5-19-2012

Proceedings of International Conference on Computer Science and Information Technology

Prof. Srikanta Patnaik Chief Mentor
IRNet India, patnaik_srikanta@yahoo.co.in

Recommended Citation

Patnaik, Prof. Srikanta Chief Mentor, "Proceedings of International Conference on Computer Science and Information Technology" (2012). *Conference Proceedings - Full Volumes*. 16.
https://www.interscience.in/conf_proc_volumes/16

Proceedings Of International Conference on

COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

(ICCSIT-2012)

19TH MAY, 2012

Organized by:

Interscience Research Network
Bhubaneswar, India

About **ICCSIT-2012**

Computer Science and Information Technology have a profound influence on all branch of science, engineering, management as well. New technologies are constantly emerging, which are enabling applications in various domains and services. International Conference on Computer Science and Information Technology (CSIT) is organized by IRNet for the presentation of technological advancement and research results in the fields of theoretical, experimental, and applied area of Computer Science and Information Technology. CSIT aims to bring together developers, users, academicians and researchers in the information technology and computer science for sharing and exploring new areas of research and development and to discuss emerging issues faced by them. ***Topics of interest for submission include, but are not limited to*:**

Algorithms
Automated Software Engineering
Bioinformatics and Scientific Computing
Compilers and Interpreters
Computer Animation
Computer Architecture and Embedded Systems
Computer Games
Computer Graphics and Multimedia
Computer Networks
Computer Security
Computer Vision
Computing Ethics
Control Systems
Data Compression
Data Mining
Digital Library
Digital System and Logic Design
Distributed Systems
Event Driven Programming
High Performance Computing
Image Processing
Information Systems
Knowledge Data Engineering
Multimedia Applications
Neural Networks
Pattern Recognition
Programming Languages
Robotics and Automation
Software Engineering & CASE
Technology in Education
Theoretical Computer Science

Artificial Intelligence
Bio-informatics
Biomedical Engineering
Computational Intelligence
Computer Architecture & VLSI
Computer Based Education
Computer Graphics & Virtual Reality
Computer Modeling
Computer Networks and Data Communication
Computer Simulation
Computer-aided Design/Manufacturing
Computing Practices & Applications
Data Communications
Data Encryption
Database Systems
Digital Signal and Image Processing
Distributed and Parallel Processing
E-commerce and E-governance
Expert Systems
Human Computer Interaction
Information Retrieval
Internet and Web Applications
Mobile Computing
Natural Language Processing
Parallel and Distributed Computing
Performance Evaluation
Reconfigurable Computing Systems
Security & Cryptography
System Security
Technology Management
Ubiquitous Computing

Organizing Committee

Programmme Chair

Dr. Harsh K Verma
Associate Professor and Head of the Department,
Department of Computer Science and Engineering.
NIT, Jalandhar.

Secretary Bangalore Chapter:
Miss. Suchana Mishra
Dayananda Sagar College of engineering
Bangalore, Karnataka

Team Members:

Prof. Sushanta Kumar Panigrahi
Prof. Mritunjay Sharma
Prof. Sharada Prasad Sahoo.
Prof. Sanjay Sharma

First Impression : 2012

Proceedings of International Conference On

Computer Science and Information Technology

DISCLAIMER

The authors are solely responsible for the contents of the papers complied 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.

ISBN : 978-93-81693-28-5

Published by :

IPM PVT. LTD., Interscience Campus
At/PO.: Kantabada, Via: Janla, Dist: Khurda, Pin- 752054
Publisher's Website : www.interscience.in
E-mail: ipm.bbsr@gmail.com

Typeset & Printed by :
IPM PVT. LTD.

TABLE OF CONTENTS

Sl. No.	Topic	Page No.
	Editorial – *Dr. Harsh K Verma*	
1	**Smart Worms Defense and Detection** – *Dasari Nagaraju & C. Shoba Bindu*	1 - 4
2	**Proofing Against ARP Spoofing - A Suggested Prevention Mechanism for ARP Spoofing on a LAN** – *Monica Sam & Daisy Raju*	5 - 6
3	**Correcting False Segmentation In Video Using Image Over-Segmentation** – *Shailaja Surkutlawar & Ramesh K Kulkarni*	7 - 12
4	**Marker Controlled Watershed Segmentation Using Bit-Plane Slicing** – *M. Sivagami & T. Revathi*	13 - 17
5	**Eon of implementingamultifaceted cloud based OCR in Apple's compassionate App Store milieu** – *C. Infant Louis Richards, T.Yuvaraj & J.Sylvester Britto*	18 - 22
6	**Mitigating Packet Dropping Misbehaviour in Multicast MANET Environment** – *K. Praneeth Reddy & Padam Kumar*	23 – 31
7	**A Methodology for Test Source Reuse in Virtual System Prototype (VSP)** – *Bhargava C R, Pawankumar B, Narayanan S, Kariyappa B S & Kamalakar R*	32 -36
8	**Management of Network Elements using NETCONF** – *S.Sindhu, G.Sadashivappa & Vignesh C.R*	37 - 41
9	**Challenges and Best Practices in Integration Testing For Service Oriented Web Applications** – *Imran Akhtar Khan & Roopa Singh*	42 - 46
10	**Association Rule Mining for Gene Expression Data** – *O. V. Kale & B. F. Momin*	47 - 51
11	**Performance Analysis And Tuning Tool For IBM DB2** – *Ankit Chouksey*	52 - 56

12 **EyeBlink Controlled Human Computer Interface for Physically Challenged People** 57 - 61

– *Naresh E & Praveen Kumar B P*

13 **Watermarking of Grayscale Text Document Images Using Histogram Swapping** 62 - 65

– *Jayanth Silesh S, Anvitha Jaishankar & Rashmi N*

14 **A Probabilistic Clustering Based Algorithm For Textual Data Categorization** 66 - 72

– *Keerthi Thota & DS Bhupal Naik*

15 **Hop-by-hop Congestion Control over a Wireless Multi-hop Network** 73 - 76

Swapna Priya Lanka & Poorna Satyanarayana. B

16 **Record Matching for Web Databases by Domain-specific Query Probing** 77 - 82

– *V.Alekhya & Ds.Bhupal Naik*

17 **Providing the Boundary line Controlled Request with Adaptable Transmission Rates in WDM Mesh Networks** 83 - 88

– *AnushaAnnapureddy & B.Balaji*

18 **Artificial Bee Colony algorithm and its Application for Image Fusion** 89 - 94

– *Bhavya V S,Navyashree K M, Prabhat Kumar Sharma, Sunil K S, Pavithra P*

19 **Generating Spring Graph to Infer RDF Data** 95 - 100

‾ *Meera Mudholkar, Saloni Peswani, Seema Kela, Jagannath Aghav, Harish Jadhao & Prashant Gaikwad*

20 **Self-Managing Performance in Application Servers – Modelling and Data Architecture** 101 - 106

– *Ravi Kumar G, C.Muthusamy & A.Vinaya Babu*

21 **Decision Tree-Rough Set Based System with HHMM For Predicting the Stock Market Trends** 107 - 113

– *Shweta Tiwari & Rekha Pandit*

22 **Computer Aided Visualization Model for Speech Perception and Intelligibility** 114 - 118

– *A Rathinavelu [AR] & M. Karthikeyan*

23 **Frequent Pattern Analysis For Effective Classification Using Parallel Apriori** 119 - 123

‾ *N.S. Jagadeesh & E.Madhusudhana Reddy*

24 **Secured Data Tranamission using Snort Rules and Mining Techniques** 124 - 132

– *R.Venkatramana & M.Sreedevi*

25 **SAT: A Security Architecture Achieving Routing Anonymity in Wireless Mesh Networks** 133 - 136

– *Tanuja Chinta & K. Madhavi*

26 **Image Enhancement Based on Lazy Wavelet Transform Using Adaptive Wiener With Threshold** 137 - 142

¯ *G. M. Rajathi, M.Sarathkumar & R. Rangarajan*

27 **Performance Evaluation of Distribution Localization Techniques for Mobile under Water Sensor Networks** 143 - 149

– *Mohan V & Mithun T P*

28 **Exypnos Office** 150 - 154

– *Maruthi* Naik R.K & M. Nagaraja

29 **Provably Secure and Higher Efficient Auditing for Data Storage Security in Cloud Computing** 155 - 165

¯ *Maragatham. S & Ravikumar. P*

30 **Continuous Integration For Enterprise Systems Using Heterogeneous Techniques** 166 - 170

– *Raksha SV*

31 **Maximum Likelihood Factor Analysis Based Speaker Recognition Using Pitch Strength** 171 - 176

– *Rama Koteswara Rao P, Srinivasa Rao Y & Vijaya Kumar D*

32 **Computer Aided Analysis of Lung CT Based on Transform Domain Features** 177 - 182

– *V.Minnal*

33 **Improve the Processing Performance of Top-K Spatial Keyword Query on Road Networks** 183 - 188

– *M. Sreenath Yadav & K.R.K.Satheesh*

34 **Automated SWIM (Single Window Infrastructure Management) for Cloud** 189 - 192

– *Jyothi Ganig B S & Ramesh Namburi*

35 **Server Update Mechanism In MANETS** 193 - 196

– *Nalini Suryadevara & P.Seshu babu*

36 **Removal of High Density Salt-And-Pepper Noise Through Improved Adaptive Median Filter** 197 - 200

– *Anisha Bhatia & Ramesh K Kulkarni*

37 **Design and Implementation of Adaptive Filter for Better Throughput** 201 - 207

– *Shivaraj S B & Bhagya R*

38 **Backbone Node Election In Presence of Misbehaving Node Using Mechanism Design Approach In MANET** 208 - 214

– *Manasa K M, Shanthi M B & Jitendranath Mungara*

39 **Categorization of SMS in Android Smart Phones** 215 - 220

– *Mohammed Mateenuddin Siddiqui & K. V. Bhosle*

40 **Real Time Implementation of Adaptive Image Enhancement Method on Android Platform** 221 - 224

– *Jharna Majumdar, Shiva Sumanth Reddy & Manoj Kumar M*

41 **Improved Near Duplicate Matching Scheme for E-mail Spam Detection** 225 - 230

– *M. Siva Kumar Reddy & B. Krishna Sagar*

42 **Continuous Integration Software Build Enhancement System Based On Agile Model** 231 - 234

– *Harish Reddy Bappa & Poornima M*

43 **Statewide Higher Education Counseling System Using Grid Computing** 235 - 240

– *D. Ramesh & A. Krishnan*

44 **Action Recognition Using Feature Transform Descriptor from Mined Dense Spatio Temporal** 241 - 247

– *N. R .Vikram & P. M. Ashokkumar*

45 **Secured Data Transmission using Snort Rules and Mining Techniques** 248 - 256

