of successful tests.

Acceptance tests can be divided into the following categories.

Tests containing lots of information and formulas. It is efficient to represent such tests using tables.

Tests containing job rules. It is efficient to represent such tests using texts.

None of the existing EATDD tools supports both the above categories of tests. In addition, none of the existing EATDD tools keeps history of tests. We developed AcceptSoftware that adds these two features into EasyAccept.

The AcceptSoftware Tool

Fig. 1 demonstrates the test framework which is used in AcceptSoftware. A class called Façade is used to call procedures of the under-test program. All commands in test scripts must be compatible to Façade's methods. Façade helps the developer in the future when the developer implements a user interface.

AcceptSoftware contains the same internal commands used in EasyAccept except for an internal command called expectTable. We have extended this command in AcceptSoftware providing the possibility in AcceptSoftware to read a data table from a database (including Oracle, Access, MySQL, and SQL-Server databases) and then use the data to test the program.

Implementation of Acceptsoftware

In this section, we describe our implementation of AcceptSoftware.

Class AcceptSoftware

This class is the core of AcceptSoftware that manages operations such as detecting the Façade of the under-test program, detecting test script files, and doing test operations for each script file.

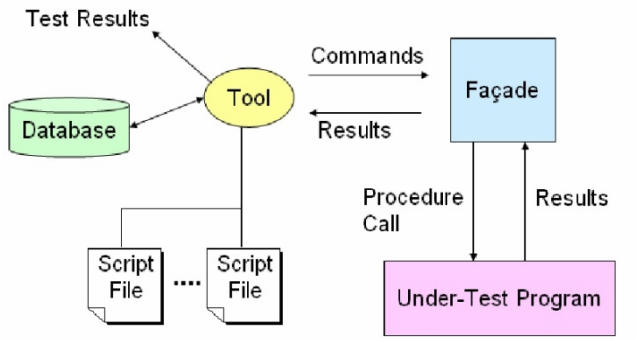


Figure 1. AcceptSoftware's Framework.

Class AcceptRunner

This class tests the program using a procedure called runnScript() considering the script file. The test operation is done using a method in the Script class.

Class Script

This class contains a method called runn() that runs the script on the under-test program and reports the result. Another method in this class called execute() helps in execution of the scripts. To properly perform the tests using the script file, this class parses the script file and associatively accesses Façade.

Class ParsedLineReader

Using method getParsedLine() in this class, the script files is parsed line by line and keywords are searched.

Tokens

Tokens are the keywords used to write the scripts. The tokens defined in AcceptSoftware include: echo, expect, expectdifferent, expecterror, expectwithin, equalfiles, quit, stringdeli, iter, stacktrace, executescript, threadpool, repeat, expectable.

Class ExpectTableProcessor

This class searches the filename or the id of the database in the script file. This operation is successful only when the tool reads keyword "expectTable" before the name of the database. Then, it connects to the database and reads the data table. It creates a new script file containing the data and executes this file. In this way, the data stored in a database can be used for testing a program.

Class DatabaseHandler

This class handles detection of database type, connecting to database, reading data from database, and creating the script file from it.

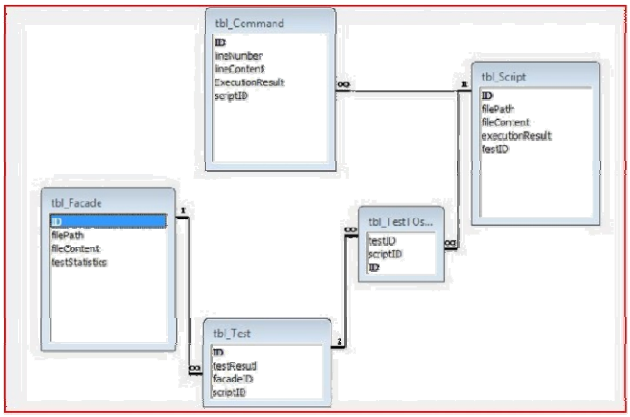


Figure 2. Defined tables and their relationship

Database Implementation to keep test history

One of the advantages of AcceptSoftware over EasyAccept is the capability of storing data and statistics related to different executions of the under-test program. To achieve this feature, we implemented a database to log all the events and statistics related to execution of the program.

We define the following five tables (Fig. 2) in the database for keeping test history.

- tbl_Test
- tbl_Script
- tbl_Command
- tbl_Facade
- tbl_TestTOscript

A record is stored in a table called tbl_Test for each under-test façade. A record is stored in a table called tbl_Script for each script file which is tested on the façade. A script file contains a number of command lines. Each command line is stored in a table called tbl_Command as a record.

Applying a command line during testing, the test result is updated in tbl_Command. This process continues until all the command lines are executed. Then, the result of executing the entire script file is updated in tbl_Script. Finally when the façade is tested by script files, the total result of this version of tests is updated in a table called tbl_Test.

A façade may be tested multiple times in the database. In this case, only one record is inserted in tbl_Facade whereas multiple records are inserted in tbl_Test. This feature avoids data redundancy and makes it easier to report a façade.

Conclusions

This paper presents AcceptSoftware which is an EasyAccept-based tool for automated acceptance testing and a self-evaluation of the tool. Existing tools are limited in supporting Acceptance Test Driven Development as they do not provide enough information to distinguish two different kinds of test failures. AcceptSoftware distinguishes these failure states by maintaining a test result history on the server, which is valuable for analyzing the existing progress and making improvements. Table I compares AcceptSoftware with the existing open-source tools of acceptance testing.

Table 1: Comparison of software testing tools.