– *R.Venkatramana & M.Sreedevi*

46 **Iris Feature Extraction Using Wavelet Maxima Components for Biometric Identification Systems** 257 - 262

– *Madhuri Sachane, V.M. Jain*

47 **Determining the Positioning Algorithm for Fingerprinting Using WLAN** 263 - 267

– *Indu Maurya , Vivek Singh Sengar & Priyanka Maurya*

48 **GLRT Design of Mimo Radar In Spherically Invariant Random Vector Clutter** 268 - 275

– *Pradeep G B & B. Roja Reddy*

49 **Predicting Medical Diseases Using Data Mining** 276 – 281

– *Shweta A.Gode & G.R.Bamnote*

50 **Density-Based Clustering with Statistical Functions in Spatial Dataset** 282 - 284

– *Sahar Zarei & Ehsan Moradi Motlagh Jonaghani*

51 **Visual Visitor Verifier** 285 - 290

– *Kishore G. R & Y Manjula*

52 **Linear Face Recognition Techniques** 291-297

– *Kavita Bramhankar, Nitin Mishra & Priti Subramanium*

53 **Adaptive Coding Techniques to Improve BER in OFDM System** 298-303

– *Rajeshree Raut, Priti Subramanium & Kavita Bramahankar*

54 **Facilitate Traffic Allocation Among Multiple Paths Based on Empirical Jamming Statistics** 304-309

– *B. Sreekanth & P. V. Naga Jayudu*

55 **An Analysis of Rigid Image Alignment Computer Vision Algorithms** 310-316

– *Rajeshree Joshi & Robert Cook*

Editorial

The mushrooming growth of the IT industry in the 21st century determines the pace of research and innovation across the globe. In a similar fashion Computer Science has acquired a path breaking trend by making a swift in a number of cross functional disciplines like Bio-Science, Health Science, Performance Engineering, Applied Behavioral Science, and Intelligence. It seems like the quest of Homo Sapience Community to integrate this world with a vision of Exchange of Knowledge and Culture is coming at the end. Apparently the quotation "Shrunken Earth, Shrinking Humanity" holds true as the connectivity and the flux of information remains on a simple command over an internet protocol address. Still there remains a substantial relativity in both the disciplines which underscores further extension of existing literature to augment the socio-economic relevancy of these two fields of study. The IT tycoon Microsoft addressing at the annual Worldwide Partner Conference in Los Angeles introduced Cloud ERP (Enterprise Resource Planning,) and updated CRM (Customer Relationship Management) software which emphasizes the ongoing research on capacity building of the Internal Business Process. It is worth mentioning here that Hewlett-Packard has been with flying colors with 4G touch pad removing comfort ability barriers with 2G and 3G. If we progress, the discussion will never limit because advancement is seamlessly flowing at the most efficient and state-of-the art universities and research labs like Laboratory for Advanced Systems Research, University of California. Unquestionably apex bodies like UNO, WTO and IBRD include these two disciplines in their millennium development agenda, realizing the aftermath of the various application projects like VSAT, POLNET, EDUSAT and many more. 'IT' has magnified the influence of knowledge management and congruently responding to social and industrial revolution.

Although the discipline like electrical engineering has narrated academic maturity in the last decades, but the limitations of the non renewable energy sources, turbulence and disturbances in the energy propagation cascades various insightfulness and stimulation in post classical electrical era. Evidence shows that there are phenomenal supplements in power generation and control after the introduction of Energy Management System (EMS) supported by Supervisory Control and Data Acquisition (SCADA). As there is increasing focus on strengthening the capacity of the power houses with the existing resources or constraints some new dimensions like FACTS, Optimal System Generation, High Voltage DC transmission system, Power Generation Control, Soft Computing, Compensation of transmission line, Protection scheme of generator, Loss calculation, economics of generation, fault analysis in power systems are emerging. Since the world is suffering with water, food, and energy crisis, energy consumption has social relevancy.

Keeping view of the ongoing energy and power issues many action research can be initiated by the research fraternity of this domain. The conference is a thought provoking outcome of all these interrelated facts.

In the quest of making this earth a better place to live we have to make a strong hold upon sustainable energy source. Sustainable energy sources include all renewable energy sources, such as hydroelectricity, solar energy, wind energy, wave power, geothermal energy, bioenergy, and tidal power. It usually also includes technologies designed to improve energy efficiency. Energy efficiency and renewable energy are said to be the twin pillars of sustainable energy. Renewable energy technologies are essential contributors to sustainable energy as they generally contribute to world energy security, reducing dependence on fossil fuel resources, and providing opportunities for mitigating greenhouse gases.

It's my pleasure to welcome all the participants, delegates and organizer to this international conference. In the process of organizing this conference IRNet family members have shown their commitment and dedication. I sincerely thank all the authors for their invaluable contribution to this conference. I am indebted towards the reviewers and Board of Editors for their generous gifts of time, energy and effort.

Editor-in-Chief
Dr. Harsh K Verma
Associate Professor and Head of the Department,
Department of Computer Science and Engineering.
NIT, Jalandhar.

Smart Worms Defense and Detection

Dasari Nagaraju & C. Shoba Bindu

CSE Dept, JNTUA College of Engineering, Anantapur, Andhra Pradesh, India
E-mail : raj2dasari@gmail.com, shobabindhu@gmail.com

Abstract **-** There are several worm attacks in the recent years, this leads to an essentiality of producing new detection technique for worm attacks. In this paper we present a spectrum based smart worm detection scheme, this is based on the idea of detection of worm in the frequency domain. This scheme uses the power spectral density of the scan traffic volume and its corresponding flatness measure to distinguish the smart worm traffic from background traffic. This scheme showed better results against the smart worms and also for the c-worm detection.

Keywords **-** *security, worm, and detection*

I. INTRODUCTION

Smart worms are malicious software that can self – propagate across the internet, i.e., compromise vulnerable hosts and use them to attack other victims. Since the early stage of the internet, worms have caused enormous damage and been a significant security threat. For example, the Morris worm infected 10% of all hosts in the internet in 1988. The Code Red worm compromised at least 359,000 hosts in one day in 2001 [1], and the Storm botnet affected the tens of millions of hosts in 2007.

Due to the substantial damage caused by worms in the past years, there have been significant efforts on developing detection and defense mechanisms against worms. In this paper, we conduct a systematic study on smart worms. The smart worms have a self propagating behavior similar to traditional worms. The smart worms are quite different from traditional worms in which it manipulates any noticeable trends in the number of infected computers such a manipulation of the scan traffic volume prevents exhibition of any exponentially increasing trends or even crossing of thresholds that are tracked by existing system [2], [3].

Based on the observation, we adopt frequency domain analysis technique and develop a detection scheme against wide-spreading of the smart worms; particularly we develop a spectrum based detection scheme that uses the power spectral density distribution of scan traffic volume in the frequency domain and its corresponding spectral flatness measure to distinguish the smart worm traffic from non worm traffic [4].

We define the metrics to evaluate the performance; the metrics are Detection time and Detection rate. Our evaluation data clearly demonstrate that spectrum based detection scheme achieves much better performance against the smart worm propagation.

The remainder of this paper is organized as follows. In Section 2, we introduce the spread of smart worms. In Section 3, we introduce the smart worm propagation model. In Section 4, we introduce the spectrum based detection scheme against the smart worms. In Section 5, the performance evaluation and the results of our detection scheme is provided. We conclude the paper in Section 6.

II. SPREAD OF SMART WORMS

In this section, we describe how smart worms spread, then introduce the parameters used in the spread of smart worms. Finally, we present pure random scanning model.

When a smart worm is fired into the Internet, it simultaneously scans many machines in an attempt to find a vulnerable host to infect. When it finally finds its prey, it sends out a probe to infect the target. If successful, a copy of this worm is transferred to this new host. This new host then begins running the worm and tries to infect other machines. During the worm's spreading process, some machines might stop functioning properly, forcing the users to reboot these computers or at least kill some of the processes that may have been exploited by the worm. Then these infected machines become vulnerable machines again, and are still inclined to further infection. When the worm is detected, people will try to slow it down or stop it. A patch, which repairs the security hole of the machines, is used to defend against worms. When an infected or vulnerable machine is patched, it becomes an invulnerable machine.

III. SMART WORM PROPAGATION MODEL

To understand the characteristics of smart worms, we adopt the epidemiological model, which has been extensively used for worm propagation modeling [5]. This model matches the dynamics of real worm propagation over the hosts quite well.

Particularly, the epidemiological model assumes that any given computer is in one of the following states: immune, vulnerable, or infected. An immune computer is one that cannot infect by a worm; a vulnerable computer is one that has the potential of being infected by a worm; an infected computer is one that has been infected by a worm. The epidemiological model can expressed as,

$$\frac{dn}{dt} = \beta . n . [N - n], \qquad (1)$$

TABLE 1

THE PARAMETERS FOR THE SPREAD OF SMART WORMS

Notation	Explanation
n	The number of infected computers at time t.
N	(=T.p1.p2) the number of vulnerable computers on the internet.
T	Total number of IP addresses on the internet.
P1	The ratio of total number of computers on the internet over T.
P2	Ratio of total number of vulnerable computers on the internet over the total number of computers on the internet
β	Pair wise infection rate.

This epidemiological model works with the principle of disease propagation, which has been used for the worm propagation modeling. It uses a continuous time differential equations.

IV. SMART WORM DETECTION SCHEME

In this section we develop a spectrum based detection scheme. Our detection scheme captures the distinct patterns of the smart worms in the frequency domain, and thereby has the potential of effectively detecting the smart worm propagation.

In order to identify the smart worms propagation in the frequency domain, we use the distribution of Power spectral density (PSD) and its corresponding Spectral flatness measure (SFM) of the scan traffic. Particularly, PSD describes how the power of a time series is distributed in the frequency domain. Mathematically, it is defined as the Fourier transform of the auto correlation of a time series. The time series corresponding to the changes in the number of worm instances that actively conduct scans over time.

4.1 SPECTRUM BASED DETECTION

We know present the details of spectrum based detection scheme. Similar to other detection schemes [6], [7], we use the destination count as the number of unique destination hosts targeted by launching scans during worm propagation. The distribution of PSD and its corresponding SFM are used to distinguish the smart worm scan traffic and non worm scan traffic. In our detection scheme, the detection data is further processed in order to obtain its PSD and SFM. In the following we detail how the PSD and SFM are determined during the processing of the detection data.

4.1.1 Power Spectral Density (PSD)

To obtain the PSD distribution for worm detection data, we need to transform the data from time domain into the frequency domain. To do so, we use source count Z (t) is obtained by counting the unique hosts. Assuming that Z (t) is the source count in the time period [t-1, t] (t ϵ[1, n]), we define the auto-correlation of Z (t) by

$$C_x \text{ (L)} = \text{M [Z (t) Z (t + L)]}. \qquad (2)$$

In formula (2), $C_x\,(L)$ is the correlation of worm detection at time interval L. if a recurring behavior exists, a Fourier transform of auto correlation function reveal such behavior. The discrete Fourier transform of the auto correlation for the function is given as follows,

$$\psi(C_x[L], P) = \sum_{n=0}^{N-1} (C_x\,[L]) \,.\, e^{-j2\pi Pn/N} \qquad (3)$$

Where P=0, 1… N-1.

As the PSD inherently captures any recurring pattern in the frequency domain, the PSD function shows a comparatively even distribution across a wide spectrum range for the normal non-worm scan traffic.

4.1.2 Spectral Flatness Measure (SFM)

We measure the flatness of PSD to distinguish the scan traffic of the smart worm from the normal non worm scan traffic. For this we introduce the Spectral Flatness Measure which can capture anomaly behavior in certain range of frequencies [8], [9]. The SFM is defined as the ratio of the geometric mean to the arithmetic mean of the PSD coefficients. it can be given as,

$$SFM = \frac{[\Pi_{k=1}^{n} S(f_k)]^{\frac{1}{n}}}{\frac{1}{n}\Sigma_{k=1}^{n} S(f_k)} \qquad (4)$$

Where $S(f_k)$a PSD coefficient is for the PSD obtained from the results in Formula (3), SFM is a widely existing measure for discriminating frequencies in various applications.

V. PERFORMANCE EVALUATION

In this section, we report our evaluation results that illustrate the effectiveness of our spectrum based detection scheme against both the smart worms and the c-worm.

5.1.1 Detection Performance for Smart worms

Table 2 shows detection results of detection scheme against the smart worm. The results have been average over 100 smart worm attacks. From this table we can observe that this detection scheme has succeeded in find outing the rate of 84% within the time period of 1000 seconds. We evaluate the detection performance of detection scheme against the C-worm. The detection performance has been averaged over the 100 C-worm attacks. We observe that this scheme has succeeded in find outing the rate of 99 % in the case of C-worm attacks within the time period of 1000 seconds.

TABLE 2

Detection results for the Smart worms and C-worm

worms	Detection time(sec)	Detection rate
Smart worms	1000	84%
C-worm	1000	99%

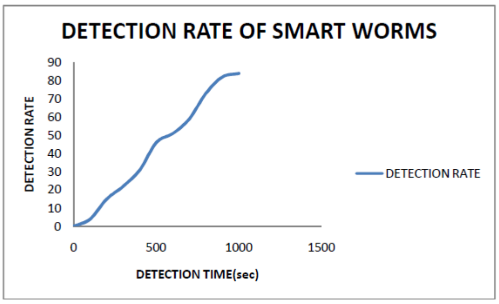


Fig. 1 : Detection Rate of detection schemes against the Smart worms

In view of emphasizing the relative performance of our spectrum based detection scheme, we plot the DT and DR results of smart worms and c-worm in Figs 1 and 2. We can observe from these figures that our spectrum based detection scheme showed better results for smart worms detection and as well as C-worm detection.

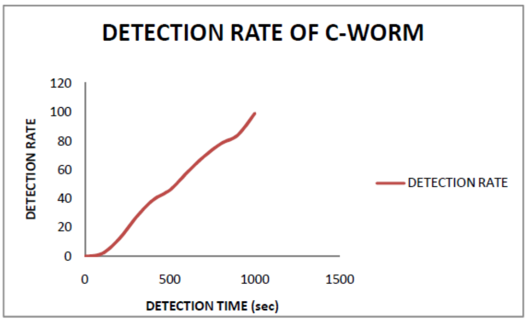


Fig. 2. Detection Rate of detection schemes against the C-worm

VI. CONCLUSIONS

We present the spectrum based detection scheme for the smart worms. Our investigation showed that this detection scheme achieved good performance against the Smart worms and also for the C-worm attacks.

REFERENCES

[1] D. Moore, C. Shanoon, and J. Brown, “Code-red: a case study on the spread and victims of an internet worm”, in Proceedings of the 2-th Internet Measurement Workshop(IMW), Marseille, France, November 2002.

[2] Z. S. Chen, L.X. Gao, and K. Kwait, “Modelling the spread of active worms,” in proceedings of the IEEE Conference on Computer Communications[INFOCOM], San Fransisco, CA, March 2003.

[3] M. Garetto, W. B. Gong, and D. Towsley, ”Modelling malware spreading dynamics”, in proceedings of the IEEE Conference on Computer Communications[INFOCOM], San Fransisco, CA, March 2003.

[4] Wei Yu, Xun Wang, Prasad Calyam, Dong Xuan, and Wei Zhao, ”Modelling and detection of camouflaging worm,” in Proceedings of IEEE transactions on dependable and secure computing, vol. 8, no. 3, MAY-JUNE2011.

[5] D. Moore, V. Paxson, and S. Savage, “Inside the slammer worm,” in IEEE Magazine of Security and Privacy, July 2003.

[6] C. Zou, W. B. Gong, D. Towsely, and L.X. Gao, "Monitoring and early detection for internet worms," in Proceedings of the 10-th ACM Conference on Computer Communications Security (CCS), Washington DC, October 2003.

[7] J. Wu, S. Vangala and L. X. Gao, "An effective architecture and algorithms for detecting worms with various scan techniques," in Proceedings of the 11-th IEEE Network and Distributed System Security Symposium (NDSS), San Diego, CA, February 2004.

[8] S. Soundararajan and D. L. Wang, "A schema-based model for phonemic restoration," Tech. Report, OSU-CISRC-1/04-TRO#, Department of Computer science and Engineering, The Ohio State Univeresity, January 2004.

[9] N. S. Jayant and P.Noll, Digital Coding of Waveforms, Prentice-Hall,1984.

◈◈◈

Proofing Against ARP Spoofing

- A Suggested Prevention Mechanism for ARP Spoofing on a LAN

Monica Sam & Daisy Raju

E-mail : Msam04@gmail.com, Daisy.rajus@gmail.com

I. INTRODUCTION

ARP Spoofing by an attacker on a LAN is one of the well-known vulnerabilities of the ARP protocol. This paper presents a solution to this problem by the use of ICMP ping.

II. DETAILS PROBLEM DESCRIPTION

ARP spoofing is a computer hacking technique whereby an attacker sends fake Address Resolution Protocol (ARP) messages onto a Local Area Network. Generally, the aim is to associate the attacker's MAC address with the IP address of another host (such as the default gateway), causing any traffic meant for that IP address to be sent to the attacker instead.

ARP spoofing may allow an attacker to intercept data frames on a LAN, modify the traffic, or stop the traffic altogether. Often the attack is used as an opening for other attacks, such as denial of service, man in the middle, or session hijacking attacks.

The attack can only be used on networks that make use of the Address Resolution Protocol (ARP), and is limited to local network segments.

The current defenses for ARP spoofing are three-fold:

Static ARP entries

IP-to-MAC mappings in the local ARP cache can be statically defined, and then hosts can be directed to ignore all ARP reply packets. While static entries provide perfect security against spoofing if the operating systems handle them correctly, they result in quadratic maintenance efforts as IP-MAC mappings of all machines in the network have to be distributed to all other machines.

ARP spoofing detection software

Software that detects ARP spoofing generally relies on some form of certification or cross-checking of ARP responses. Uncertified ARP responses are then blocked. These techniques may be integrated with the DHCP server so that both dynamic and static IP addresses are certified. This capability may be implemented in individual hosts or may be integrated into Ethernet switches or other network equipment. The existence of multiple IP addresses associated with a single MAC address may indicate an ARP spoof attack, although there are legitimate uses of such a configuration. In a more passive approach a device listens for ARP replies on a network, and sends a notification via email when an ARP entry changes.

OS security

Operating systems react differently, e.g. Linux ignores unsolicited replies, but on the other hand uses seen requests from other machines to update its cache. Solaris only accepts updates on entries after a timeout.

AntiARP also provides Windows-based spoofing prevention at the kernel level. ArpStar is a Linux module for kernel 2.6 and Linksys routers, which drops invalid packets that violate mapping, and contains an option to repoison/heal.

The simplest form of certification is the use of static, read-only entries for critical services in the ARP cache of a host. This only prevents simple attacks and does not scale on a large network, since the mapping has to be set for each pair of machines resulting in (n*n) ARP caches that have to be configured.

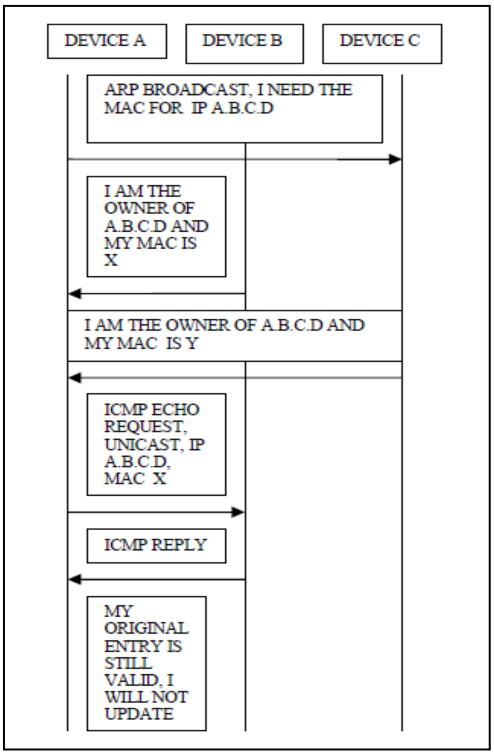
III. SUGGESTED SOLUTION

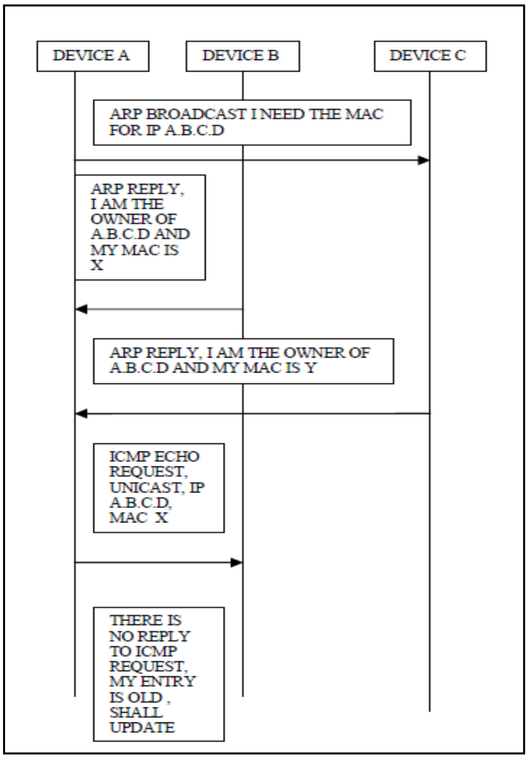
This paper suggests a solution to ARP spoofing that is two-fold.