Tool		TextTest	Exactor	EasyAccept	Selenium	FIT	AcceptSoftware
Acceptance Testing Criteria	Edit/Run	*	*	*	*	*	*
	Supporting the text format	*	*	*	-	-	*
	Supporting the HTML format	-	-	-	*	*	*
	Supporting the Excel format	-	-	-	-	*	*
	SQL, Oracle, XML format	-	-	-	-	-	*
Test Result Criteria	Detection of regression errors and non-implemented features	-	-	-	-	-	*
	Presenting test result history	-	-	-	-	-	*
Other Criteria	Open source	*	*	*	*	*	*
	Being user friendly	*	*	*	*	*	*
	Containing a Façade	-	-	*	-	-	*

As a tool supporting agile methodology, it will be helpful to integrate this work with other practices in Agile. For instance, acceptance tests can be used in conjunction with story card management to provide more meaningful reports for the customers.

The work presented in this paper is a preliminary step in constructing an effective tool for supporting EATDD in Agile software development environment. There is still a lot of room in this research area for future work.

From the self-evaluation, we can see that AcceptSoftware can provide useful support for EATDD. However, this self-evaluation is limited in time and the number of acceptance tests. Therefore, the next research step is to conduct a more formal evaluation of the approach to assess if AcceptSoftware as a whole is useful for development teams to practice Executable Acceptance Test Driven Development.

Another idea for future work involving AcceptSoftware is a comparison to other ATDD approaches, particularly those that use different formats of acceptance tests such as FiT tables. Such a comparison would allow us to abstract away which ATDD patterns and techniques are tool-dependent and which are general, improving the state-of-the-art of acceptance testing.

References

[1] K. Beck, C. Andres, Extreme Programming Explained, 2nd Edition, Addison-Wesley, 2005.

[2] L. Koskela, Test Driven: practical TDD and acceptance TDD for Java developers, Manning Publications, 2007.

[3] L. Crispin, T. House, and C. Wade, “The Need for Speed: Automating Acceptance Testing in an Extreme Programming Environment”, Proc. Second Int'l Conf. eXtreme Programming and Flexible Processes in Software Eng., pp.96-104, 2001.

[4] FiT, http://agile.csc.ncsu.edu/SEMaterials/tutorials/fit/.

[5] Calgary Agile Method User Group home page: http://www.agilenetwork.ca/camug/, 2007.

[6] A. Neto, J. P. Sauvé, and A. Dantas, “Patterns for Scripted Acceptance Test-Driven Development”, Proceedings of EuroPLoP'07, 2007.

[7] R. Mugridge, and W. Cunningham, Fit for Developing Software: Framework For Integrated Tests, Prentice Hall, 2005.

[8] J. P. Sauvé, A. Neto, W. Cirne, “EasyAccept: a tool to easily create, run and drive development with automated acceptance tests”, Proceedings of the 2006 international workshop on Automation of software test, 2006.

[9] Selenium homepage on OpenQA, website: http://www.openqa.org/selenium/, 2006.

[10] J. Andersson, G. Bache, P. Sutton, “XP with Acceptance-Test Driven Development: A rewrite project for a resource optimization system”, Proceedings of the 4th International Conference on Extreme Programming, 2003.

[11] EasyAccept Homepage: http://easyaccept.org.

[12] Exactor, http://exactor.sourceforge.net/.

[13] TextTest, http://texttest.carmen.se/.

[14] Jaccept, http://maven.agilos.org/sites/jaccept/released/.

[15] Systir, http://systir.rubyforge.org/.

Queuing Algorithms Performance against Buffer Size and Attack Intensities

Santosh Kumar[1], Abhinav Bhandari[2], A.L. Sangal[3] and Krishan Kumar Saluja[4]

[1-3]Computer Science and Engineering, Dr. B. R. Ambedkar NIT, Jalandhar, India
[4]S.B.S.C.E.T Firojpur, India
E-mail: [1]santosh.iet06@gmail.com, [2]bhandarinitj@gmail.com, [3]sangal62@yahoo.com, [4]k.saluja@rediffmail.com

Abstract
Distributed Denial of Service (DDoS) attack is one of the biggest threats now days. This paper aims at providing the simulation results of buffer size and attack intensities effect on various queuing algorithms such as DropTail, Fair Queuing (FQ), Stochastic Fair Queuing (SFQ), Deficit Round Robin (DRR) and Random Early Detection (RED) using ns-2 as a simulation environment. The results in this paper indicate that Stochastic Fair Queuing is the best algorithms in terms of providing maximum bandwidth to legitimate users against various attack intensities. It is also cleared from simulation results that there is no effect of variation in buffer size on queuing algorithms such as Fair Queuing, Stochastic Fair Queuing and Deficit Round Robin while DropTail and Random Early Detection algorithms are giving the best performance on buffer size 60 against various attack intensities. This paper also covers the basic overview of Denial of Service Attack (DoS), Distributed Denial of Service attack (DDoS), attacking methods, DDoS defense approaches and Queuing Algorithms.

Keywords*:* DDoS, Queuing algorithms, Buffer size, Attack intensities

Introduction

Denial of service attack is an attempt to prevent the legitimate users from accessing the network resource such as website, computer system or web service [1]. The aim of DoS attack is to send a vast number of messages to the destination so that it can be crashed, reboot or not be able to full fill the legitimate users' request [2]. Distributed Denial of Service attack is a coordinated Denial of Service attack that uses so many computers to launch an attack against one or many destinations [3]. To launch a coordinated attack DDoS uses many compromised systems to degrade the performance of target. The target of the Distributed Denial of Service attack is called "primary victim" while the compromised systems that are used to launch DDoS attack are often called "secondary victims". Fig. 1 shows the architecture of DDoS attack.

Thousand of attacks occur on regular basis and few of them get caught or traced. There are many types of attackers who participate in DoS attack. Sophisticated attackers are those who hide their identities by several means during the attack. Script kiddies are those who use few kinds of attacking tools available on the internet, such attacker sometimes get caught or easily traced because they left sufficient trails to trace [4]. Another type of attacker is potential attackers who use their own developed scripts or tool for attacking purpose they are smart enough to write their own code.

Two types of attacks occur in the network (i) Vulnerability attack is an attack in which attackers exploits the vulnerability available in the system. (ii) Flood attack is an attack in which attackers send a vast number of messages to overwhelm the network bandwidth in terms of consuming bandwidth.