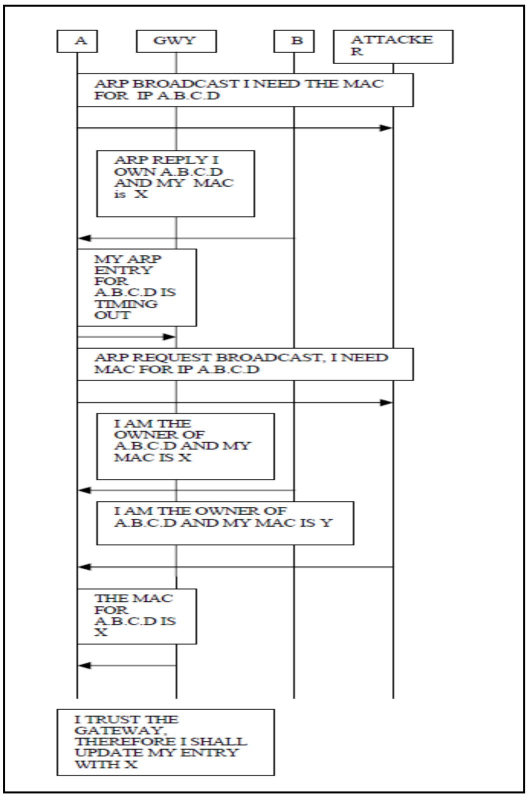
1) The following intelligence can be built into all the devices on which such an ARP spoofing attack is a possibility. ARP operations work as usual – a device sends a broadcast on the LAN to get the MAC address associated with an IP address and the owner of that IP address sends a unicast reply to the requestor. In the case of an attempt by a second device (possible attacker) to over-write an entry on any of these devices, the device simply executes an ICMP ping to the original owner of the IP. That is an ICMP request is sent with the destination MAC (already existing entry) and IP address. If it receives a reply from the original owner, it does not over-write the entry. If there is no reply, it means that the MAC/IP mapping is no longer valid and the entry is over-written. This prevents the devices, especially the gateway from having incorrect entries in its ARP table.

2) In the case of ARP timeout , just before an ARP entry gets timed out , an ARP timeout message can be unicast from the device on which the ARP entry is going to timeout to the gateway. When the MAC for this IP has to be learnt again, the gateway also responds to the broadcast request message. The response from the gateway takes more precedence over any other reply, since the gateway can be trusted more than a new random device. For this, the intelligence to send an ARP reply to the device from which a timeout has been received for the IP the timeout was sent for, should be built into the gateway.

A diagrammatic representation:


DEVICE A
DEVICE B
DEVICE C
ARP BROADCAST, I NEED THE MAC FOR IP A.B.C.D
I AM THE OWNER OF A.B.C.D AND MY MAC IS X
I AM THE OWNER OF A.B.C.D AND MY MAC IS Y
ICMP ECHO REQUEST, UNICAST, IP A.B.C.D, MAC X
ICMP REPLY
MY ORIGINAL ENTRY IS STILL VALID, I WILL NOT UPDATE


DEVICE A
DEVICE B
DEVICE C
ARP BROADCAST I NEED THE MAC FOR IP A.B.C.D
ARP REPLY, I AM THE OWNER OF A.B.C.D AND MY MAC IS X
ARP REPLY, I AM THE OWNER OF A.B.C.D AND MY MAC IS Y
ICMP ECHO REQUEST, UNICAST, IP A.B.C.D, MAC X
THERE IS NO REPLY TO ICMP REQUEST, MY ENTRY IS OLD , SHALL UPDATE


A
GWY
B
ATTACKE R
ARP BROADCAST I NEED THE MAC FOR IP A.B.C.D
ARP REPLY I OWN A.B.C.D AND MY MAC is X
MY ARP ENTRY FOR A.B.C.D IS TIMING OUT
ARP REQUEST BROADCAST, I NEED MAC FOR IP A.B.C.D
I AM THE OWNER OF A.B.C.D AND MY MAC IS X
I AM THE OWNER OF A.B.C.D AND MY MAC IS Y
THE MAC FOR A.B.C.D IS X
I TRUST THE GATEWAY, THEREFORE I SHALL UPDATE MY ENTRY WITH X

REFERENCES

[1] http://tools.ietf.org/html/rfc826
[2] http://tools.ietf.org/html/rfc5227

Correcting False Segmentation In Video Using Image Over-Segmentation

Shailaja Surkutlawar & Ramesh K Kulkarni
EXTC
Vivekananda education society of information technology, Mumbai, India
E-mail : shaili_ayush@yahoo.com, rk1_2002@yahoo.com

***Abstract* -** Moving objects detection is a fundamental step in many vision based applications. Background subtraction is the typical method. When scene exhibits pertinent dynamism method based on mixture of Gaussians is a good balance between accuracy and complexity, but fails due to two kinds of false segmentations i.e moving shadows incorrectly detected as objects and some actual moving objects not detected as moving objects. In computer vision, segmentation refers to process of partitioning a digital image in to multiple segments and goal of segmentation is to simplify and/or change representation of image in to something that is more meaningful and easier to analyse. A colour clustering based on k-means and image over-segmentation are used to segment the input frame into patches and shadow suppression done by HSV colour space, the outputs of mixture of Gaussians are combined with the colour clustered regions to a module for area confidence measurement. In this way, two major segment errors can be corrected. Experimental results show that the proposed approach can significantly enhance segmentation results.

Keywords*— Adaptive Mixture of Gaussian, K-means ,HSV colour space , image over-segmentation.*

I. INTRODUCTION

Moving Objects segmentation is a fundamental and critical task in many vision based applications, such as automated visual surveillance, human-machine interface, and very low-bandwidth telecommunications. A common approach is to perform background subtraction, which identifies moving objects from the difference between the current frame and a reference frame (which often called “background model”). The background model must be representation of the scene with no moving objects and must be kept regularly updated because for some cases, the background is changing when time passes by. Such as view captured by an outdoor surveillance camera, the background is different when sun-light or weather is different. With respect to the state of the art [1-3], a wide variety of approaches performing background subtraction have been developed. A good review for these methods can be found in [4]. Referring to the conclusions of [4], Mixture of Gaussians [5-7] and Kernel density estimation (KDE) [8] can model well the background pdf in general cases and provide higher accuracy compared to other reviewed methods. If speed is concerned, they both have a constant complexity. But KDE has a much higher memory requirement (in order of a 100 frames).

So in the real applications, Mixture of Gaussians is the most frequently used method, as witnessed by the huge amount of literature on it. However, this method also suffers from slow learning at the beginning [9], incapability of identifying moving shadows from the objects casting them [6,10] and unsatisfied results in some cases.The video of the background model is not a literal visualization but it's simply a weighted sum of all components, whether they're part of the background model or not. From the results we can find that due to low rate of background updating, moving shadows, and possible influence by noise, the performance is poor. Though efforts have been imposed by many researchers to improve the algorithm in different senses, we have to say that a comprehensive physical model of the background is really difficult to develop. Therefore, a good post processing may be more suitable in general cases. In this paper, a novel color clustering based post-processing method is proposed, and will be discussed in details in the following sections.

II. BACKGROUND SUBSTRACTION

In the model of Mixture of Gauss [5-7], the background is not a single frame without any moving objects. Gaussian Mixture Model (GMM) is thought to be one of the best background modeling methods and works well when gradual changes appear in the scene . The GMM method models the intensity of each pixel with a mixture of *K* Gaussian distributions. The probability that a certain pixel has a value X_t at time can be written as

$$P(X_t) = \sum_{k=1}^{k} \omega_{k,t} . \eta(X_t, \mu_{k,t}, \Sigma_{k,t}) \qquad (1)$$

where *K* is the number of distributions (currently, from 3 to 5 is used), $\omega_{k,t}$ is the weight of the *k*th Gaussian in the mixture at time *t* , and $\eta(X_t, \mu_{k,t}, \Sigma_{k,t})$ is the Gaussian probability density function. $\eta(X_t, \mu_{k,t}, \Sigma_{k,t}) = \frac{1}{(2\pi)^3/^2 |\Sigma_{k,t}|^1/^2} e^{\left\{\frac{-1(X_t-\mu_{k,t})^T \Sigma_{k,t}^{-1} (X_t-\mu_{k,t})}{2}\right\}}$

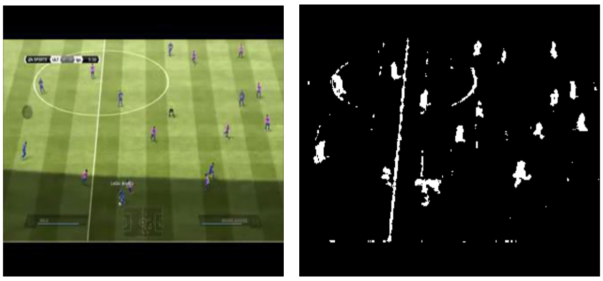
$$(2)$$

where $\mu_{k,t}$ is the mean value and $\sum_{k,t}$ is the covariance of the k_{th} Gaussian at time *t*. For computational reasons, the covariance matrix is assumed to be of the form.

$$\Sigma_{k,t} = \sigma^2 . I \qquad (3)$$

Where σ is the standard deviation.

This assumes that the red, green, and blue pixel values are independent and have the same variance, allowing us to avoid a costly matrix inversion at the expense of some accuracy.



(a) (b)

Fig.1 (a) original frame (b) Extracted moving regions by mixture of Gaussians

III. MOVING SHADOW SUPPRESSION

Shadows are due to the occlusion of light source by an object in the scene. In particular, that part of the object not illuminated is called self-shadow, while the area projected on the scene by the object is called cast shadow [2]. This last one is more properly called moving cast shadow if the object is moving. In literature, many works have been published on shadow detection topic. Jiang and Ward [2] extract both self-shadows and cast shadows from a static image. They use a three level processes approach: the low level process extracts dark regions by thresholding input image; the middle level process detects features in dark regions, such as the vertexes and the gradient of the outline of the dark regions and uses them to further classify the region as penumbra (part of the shadow where the direct light is only partially blocked by the object), self-shadow or cast shadow; the high level process integrates these features and confirms the consistency along the light directions estimated from the lower levels.

Since our work addresses the problem of segmentation of moving objects, we aim to define an approach for detecting moving cast shadows on the background, without computing static shadows (due to static objects). In [3], the authors detail the shadow handling system using signal processing theory. Thus, the appearance of a point belonging to a cast shadow can be described as:

$$S_k(x, y) = E_k(x, y)\ \rho_k(x, y) \qquad (4)$$

where $\boldsymbol{S_k}$ is the image luminance of the point of coordinate (x,y) at time instant *t*. $E_k(x, y)$ is the irradiance and it is computed as follows:

$$E_k(x, y) = \begin{cases} C_A + C_P \cos \angle(N(x, y), L) & \textit{lluminate} \\ C_A & \textit{shadowed} \end{cases} \qquad (5)$$

where $\boldsymbol{C_A}$ and $\boldsymbol{C_P}$ are the intensity of the ambient light and of the light source, respectively, $\boldsymbol{L}$ the direction of the light source and $\boldsymbol{N(x,y)}$ the object surface normal. $\boldsymbol{\rho_k(x, y)}$ is the reflectance of the object surface.

In [3], some hypotheses on the environment are outlined:

I. strong light source

II. static background (and camera)

III. planar background

Most of the papers take implicitly into account these hypotheses. In fact, typically the first step computed for shadow detection is the difference between the current frame and a reference image, that can be the previous frame, as in [3], or a reference frame, typically named background model [4][5][6][1].

we can write this difference $D_k(x, y)$ as:

$$D_k(x,y)= S_{k+1}(x,y) - S_k(x,y) \qquad (6)$$

Let us consider that a previously illuminated point is covered by a cast shadow at frame ***k + 1***. According to the hypothesis 2 of a static background, reflectance $\boldsymbol{\rho_k(x, y)}$ of the background does not change with time, thus we can assume that

$$\rho_{k+1}(x,y) = \rho_k(x,y) = \rho(x, y) \qquad (7)$$

$$D_k(x , y) = \rho(x , y) C_P \cos \angle (N(x , y) , L) \qquad (8)$$

Thus, if hypothesis 1 holds, C_p in eq.8 is high. Summarizing, if hypotheses 1 and 2 hold, difference in eq. 6 is high in presence of cast shadows covering a static background. This implies (as assumed in many papers) that shadow points can be obtained by thresholding the frame difference image. Eq. 8 detects not only shadows, but also foreground points. The papers in literature mainly differ in the way they distinguish between those points. In [4] Kilger uses a background suppression technique to find the moving objects and moving cast shadows in the scene. Then, for each object, it exploits the information on date, time and heading of the road computed by its system to choose whether to look for vertical or horizontal edges to separate shadows from objects. In [7], a the statistical a-posteriori estimation of the pixel probabilities of membership to the class of background, foreground or shadow points. The authors use three sources of information: local, based on the assumption that the appearance of a shadowed pixel can be approximated using a linear transformation of the underlying pixel appearance, according with the fact that the difference of eq. 8 should be positive; spatial, which iterates the local computation by re-computing the a-priori probabilities using the a-posteriori probabilities of the neighborhood; temporal, which predicts the position of shadows and objects from previous frames, therefore adapting the a-priori probabilities. The approach in [3] exploits the local appearance change due to shadow by computing the ratio $R_k(x, y)$ between the appearance of the pixel in the actual frame and the appearance in a reference frame.

$$R_k(x,y) = \frac{S_{k+1}(x,y)}{S_k(x,y)} \qquad (9)$$

that can be rewritten as ratio between irradiance and reflectance by using eqs. 4 and 7 :

$$R_k(x,y) = \frac{E_{k+1}(x,y)}{E_k(x,y)} \qquad (10)$$

If a static background point is covered by a shadow, we have:

$$R_k(x,y) = \frac{C_A}{C_A + C_p \cos \angle (N(x,y), L)} \qquad (11)$$

This ratio is less than one. In fact, the angle between N(x, y) and L is in the range between $-\pi/2$ to $\pi/2$ therefore the Cos function is always positive. Moreover, due to hypothesis 3, we can assume N(x, y) as spatially constant in a neighbourhood of the point, because the background is supposed planar in a neighbourhood .In [3], authors exploit the spatial constancy of N to detect shadows by computing the variance in a neighbourhood of the pixel of the ratio R_k (x, y): a low variance means that assumption 3 holds, then they mark that pixel as “possible shadow”. Moreover, authors use a lot of other techniques in order to exploit all the four assumptions (such as edge detection and gradient calculation). eq.11 can be seen as the ratio between the luminance after and before shadow appears. In a similar way, Davis et al. ([5][8]) define a local assumption on the ratio between shadow and shadowed point luminance. This is based on the hypothesis that shadows darken the covered point, as eq. 11 and the considerations above confirm. This approach has been improved in [6] where the authors state that shadow has similar chromaticity but lower brightness than that of the same pixel in the background image. They base this statement on the notion of the shadow as a semitransparent region in the image, which retains a representation of the underlying surface pattern, texture or colour.

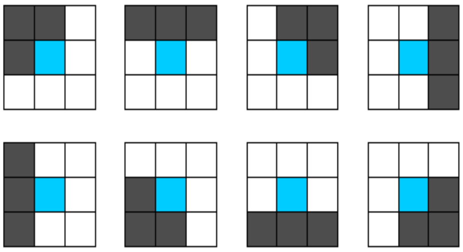
IV. COLOR CLUSTERING

From the result obtained from background subtraction we can see that the resulting contours of moving objects have been drawn roughly. But if we inspect the result carefully, it can be seen that there are at least two kinds of false segmentation lying near the contours of moving objects. The first is that background areas are falsely categorized to moving objects. The second is on the contrary. The reasons behind it may possibly be that the updating rate of background is not fast enough so the background model is not clean enough to extract the moving objects, and it may also be caused by image noise. These false segmentations will certainly degrade the accuracy of further processes, such as objects tracking, and be even worse when objects are close to each other. Moreover, some of these errors, which connecting with moving objects, can’t be eliminated by general processing, such as smoothing, de-noising and erosion dilation based morphologic operations. In order to solve this problem without adding too much computational burden, we proposed a novel colour clustering based method as a post-processing to correct the false segmentations in the initial results Colour based image segmentation is a process of dividing an image into different regions such that each region has homogeneous colour. It is an important operation in many applications of image processing and computer vision, and has been extensively studied [18]. With colour based image segmentation, it can provide relatively complete boundaries of objects. Experiences tell us that in most cases, neighbouring pixels with similar colours should belong to the same objects, but the reverse deduction may not be true. So the goal of segmentation is to split each image into regions that are likely to belong to the same object. These regions or segments should be as precise as possible to distinguish the foreground objects from the background areas. There are many algorithms existing in the literature, for the sake of real-time characteristic, in this paper, we have tried two methods: K-mean algorithm implemented in OpenCV[9] and the method of over-segmentation[19,20]. The main difference between these two methods is the size of segment. The effect of using large segment is that it may straddle more than two objects or between the object and background area. It is undesirable. On the other hand, if the segment is too small, it may not provide sufficient information to distinguish the object from the other object or background. The use of over-segmentation strikes a good balance between providing segments that contain enough information for distinguishing and reducing the risk of a segment spanning multiple objects or over the background and the foreground area.

OVER-SEGMENTATION

The segmentation algorithm has two steps [19].

1. Image is smoothed using a variant of anisotropic diffusion. The purpose of smoothing is to remove image noise.
2. Then, the image is segmented based on neighboring color values.



SAD ALGORITHM

The smoothing algorithm iteratively averages along one of the eight directions as shown in Fig.2 The direction is determined by which direction has the minimum sum-of-absolute-differences (SAD) in color from the center pixel. After smoothing, each pixel is assigned its own segment. Two neighboring 4-connected segments are merged if the Euclidean distance between their average colors varies by less than a threshold (in [19], the value is 66). If the segment is too small, it will be merged with their most similarly colored neighbors. And if the segment is too large, it will also be divided. For more details, please refer to [19]. A result of the over-segmentation algorithm can be seen in Fig. For comparison, images with the averaged color value per segment of K-means and over-segmentation are also illustrated in Figure. Form the results comparisons in Figure 6, we can see that the result of over-segmentation looks more naturally. The reason is that the number of segments in the over-segmentation result is much larger than that of K-means. Though we can enforce the K-means algorithm to cluster more colors, but the price of time-consuming will increase greatly. On the contrary, the over-segmentation algorithm can run very fast, it is important for video surveillance applications.

K-MEANS

K-Means algorithm is an unsupervised clustering algorithm that classifies the input datapoints into multiple classes based on their inherent distance from each other. The algorithm assumes that the data features form a vector space and tries to find natural clustering in them. The points are clustered around centroid which are obtained by minimizing the objective where there are k clusters Si, i = 1; 2; : : : ; k and _i is the centroid or mean point of all the points

As a part of this project, an iterative version of the algorithm was implemented. The algorithm takes a 2 dimensional image as input. Various steps in the algorithm are as follows:

1. Compute the intensity distribution(also called the histogram) of the intensities.
2. Initialize the centroids with k random intensities.
3. Repeat the following steps until the cluster labels of the image does not change anymore.
4. Cluster the points based on distance of their intensities from the centroid intensities.
5. Compute the new centroid for each of the clusters.where k is a parameter of the algorithm (the number of clusters to be found), i iterates over the all the intensities, j iterates over all the centroids and are the centroid intensities.

I. EXPERIMENTAL RESULTS

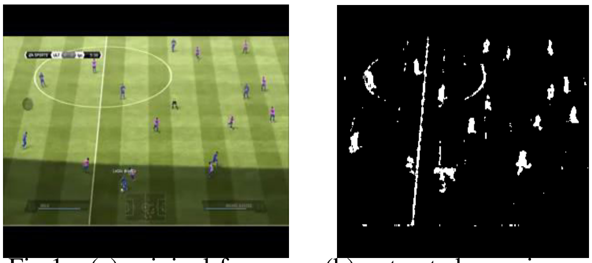


Fig.1 (a) original frame (b) extracted moving regions by mixture of Gaussians

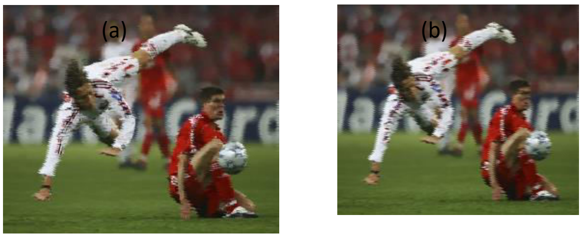


Fig.2 (a) original frame (b) image after edge smoothing

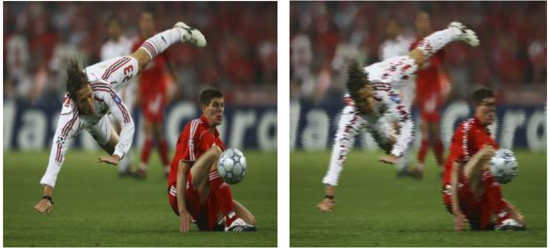


Fig.3 (a) over-segmentation result (threshold 6) (b) over- segmentation result (threshold 66)

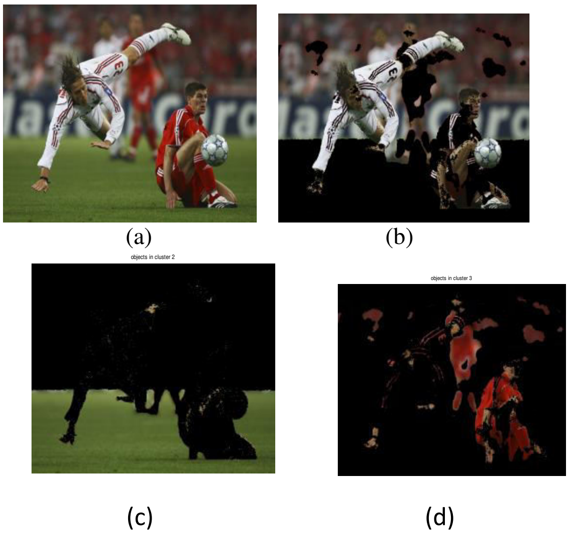


Fig.4 (a)original image (b) pixels in k-means cluster 1 (c) pixels in k-means cluster 2 (d) pixels in k-means cluster 3.