There are many attacking methods such as smurf attack, ping flood, ping of death, SYN flood, teardrop attack, permanent denial of service attack, degradation of service attack and nuke are used in DoS and DDoS attack.

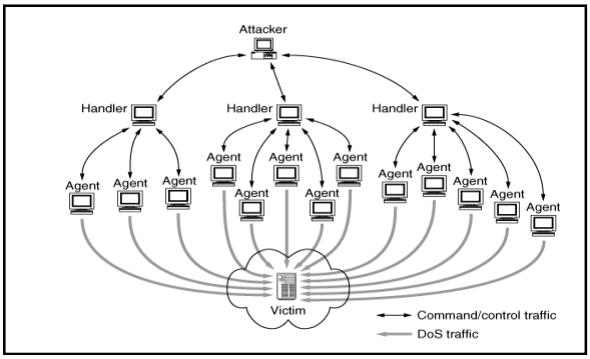


Fig 1: Handler/Agent Architecture [2].

Some attackers are smart enough to create their own attack code, most commonly they use code written by others. Such code is typically built into a general, easily used package called an attack toolkit. It is very common today for attackers to bundle a large number of programs into a single archive file, often with scripts that automate its installation. Now days the attacking tools such as Trinoo, Tribe Flood Network (TFN), Stacheldraht, Shaft and Tribe Flood Network 2000 (TFN2K) are used as DDoS attack tool kits [5-8].

DDOS Defense Approaches

The aim of DDoS defense approach is to improve the security level of a computer system or network. Few of them are explained below.

Disabling unused Services

If there are less application services and open ports in hosts, there will be less chance of exploiting the vulnerabilities by attackers. Therefore the best way to reduce the chances of occurring DDoS attack is to disable the services that are not in use, e.g. UDP echo, character generation services [9].

Install latest security patches

Nowadays attacks are based on exploiting the vulnerability in the target system. So by installing latest security patches we can prevent the re-exploitation of vulnerabilities available in the target system [9].

Disabling IP broadcast

Generally attackers use intermediate broadcasting nodes to consume the network bandwidth. Smurf and ICMP flood attacks are based on broadcasting. So defense against attack will be successful if the host computer and all intermediate nodes disable the IP broadcast [10].

Firewalls

A firewall is a device that is used to protect the network from unauthorized access while permit the legitimate services to pass through it. Fire wall have some policies such as to allow or deny protocols, ports or IP addresses [11]. But some complex attack, such as attack on port 80 (web services) firewalls cannot prevent that attack because they cannot distinguish good traffic from DoS attack traffic.

IP Hopping

In IP hopping, the IP address of active servers is proactively changed within the range of a pre-specified set of IP addresses [12]. The victim computer's IP address is invalidated by changing it with a new one. Once the IP address is change all the routers in the network is informed and edge router will drop the attacking packets. So we can prevent the attack by using the IP hopping. Drawback of this technique is that another system may be victim of attacker if it is allocated the previous IP address of active server.

Ingress/Egress Filtering

Ingress Filtering is a restrictive mechanism to drop any incoming packet if its IP address does not match with a domain prefix connected to the ingress router. Egress filtering ensures that the packet leaving from any network having IP address claims to that network really match within the range of IP addresses of that network [14].

First most important thing is required for Ingress/Egress filtering to have knowledge of expected IP addresses at particular port but for some networks with complicated topologies it is not easy to obtain this knowledge. Reverse path filtering [15] is a technique that is used to build this knowledge. In this technique each router knows which network is accessible via which interface of that router. If coming packet at any particular interface of the router claims that its source IP address belongs to a particular network then we do cross check. Router again tries to find out whether that source address using the particular interface is accessible or not. If yes then packet is allowed to pass through that router otherwise dropped.

Defense against IP spoofing

It is a defense mechanism against IP spoofing based on trusted nodes and traceroute [16]. Consider a network consists of trusted nodes. Each trusted node in a network contains the access information about all other nodes such as node name and IP address, hop count and traceroute from itself to the other trusted nodes. IP spoofing is a process in which hacker sends the request to any destination node while having source address spoofed. In this method, whenever any node send the request to other node in the network for establishing the communication. The node that gets the request from any particular node first verifies that node by using traceroute whether it is trusted node or not. In traceroute method, if any node gets the request then it tries to access that IP address to check whether it is accessible or not. If the node is not accessible then receiver node of request gets message "host is unreachable". In this case the receiver does not respond to that IP address.

Queuing Algorithms

A queuing algorithm allows us to manage access to the fixed amount of out port bandwidth by selecting which packet should be transferred and which one should be dropped when queue limit is fully occupied. There are many different queue scheduling algorithms to provide the balance between complexity, control and fairness. Congestion occurs when packets arrive at out port faster than they can be transmitted. In this case router interface become congested if just a single packet has to wait for another packet to complete its transmission. The task of queue scheduling algorithms is to minimize the congestion and to provide fair bandwidth to each of different services competing for bandwidth on the output port. It also furnishes protection between different services on output port, so that poorly behaved service in one queue can not impact the bandwidth delivered to the other services. In our simulation we are using the DropTail, Fair Queuing (FQ), Stochastic Fair Queuing (SFQ), Deficit Round Robin (DRR) and Random Early Detection (RED) available in ns-2.

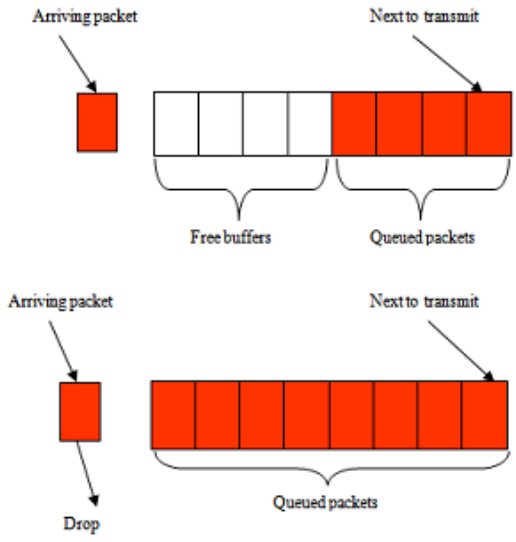


Fig 2: DropTail [17].