VI. CONCLUSION

Moving objects detection and segmentation is a fundamental step in many applications based on vision. Mixture of Gaussians is the frequently used method to subtracting moving objects from background. But its results are not good enough in some cases. In this paper, a post-processing method is proposed to solve this problem. The results with more complete boundaries provided by the color clustering is used to verify the outputs of mixture of Gaussians, and thus two possible false segmentations can be corrected effectively. Moving shadow suppression and small region filter are also adopted. Using these methods, the results can be greatly improved. Experiments have been done to prove the effectiveness of our work. As a general post-process procedure, the proposed method can also be used for other background subtraction related methods and the results can be used in next step-moving objects tracking.

REFERENCES

[1] C Niu, Y Liu, Moving object segmentation in the H. 264 compressed domain, Lecture Notes in Computer Science,vol 5995,2010:645-654.

[2] W Wang, J Yang, W Gao,Modeling Background and Segmenting Moving Objects from Compressed Video, IEEE Transactions on Circuits and Systems for video technology, Vol. 18, No. 5, 2008:670-681.

[3] Jens Klappstein, Tobi Vaudrey, Clemens Rabe, Andreas Wedel at el,Moving Object Segmentation Using Optical Flow and Depth Information,Lecture Notes in Computer Science, Vol 5414,2009:611-623

[4] Piccardi M. Background subtraction techniques: a review. IEEE International Conference on Systems, Man and Cybernetics, 2004, vol.4: 3099- 3104.

[5] Stauffer C, Grimson W.E.L. Adaptive background mixture models for real-time tracking. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Ft. Collins, 1999: 246-252.

[6] Horprasert T, Harwood D, Davis L S. A Statistical Approach for Real-time Robust Background Subtraction and Shadow Detection. Proceedings of IEEE ICCV' 99 Frame-Rate Workshop, 1999, pp.1-19.

[7] KaewTraKulPong P., Bowden R. An Improved Adaptive Background Mixture Model for Realtime Tracking with Shadow Detection. In

Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems, Sept 2001, Pages:1-5.

[8] Elgammal A., Harwood D., Davis L. Non-parametric Model for Background Subtraction. in Proc. 6th Eur. Conf. Computer Vision, vol. 2, 2000, pp. 751-767.

[9] Dar-Shyang Lee. Effective Gaussian mixture learning for video background subtraction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5):827 - 832.

[10] KaewTraKulPong P., Bowden R. An improved adaptive background mixture model for real-time tracking with shadow detection. in Proceedings of the 2nd European Workshop on Advanced Video-Based Surveillance Systems, Sept. 2001.

[11] http://www.cvg.rdg.ac.uk/PETS2009/a.html

[12] Intel Open Source Computer Vision Library. URL http://www.intel.com/ research/mrl/ research/opencv/.

[13] Gao X., Boult T., Coetzee F., Ramesh V. Error analysis of background adaption. in Proceedings IEEE conference on computer vision and pattern recognition, 2000, vol.1, pp. 503-510.

[14] Power P. W., Schoonees J. A. Understanding background mixture models for foreground segmentation. In Proceedings Image and Vision Computing, 2002, pp:267-271.

[15] Lee D.S., Hull J., Erol B. A Bayesian framework for gaussian mixture background modeling. in Proceedings of IEEE International Confererence on Image Processing, 2003, pages:973-976

[16] Mittal A., Huttenlocher D. Scene modeling for wide area surveillancd and image synthesis. in Proceedings IEEE conference on computer vision and pattern recognition, 2, pp. 160-167, (Hilton Head Isand, SC), June 2000.

[17] Cucchiara R., Grana C., Piccardi M., Prati A., Sirotti S. Improving shadow suppression in moving object detection with HSV color information. Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE, 25-29 Aug. 2001 Page(s):334 - 339.

[18] Lucchese L., Mitra S. K. Color image segmentation: A state-of-the-art survey. in Proc. Indian National Science Academy(INSA-A), vol. 67, A, New Delhi, India, Mar. 2001, pp. 207–221.

[19] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, Richard Szeliski. Highquality video view interpolation using a layered representation. Proceedings of ACM SIGGRAPH 2004, Pages: 600 – 608.

[20] C. Lawrence Zitnick, Sing Bing Kang. Stereo for Image-Based Rendering using Image Over-Segmentation.International Journal of Computer Vision, 2007, 75(1):49–65.

◈◈◈

Marker Controlled Watershed Segmentation Using Bit-Plane Slicing

M. Sivagami & T. Revathi

School of Computer Science, VIT University, Chennai, India
E-mail : msivagami@vit.ac.in, revathi.theerthagiri@vit.ac.in

***Abstract* -** Image segmentation is the basis for computer vision and object recognition. Watershed transform is one of the common methods used for region based segmentation. The previous watershed methods results in over segmentation. In this paper we present a novel method for efficient image segmentation by using bit-plane slicing and marker-controlled watershed. Bit-Plane slicing method produces the sliced image by taking the most significant bit of the image as the input to the bit-plane slicing algorithm. The output of the Bit-Plane slicing algorithm is used to produce the gradient image .The watershed segmentation algorithm is applied to the average of the marker image and the gradient image so as to get efficient segmentation result. Experimental results, shows that the proposed method reduces the memory consumption and computation**.**

***Keywords* -** *component; Image Segmentation, Marker- controlled Watershed Transform, Bit-Plane Slicing, Multi-scale gradient.*

I. INTRODUCTION

Image segmentation involves partitioning an image into groups of pixels which are homogenous with respect to some criterion. The segmentation is based on measurements taken from the image and might be gray level, color, texture, depth or motion. The purpose of image segmentation is to partition an image into meaningful regions with respect to particular application. Image segmentation is an initial and vital step for overall image understanding. The classical image segmentation is bounded by various aspects like Shadow problem in object variability and noise leading to over segmentation. To overcome these problems, examining the image at multi-resolution level is being considered.

Segmentation algorithms are generally based on one of the two basis properties on intensity values: They are discontinuity and similarity. Discontinuity: To partition an image based on sharp changes in intensity (such as edges). Similarity: To partition an image into regions that is similar according to a set of predefined criteria. Other conventional segmentation are spit-and-merge and morphological method. Among them morphological segmentation is commonly used because they deal with geometric features, such as size, shape, contrast or connectivity. Thus morphological transformations can be considered as object-oriented, and therefore segmentation oriented [7]. Until now, a variety of techniques and algorithms have been proposed for image segmentation. The three main categories of image segmentation are: edge detection, clustering and region extraction [1],[2],[8]. Clustering consists of classifying a homogenous cluster and naming each cluster as different region. Drawback of this method is that, the number of cluster is not known. Edge detection identifies the points in a digital image at which the image brightness changes sharply or more formally, has discontinuities. Discontinuities of edge detection correspond to depth, surface orientation; etc. The purpose of detecting sharp changes in image brightness is to capture important events and changes in the properties of the world. Finally, region extraction groups pixels into a set of regions based on similarity [2].Most segmentation techniques are based on region extraction [2],[3],[4]. Image enhancement is the method to enhance the low contrast image. Disadvantages of this method is that all the total pixels in the image are brightened and so that this method may not be suitable for some applications. Bit- plane slicing (BPS) method is used to solve this problem. Bit-plane slicing is a technique in which the image is sliced at different planes. It bit level ranges from 0(LSB) to 7 (MSB). The input to this method is an 8-bit per pixel [5].We proposed an efficient image segmentation method is based on BPS and Marker controlled watershed segmentation (MCWS) algorithm. In this approach the over segmentation is avoided by marker-controlled watershed (MCW). The computation and along with memory is reduced using BPS.

II. RELATED WORKS

Watershed transformation is a morphological based tool for image segmentation. In grey scale mathematical morphology the watershed transformation for image segmentation is originally proposed by Digabel and Lantuejoul [6],[7]. The watershed transform can be classified as region based segmentation approach. The watershed algorithm is known to be a very fast algorithm for segmenting the images. Regions of the image characterized by small variations in grey levels have small gradient values, so watershed segmentation is applied on the gradient of the image rather than the actual image. The major problem with the watershed segmentation is that it produces over segmentation due to the large number of minima. The drawbacks of normal watershed transform are sensitivity to strong noise and high computation. To overcome these problems, a strategy was proposed by Meyer and Beucher (1990). The strategy is called marker-controlled segmentation [2].The goal of the marker controlled segmentation is to detect the presence of the homogenous regions from the image by a set of morphological operations. Markers are connected components belonging to an image [9] ,[10]. The marker image used for watershed segmentation (WS) is a binary image consisting of either single marker points or larger marker regions, where each connected marker is placed inside an object of interest. Each marker has one-to-one relationship to specific watershed regions. After segmentation the boundaries of the watershed regions are arranged on the desired ridges, thus separating each object from its neighbors. The multi-resolution image is generated by the two-scale Daubechies 4-tap wavelet transform and the markers for the WS algorithm were extracted from a low-resolution image. The flat regions larger than 85 pixels were extracted as markers [2]. Multi-resolution framework watershed segmentation is mainly used to reduce the noise related problems and computation [2].

III. PROPOSED METHODOLOGY

The proposed method is concerned with satellite images. The methods used in this system are BPS and MCWS. BPS slices the image into eight planes and the most significant bit plane is used in this system. The image is segmented by using MCWS, which reduces over segmentation .The proposed algorithm is given below:

- Apply bit-plane slicing to input image.
- Use marker- controlled watershed transform to the sliced image.
- Compute morphological gradient for the sliced image.

$$G(f) = (f \oplus B) - (f \ominus B) \qquad (1)$$

Where $G\ (f)$ = Morphological gradient,

f = given image

B = Structuring element.

- Compute multi-scale morphological gradient.

$$MG\ (f) = 1/n\sum^{n}_{i=1}(G\ (f)\ \ominus B\ ) \qquad (2)$$

- Compute final gradient image is obtained by reconstructing the multi-scale gradient image, with its dilated image as a reference image.

$$FG\ (f) = \Phi_{rec}((MG\ (f) \oplus B)\ \ MG(f)) \qquad (3)$$

∟ ┘

∟

- Extract Markers using top-hat transform and bottom-hat transform.
- The segmented image is obtained by applying the morphological watershed to the average of the marker image and the final gradient image.

IV. STUDY AREA

The study areas used Mumbai city and Rome city images from quick bird satellite images which has the resolution of 215x215 and 216x215 respectively.

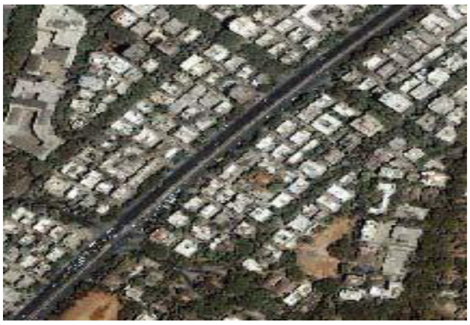


Fig. 1 : Mumbai City Image

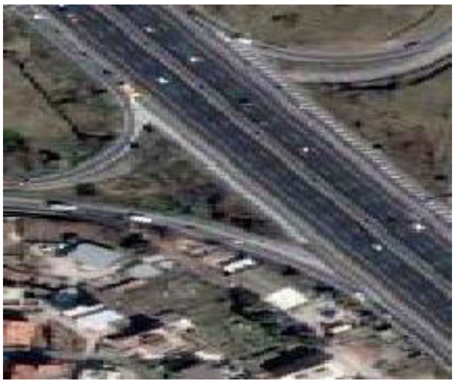


Fig. 2 : Rome City Image

A. *Figures and Tables*

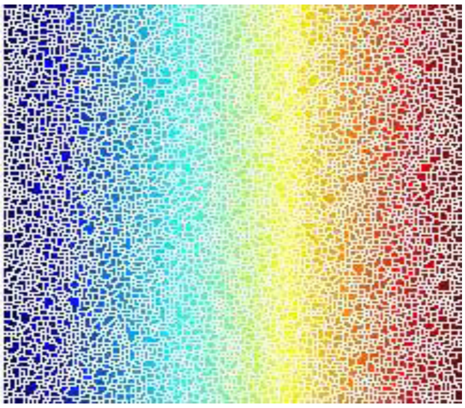


Fig. 3 : Watershed algorithm applied for Mumbai city image.

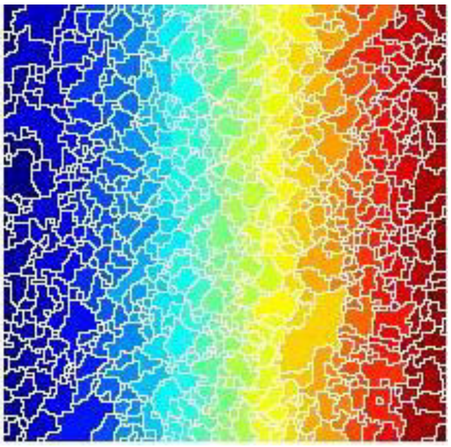


Fig. 4 : Morphological Watershed algorithm applied for Mumbai city image.

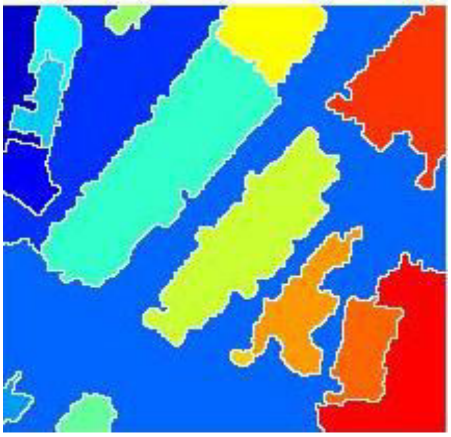


Fig. 5 : Marker-controlled Watershed algorithm applied for Mumbai city image.

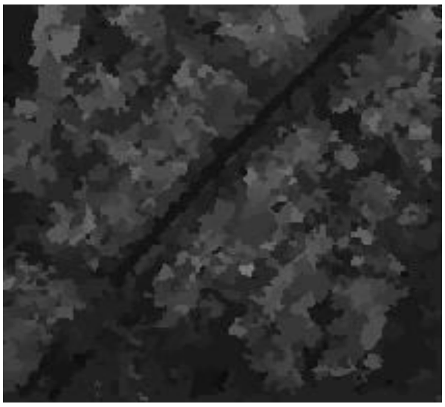


Fig. 6 : Multi-resolution Marker-controlled Watershed algorithm applied for Mumbai city image.

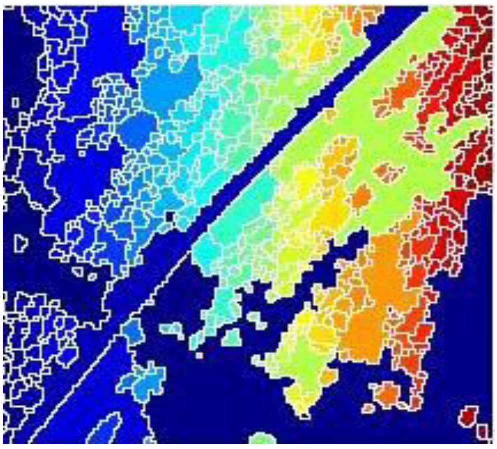


Fig. 7 : Bit-Plane slicing and Marker-controlled Watershed algorithm applied for Mumbai city image.

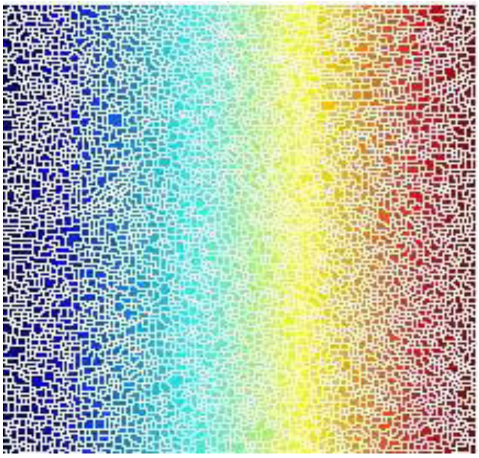


Fig. 8 : Watershed algorithm applied for Rome city image.

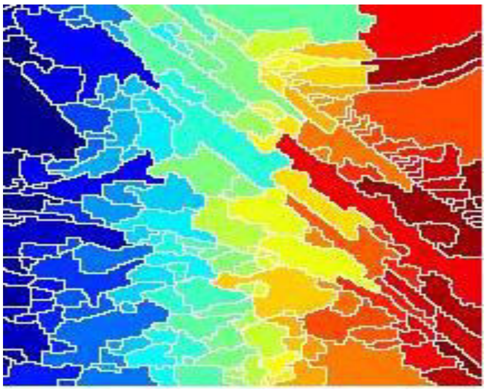


Fig. 9 : Morphological Watershed algorithm applied for Rome city image.

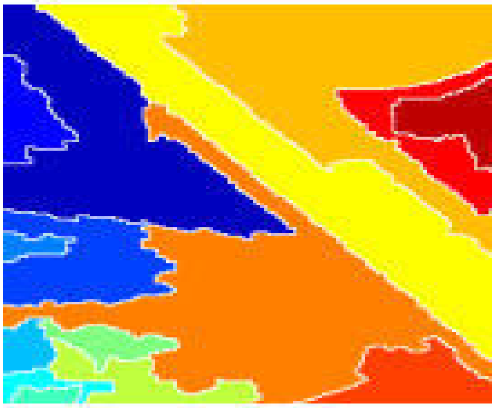


Fig. 10 : Marker-controlled Watershed algorithm applied for Rome city image.

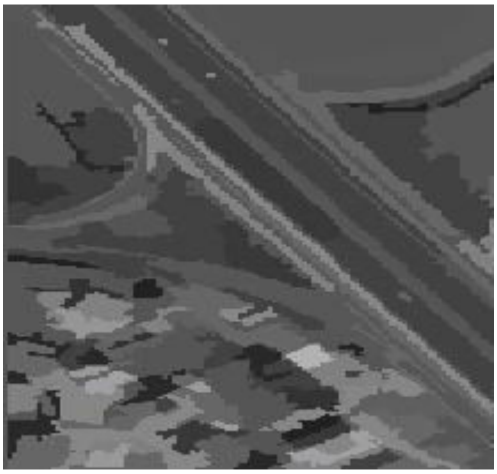


Fig. 11 : Multi-resolution Marker-controlled Watershed algorithm applied for Rome city image.

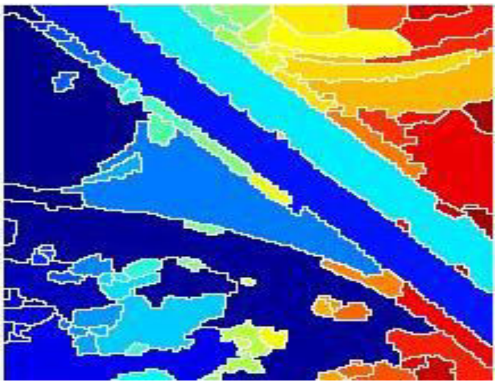


Fig.12 : Bit-Plane slicing and Marker-controlled Watershed algorithm applied for Rome city image.

The figure3 to figure 6 shows the previous watershed method results of Mumbai city image. The figure7 shows the proposed method result for Mumbai city image. The figure8 to figure 11shows the previous watershed method results of Rome city image. The figure12 shows the proposed method result for Rome city image. Comparing with the previous watershed methods, the proposed method gives the good segmentation result.