DropTail

DropTail is one of the simplest algorithms mostly used in the internet routers. It is based on first in first out (FIFO) queue [17] policy. The entire incoming packets are stored in a buffer or queue of limited size. And router serves the packets stored in queue in the same order as they were placed. Fig. 3 shows the function of DropTail algorithm.

Fair Queuing

Fair Queuing is an algorithm having motive to allocate fair bandwidth among different flows [10]. This algorithm maintains a separate queue for each flow and discrimination of traffic sources may be based on packet size or sending rate of source computers. These queues are served by the router in sort of round robin. Fair Queuing is based on finishing time of each packet. It calculates the finishing time of each packet residing at the head of each queue and compares this finishing time. The packet having shortest time is transmitted first.

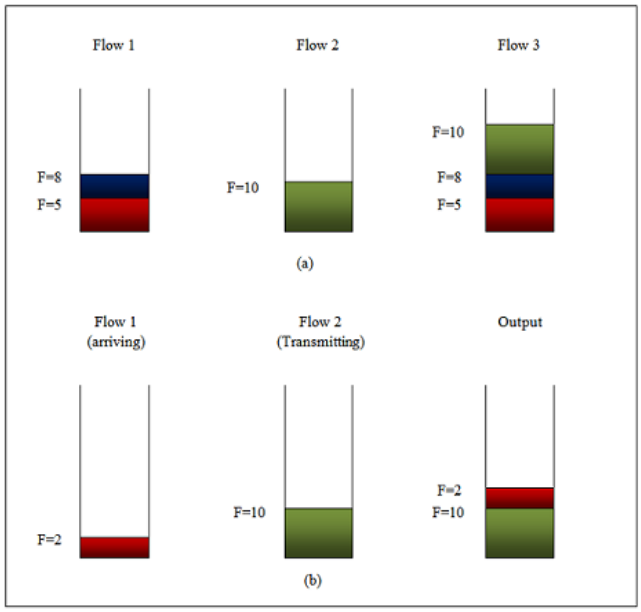


Fig 3: fair queuing example: (a) packet with shortest finishing times transmitted first; (b) already sending packet completed first [17].

Consider an example of Fair Queuing algorithm shown in the Fig. 3. Router discriminates the incoming traffic into different flows, Flow 1 and Flow 2. And the arriving packets are stored into the flow in which they belong. In Fig 3 (a), flow 1 stores two packets one having the finishing time F=8 and another one having 5 and flow 2 stores one packet having finishing time F=10.The finishing time of packet residing at the head of each queue is compared. The packet with finishing time F=8 of flow 1 is compared with the packet with finishing time F=10 of flow 2 and packet with finishing time F=8 is transmitted first because it is shortest finishing time. After fully transmission of packet having finishing time F=8, again it compares packet of flow 1 with packet of flow 2 and finds that packet having finishing time F=5 is shortest so it is transmitted first and then the packet having finishing time F=10 of flow 2 is transmitted. In Fig 3 (b), the packet of flow 2 having finishing time F=10 is being transmitted and a packet in flow 1 arrives having finishing time F=2 but transmission of the packet of flow 2 is not halted and after completion of this transmission it will send the packet with finishing time F=2.

Stochastic Fair Queuing

Stochastic Fair Queuing is an implementation of Fair Queuing. Stochastic Fair Queuing uses a hash algorithm to divide the traffic over a limited number of queues [17]. Due to the hashing in SFQ multiple sessions might end up into the same bucket. SFQ changes its hashing algorithm so that any two colliding sessions will only work for a small number of seconds.

Deficit Round Robin

Deficit Round Robin uses three parameters, weight, DeficitCounter and quantum [18]. Weight decides percentage of output port must be allocated to the queue. DeficitCounter decides whether a queue is permitted to send data packet or not. Quantum is proportional to the weight of a queue and also represented in terms of bytes [19]. Function of Deficit Round Robin is shown in figure 4.

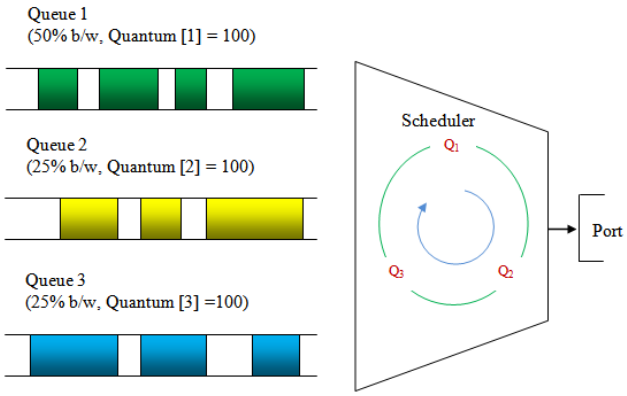


Fig 4: Deficit Round Robin [19].

Random Early Detection

The objective of Random Early Detection (RED) algorithm is to fairly distribute the effect of congestion among all traffic sources competing for the bandwidth by random dropping the packet from the queue. To avoid the congestion, packet is early dropped when the congestion is imminent. To achieve these objectives, it monitors the average queue size to find out whether it lies between some minimum threshold value and maximum threshold value. If it is true then the arriving packet is marked or dropped with some probability that is increasing function of average queue size. All the arriving packets are dropped when the variable does not lie between minimum and maximum threshold values.

Simulations for Studying the Effect of Attack Intensities and Buffer Size on Various Queuing Algorithms

Fig. 5 shows the simulation structure for checking the performance of different queuing algorithms. Node 0, node 1,

node 2, node 3, node 4, node 5 and node 6 represent the legitimate TCP user, legitimate UDP user, attacker1, attacker2, attacker3, router and receiver respectively. All the links between nodes have 1Mbps bandwidth and propagation delay of 100ms. These nodes send data packets to receiver and packets first stored on router (node 5) and then forwarded. Each router in internet maintains queues to store data packets and the size of queue may vary. Attackers are using UDP type flood attack. Here we are going to study the effect of variation in attack intensities and buffer size on different queuing algorithms. Suppose there is 1Mbps link between any particular router and destination node. Entire incoming packets at router are capable of 1.3Mbps. Router can transfer only 1Mbps of data at a time all other data capable of 0.3Mbps or 30% of data will be dropped by router and also known as 30% attack intensity. Table 1 shows the details of simulation parameters.

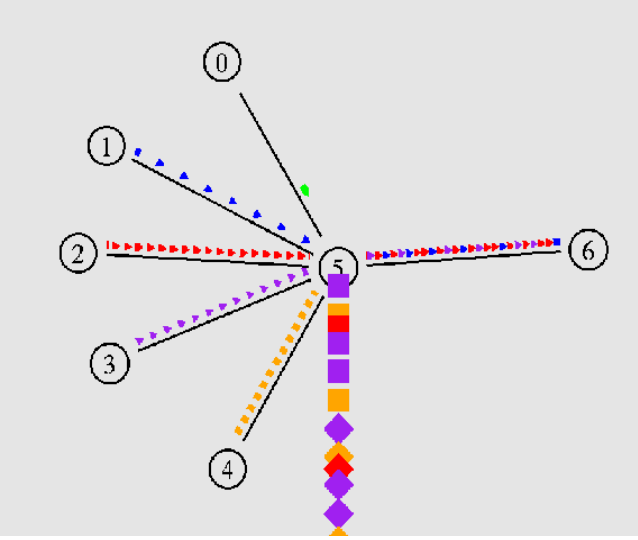


Fig 5: Simulation structure

Table 1: Simulation Parameters

Number of Nodes	7
Link bandwidth between Nodes	1Mbps
Propagation Delay	100ms
Simulation Time	50 seconds
Attack intensity's range	20% to 60%

Buffer size effect on DropTail

Figure 6 (a), 6 (b) and 6 (c) show the effect of buffer size on DropTail algorithm against 20%, 60% and 120% attack intensities respectively.

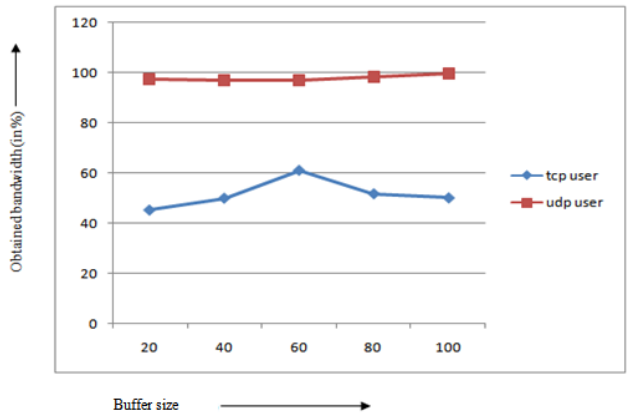


Fig 6 (a): Buffer size effect on DropTail against 20% attack intensity

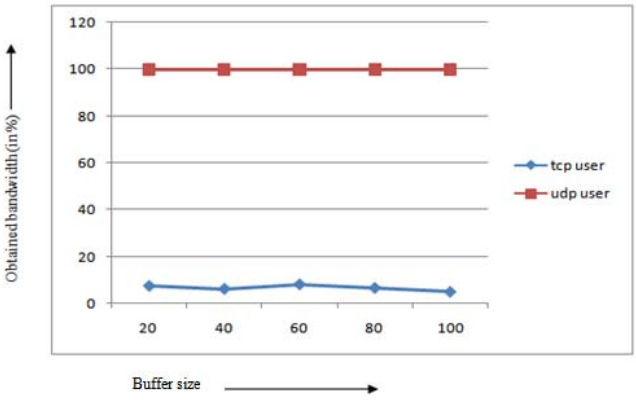


Fig 6 (b): Buffer size effect on DropTail against 60% attack intensity

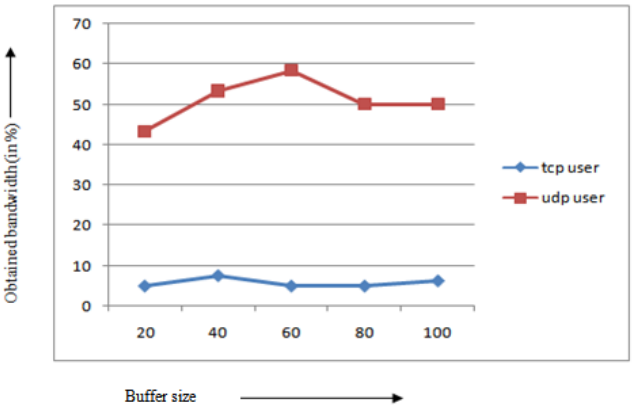


Fig 6 (c): Buffer size effect on DropTail against 120% attack intensity

Figure 6 (a) shows that on increasing the buffer size gradually from 20 to 60, there is no much effect on bandwidth obtained by legitimate UDP user but bandwidth obtained by legitimate TCP user is gradually increasing. And during the

buffer size from 60 to 80, bandwidth obtained by TCP user is gradually decreasing. Figure 6 (b) shows that there is no effect on legitimate UDP user but legitimate TCP user is getting maximum bandwidth while having buffer size 60. Figure 6 (c) also shows that there is no much effect on legitimate TCP user but legitimate UDP user is getting maximum bandwidth while having buffer size 60. Now the conclusion is that DropTail performance is best when buffer size is 60 against various attack intensities.

Buffer size effect on Random Early Detection

Figure 7 (a), 7 (b) and 7 (c) show the effect of buffer size on SFQ algorithm against 20%, 60% and 120% attack intensities respectively. Figure 7 (a) shows that there is no effect of buffer size on legitimate TCP user and legitimate UDP user gets the maximum bandwidth when buffer size is greater than or equal to 60. Figure 7 (b) shows that legitimate user is getting maximum bandwidth when buffer size is 60. While there is a constant effect on legitimate TCP user during the variation in buffer size from 20 to 60. Legitimate TCP user is getting less bandwidth while having buffer size greater than 60. Figure 7 (c) shows that legitimate UDP user is getting constant bandwidth during buffer size from 40 to 100. Now the conclusion is that RED algorithm is giving the best performance in case of buffer size is equal to 60.

Buffer size effect on FQ, SFQ, DRR algorithms

From various simulation studies it is clear that there is a constant effect of buffer size on queuing algorithms FQ, SFQ and DRR against 20%, 60% and 120% attack intensities.

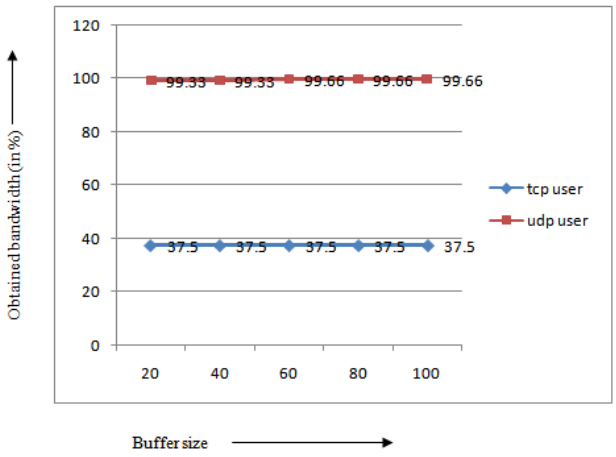


Fig 7 (a): Buffer size effect on RED against 20% attack intensity

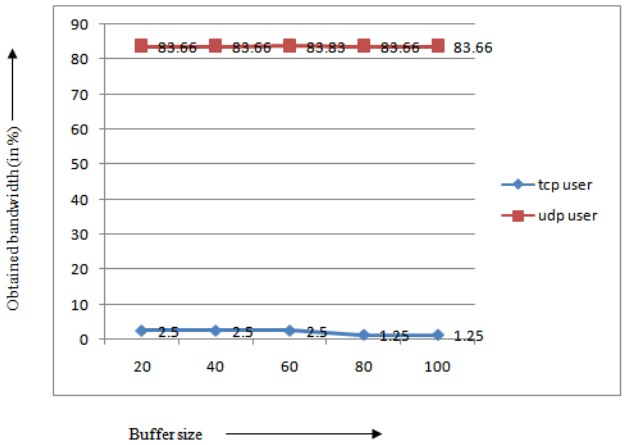


Fig 7 (b): Buffer size effect on RED against 60% attack intensity

DropTail Performance against attack intensities

In this section we are going to check the performance of DropTail algorithm on queue limit 80 against different attack intensities. Fig. 8 shows the performance of DropTail algorithm. It is clear from the graph that on increasing the attack intensity, bandwidth obtained by legitimate TCP and UDP users are gradually decreasing.

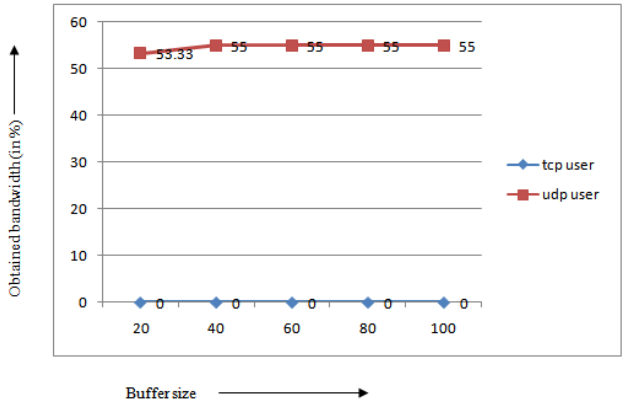


Fig 7 (c): Buffer size effect on RED against 120% attack intensity

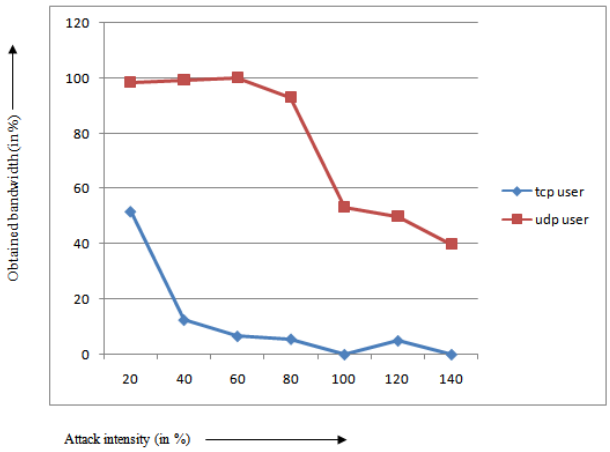


Fig. 8 DropTail performance

Fair Queuing Performance against attack intensities
Fig. 9 shows the performance of Fair Queuing algorithm. From the graph it is clear that bandwidths obtained by legitimate users are decreasing when attack intensity is increasing from 20% to 40%. And there is a constant effect of attack intensities varying from 40% to 140%.

Stochastic Fair Queuing Performance against attack intensities
Fig. 11 shows the performance of Stochastic Fair Queuing algorithm. Graph shows a constant effect of attack intensities on legitimate TCP and UDP users.

Deficit Round Robin Performance against attack intensities
Fig. 11 shows the performance of Deficit Round Robin algorithm. It shows that on increasing the attack intensity bandwidth obtained by legitimate TCP user is gradually decreasing while there is a constant effect on bandwidth obtained by UDP user during attack intensity varying from 40% to 140%.

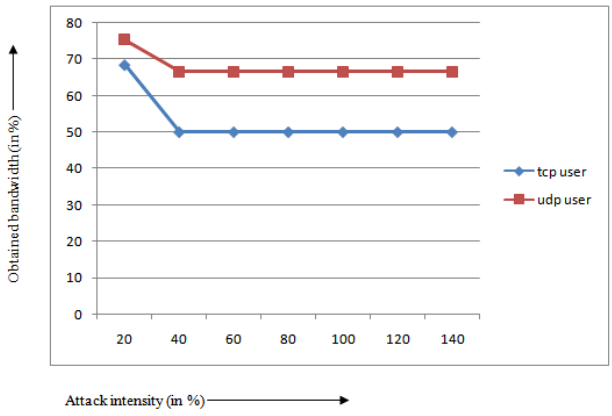


Fig. 9 Fair Queuing performance

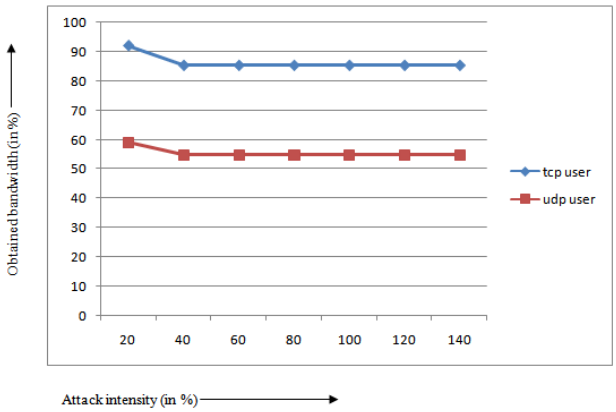


Fig. 10 Stochastic Fair Queuing performance

Deficit Round Robin Performance against attack intensities
Fig. 11 shows the performance of Deficit Round Robin algorithm. It shows that on increasing the attack intensity bandwidth obtained by legitimate TCP user is gradually decreasing while there is a constant effect on bandwidth obtained by UDP user during attack intensity varying from 40% to 140%.

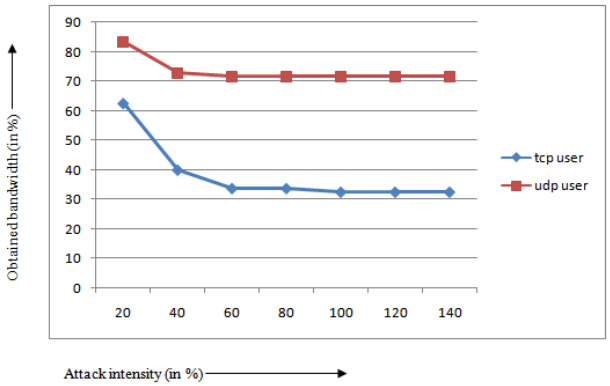


Fig. 11: Deficit Round Robin performance

Random Early Performance against attack intensities
Fig. 12 shows the performance of Random Early Detection algorithm. This algorithm is not useful for TCP user because it gets nothing when attack intensity goes above 60%. While bandwidth obtained by legitimate UDP user is gradually decreasing on increasing the attack intensity.

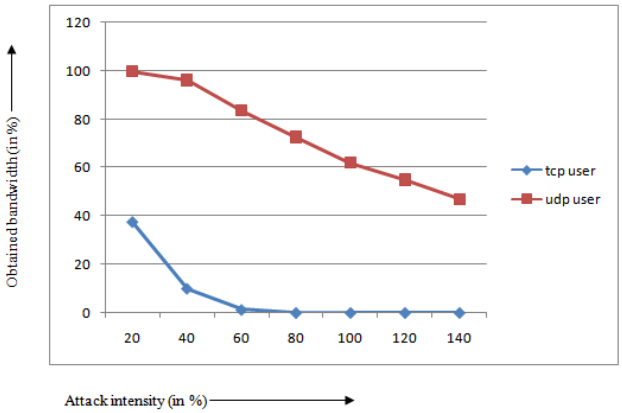


Fig. 12: Random Early Detection performance

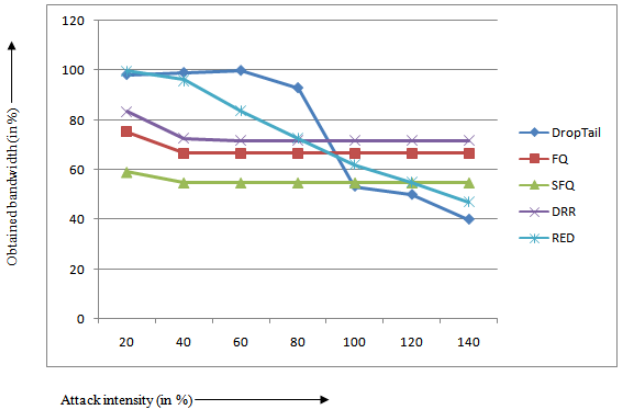


Fig. 13 (b) comparison of throughputs of UDP user on different queuing algorithms

Performance comparison of Queuing Algorithms

Fig. 13(a) and Fig. 13 (b) show the comparison of bandwidth obtained by legitimate TCP and UDP users on different queuing algorithms against different attack intensities. According to Fig. 13(a) legitimate TCP user is getting maximum throughputs in case of Stochastic Fair Queuing algorithm. Fig. 13(b) shows that legitimate UDP user is getting maximum bandwidth 75% in case of Deficit Round Robin. But in case of Deficit Round Robin legitimate TCP user is getting bandwidth 33%. So if we consider throughputs of TCP user then it is not good enough but if we consider only for UDP user then Deficit Round Robin is best algorithm. Fair Queuing algorithm is the second best algorithm to provide the maximum bandwidth to the legitimate UDP users. It is providing 70% bandwidth to legitimate UDP user and 50% to legitimate TCP user. While Stochastic Fair Queuing algorithm is providing 85% throughputs to legitimate TCP user and 55% to legitimate UDP user. So finally, Stochastic Fair Queuing algorithm is best algorithm among all algorithms in case of providing satisfactory bandwidth to the legitimate users in case of having both legitimate TCP and UDP users in network. And second best algorithm is Fair Queuing algorithm.

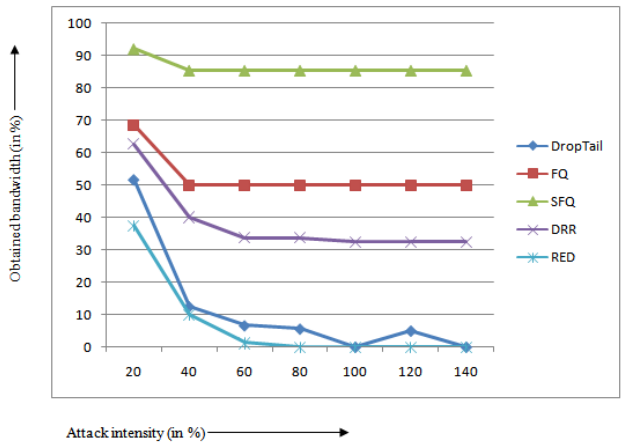


Fig. 13 (a) comparison of throughputs of TCP user on different queuing algorithms

Conclusion

We have explained some basic overview of DDoS, attacking methods; DDoS attack tool kits and DDoS prevention mechanisms. We also discussed the various queuing algorithms. Mainly we focused on buffer size's effect and attack intensities' effect on various queuing algorithms.

Simulation result shows that DropTail and Random Early Detection (RED) algorithms are giving the best performance in case of buffer size that is 60. While there is no effect on FQ, SFQ and DRR algorithms against variation in buffer size. We also found that Stochastic Fair Queuing is the best algorithm against attack intensities in terms of providing maximum bandwidth to the legitimate users. The results indicate that we must set buffer size 60 in case of DropTail and RED algorithms.

References

[1] Stephen M. Specht, Ruby B. Lee, "Distributed Denial of Service: Taxonomies of Attacks, Tools, and Countermeasures". In Proceedings of the 17th International Conference on Parallel and Distributed Computing Systems, 2004.

[2] Jelena Mirkovic, Sven Dietrich, David Dittrich, Peter Reither, "Internet Denial of Service: Attack and Defense Mechanisms", Publisher: Prentice Hall PTR, 2004.

[3] Nathalie Weiler, "Honeypots for Distributed Denial of Service Attacks". In Proceedings of the Eleventh IEEE International Workshops on Enabling Technologies 2002.

[4] http://www.pctools.com/security-news/script-kiddie.

[5] D. Dittrich, "The DoS project's 'Trinoo' distributed denial of service attack tool," Oct. 1999; "The 'Stacheldraht' distributed denial of service attack tool," Dec. 1999; "The 'Tribe Flood Network' distributed denial of service attack tool," Oct. 1999, http://www.washington.edu/People/dad.

[6] D. Dittrich, "The Stacheldraht Distributed Denial of Service Attack Tool," December 1999,

http://staff.washington.edu/dittrich/misc/stacheldraht.analysis.txt.

[7] D. Dittrich, S. Dietrich, and N. Long, "An analysis of the 'Shaft' distributed denial of device tool", 2000, http://netsec.gsfc.nasa.gov/~spock/shaft_analysis.txt

[8] C. Adams and J. Gilchrist, "RFC 2612: The CAST-256 encryption algorithm," June 1999, http://www.cis.ohiostate.edu/htbin/rfc/rfc2612.html.

[9] B. B. Gupta, R. C. Joshi, Manoj Misra,"Distributed Denial of Service Prevention Techniques", International Journal of Computer and Electrical Engineering, April, 2010

[10] Felix Lau, Rubin H. Stuart, Smith H. Michael, and et al., "Distributed Denial of Service Attacks," in Proceedings of 2000 IEEE International Conference on Systems, Man, and Cybernetics, Nashville, TN, Vol.3, pp.2275-2280, 2000.

[11] http://www.checkpoint.com/resources/firewall/.

[12] X. Geng, A.B. Whinston, Defeating Distributed Denial of Service attacks, IEEE IT Professional 2 (4) (2000) 36–42.

[13] M.Nagaratna, V.Kamakshi Prasad, S.Tanuz Kumar, "Detecting and Preventing IP-spoofed DDoS Attacks by Encrypted Marking based Detection And Filtering (EMDAF)", International Conference on Advances in Recent Technologies in Communication and Computing, 2009.

[14] P. Ferguson, and D. Senie, "Network ingress filtering: Defeating denial of ser-vice attacks which employ IP source address spoofing," RFC 2267, the Internet Engineering Task Force (IETF), 1998.

[15] Baker, F. "Requirements for IP version 4 routers," RFC 1812, Internet Engineering Task Force (IETF).Go online to www.ietf.org.

[16] Yunji Ma, "An Effective Method for Defense against IP Spoofing Attack", 2010 IEEE.

[17] http://nms.csail.mit.edu/6.829-f06/lectures/bruce-queue.pdf

[18] M. Shreedhar, George Varghese, "Eficient Fair Queuing using Deficit Round Robin", Microsoft Corporation.

[19] Chuck Semiria, "Supporting Differentiated Service Classes: Queue Scheduling Disciplines", Juniper Networks, Inc.

Technical Sponsors:

IEEE Delhi Section, IEEE Computer Society Chapter, Delhi Section & IETE Delhi Centre

IEEE

IEEE COMPUTER SOCIETY

THE INSTITUTION OF ELECTRONICS AND TELECOMMUNICATION ENGINEERS
IETE
सह वीर्यं करवावहै

A·P·I·I·T
ASIA PACIFIC INSTITUTE OF INFORMATION TECHNOLOGY
SD INDIA

ISO 9001-2008 Certified

Sponsored by:

HDFC BANK

AIRCEL

ISBN 81-87885-03-3

ABC Group of Publication
9, Indira Colony, Vikram Marg, Karnal
Tel: +91-184-6540168, +91-9215508638
Email: jobs@abccomputers.in
Website: www.abccomputers.in

₹ 2500