Evaluation of segmentation results	Number of Segments	Elapsed Time (second)	PSNR	Goodness Value
Watershed (WS)	4041	0.857296	41.2287	6.1938
Morphological Watershed (MWS)	478	2.025499	41.2964	7.4010
Marker-controlled Watershed(MCWS)	15	3.023527	40.7937	4.0761
Multi-resolution marker controlled watershed (MMCWS)	265	14.634666	30.7023	7.9766
Bitplane Slicing (BPS)with Marker-controlled Watershed(MC	208	7.975092	41.6423	1.4766

WS)				

Table 1: Mumbai City Image Evaluation Results

Evaluation of segmentation results	Number of Segments	Elapsed Time (seconds)	PSNR	Good ness Value
Watershed (WS)	4394	0.576636	41.2287	6.6117
Morphological Watershed (MWS)	942	1.264298	41.2964	7.4010
Marker-controlled Watershed(MCWS)	8	2.800380	40.7937	4.0761
Multi-resolution marker controlled watershed(MMCWS)	832	12.925644	36.6794	1.5775
Bitplane Slicing (BPS)with Marker-controlled Watershed(MCWS)	896	7.253986	41. 7937	1.2176

Table 2: Rome City Image Evaluation Results

Evaluation of segmentation results is done on number of segmented regions, PSNR and Goodness function. PSNR is calculated by the following function:

$$PSNR= 10 * \log10 (256^2 / MSE) \qquad (4)$$

Where MSE is the Mean Squared Error.

The higher the value of PSNR is better. The Goodness function is calculated by

$$F (I) =\sqrt{M}\ x \sum^{n}_{i=1\ (ei)}{}^{2}/\sqrt{A} \qquad (5)$$

Where I is the image to be segmented, M is the number of regions in the segmented image, A is the area or i^{th} region number of pixels and e_i is the sum of the Euclidean distance of the color Vectors between the original image and the segmented image of each pixel in the region. The smaller the value of F gives the good segmentation.

V. CONCLUSION

The proposed system uses bit-plane slicing thus reduces the memory when compared to other segmentation approaches. It takes less execution time when compared with the MMCW and gives the good segmentation result compared with the WS, MWS, MCWS, and MMCWS algorithms .The same work can be extended for real time video processing with minimum execution time by using Multi-threading in a multi-core machine.

ACKNOWLEDGMENT

This work has been funded by the Department of Science and Technology.

REFERENCES

[1] Gonzalez, R.C., Woods, R.E., 2002. Digital Image Processing, 2nd ed. Prentice-Hall, Reading, NJ,USA.

[2] A Rizvi, B K Mohan, P R Bhatia, “Probabilistic Multi-Resolution Segmentation of High – Resolution Remotely Sensed Imagery Using Marker-Controlled Watershed Transform”, ICWET 2011, Mumbai, India.

[3] Bhandarkar, S.M., Hui, Z., 1999. Image segmentation using evolutionary computation. IEEE Trans. Evolut. Comput. 3(1), 1–21.

[4] Kim, H.J., Kim, E.Y., Kim, J.W., Park, S.H., 1998. MRFmodel based image segmentation using hierarchical distributed genetic algorithm. IEE Electron. Lett. 34 (25),1394–1395

[5] M.Mohammed Sathik,”Feature Extraction on Colored X-ray Images By Bit-Plane Slicing technique”,2010, International Journal of Engineering Science and TechnologyVol. 2(7), 2820-2824.

[6] Digabel, H., and Lantuejoul, C. Iterative Algorithms, Actesdu Second Symposium European d’Analyse Quantitative desMicrostructuresenSciences des Materiaux, Biologie etMedecine, Caen, 4-7 October 1977, J.-L. Chermant, Ed., Riederer Verlag, Stuttgart, pp.85-99, 1978.

[7] Lantuejoul, C. La Squelettisation et Son Application AuxMesuresTopologiquesDesMosaiquesPolycristallines. PhDthesis, Ecole des Mines, Paris, 1978.

[8] Rafael C. Gonzalez, Richard E. Woods: Digital Image Processing, Third Edition, Pearson Education, pp. 117 – 119.

[9] Beucher, S., Meyer, F., 1993. The morphological approach to segmentation: the watershed transformation. In: Dougherty, E. (Ed.), Mathematical Morphology in Image Processing. Marcel Dekker, New York.

[10] Meyer, F. and Beucher, S., 1990, "Morphological Segmentation," Journal of Visual Communication and Image Representation, v.11, p. 21–46.

◈◈◈

Eon of implementingamultifaceted cloud based OCR in Apple's compassionate App Store milieu

C. Infant Louis Richards, T.Yuvaraj & J.Sylvester Britto

Dept. of CSE, Jeppiaar Engineering College, Chennai, India
E-mail : richiemdu@gmail.com, yuva_vishnu@yahoo.in, jsbrocks45@gmail.com

***Abstract* -** Cloud Architectures discourse key hitches surrounding large-scale data dispensation. In customary data processing it is grim to get as many machines as an application needs. Second, it is difficult to get the machines when one needs them. Third, it is difficult to dispense and harmonize a large-scale job on different machines, run processes on them, and provision another machine to recover if one machine fails. Fourth, it is difficult to auto scale up and down based on dynamic workloads. Fifth, it is difficult to get rid of all those machines when the job is done. Cloud Architectures solve such difficulties.Optical character recognition of cursive scripts present a number of thought-provokingsnags in both segmentation and recognition processes and this entices many researches in the arena of contraption learning. This paper presents the best approach based on a mishmash of OCR and Cloud Computing to handle with the Apple's prerequisite, to make it available in the app store to design a splendid OCR for outdoor portable documents. The enactment results on a comprehensive database show a high notch of accuracy which meets the requirements of viable use.

I. INTRODUCTION

There are many advances in the computer field, where the two main include Cloud computing and the OCR which can now be performed to be implemented in a better environment. With IT technology development, the platform for people to use software has been changed from single PC platform to multi-platforms such as PC +Web-based+ Cloud Computing + Mobile devices. After 30 years development, OCR software started to adapt to new application requirements. WebOCR also known as OnlineOCR or Web-based OCR service has been a new trend to meet larger volume and larger group of users after 30 years development of the desktop OCR. Internet and broadband technologies have made WebOCR & OnlineOCR practically available to both individual users and enterprise customers.

Since 2000, some major OCR vendors began offering WebOCR & Online software, a number of new entrants companies to seize the opportunity to develop innovative Web-based OCR service, some of which are free of charge services.

The term "cloud" is used as a metaphor for the Internet, based on the cloud drawing used in the past to represent the telephone network, and later to depict the Internet in computer network diagrams as an abstraction of the underlying infrastructure it represents.

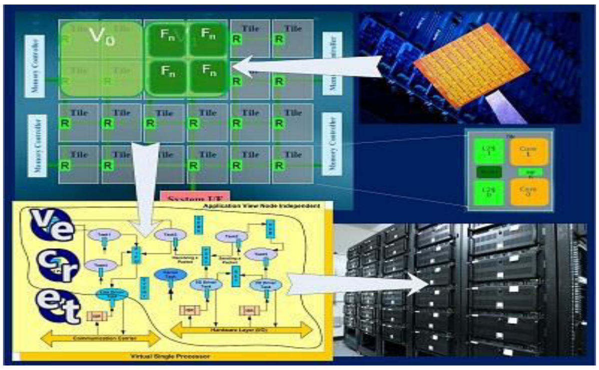


Fig. 1 : OCR Cloud concept

The ubiquitous availability of high capacity networks, low cost computers and storage devices as well as the widespread adoption of virtualization, service-oriented architecture, autonomic, and utility computing have led to a tremendous growth in cloud computing Details are abstracted from end-users, who no longer have need for expertise in, or control over, the technology infrastructure "in the cloud" that supports them.

Almost all the modern-day characteristics of cloud computing, the comparison to the electricity industry and the use of public, private, government, and community forms, were thoroughly explored in Douglas Parkhill's 1966 book, The Challenge of the Computer

Utility. Other scholars have shown that cloud computing's roots go all the way back to the 1950s when scientist Herb Grosch postulated that the entire world would operate on dumb terminals powered by about 15 large data centers.

II. OCR

Optical character recognition (OCR) is the process of converting an image of text, such as a scanned paper document or electronic fax file, into computer-editable text.The text in an image is not editable: the letters/characters are made of tiny dots (pixels) that together form a picture of text.

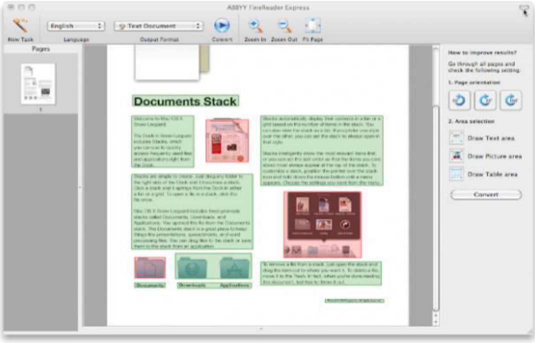


Fig. 2 : Abby fine reader for MAC

During OCR, the software analyzes an image and converts the pictures of the characters to editable text based on the patterns of the pixels in the image. After OCR, you can export the converted text and use it with a variety of word-processing, page layout and spreadsheet applications. OCR also enables screen readers and refreshable Braille displays to read the text contained in images.

In addition to device methods of handling the OCR, there are more concepts added to make it highly portable which is the goal of this paper.

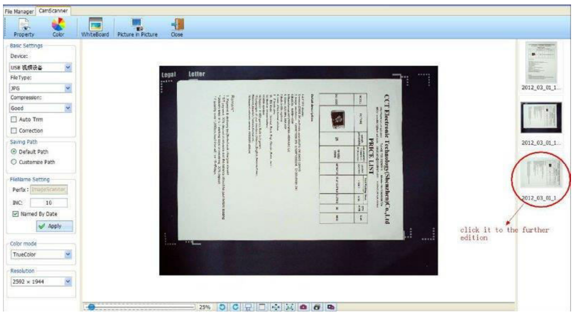


Fig. 3 : OCR portable

The MLP Network implemented for the purpose of this project is composed of 3 layers, one input, one hidden and one output layer. The input layer constitutes of 150 neurons which receive pixels, binary data from a 10x15 symbol pixel matrix. The size of this matrix was decided taking into consideration the average height and width of character image that can be mapped without introducing any significant pixel noise.

The hidden layer constitutes of 250 neurons whose number is decided on the basis of optimal results on a trial and error basis. The output layer is composed of 16 neurons corresponding to the 16-bits of Unicode encoding. To initialize the weights a random function was used to assign an initial random number which lies between two preset integers named ± weight bias. The weight bias is selected from trial and error observation to correspond to average weights for quick convergence.

Implementation can be easily done using the following algorithm.

1. Start at left top of the picture [.bmp]
2. Scan up to image height on the same x-component
   a. If black pixel is detected register x as left of the character, and y as top, Increment x, y
   b. If not continue to the next pixel
3. Scan the image(in the same character space), if y>top , update top
4. If y is equal to height register x as right of character. Increment the number of Characters.
5. Repeat step 1 to 4 till x is equal to image width.
6. Using left, top and right of each character scan character for bottom.

III. iCLOUD

The cloud computing is the latest trend among the IT sector, as it's the most cost effective solution to adapt to larger number of high end technologies and demands. Cloud Storage is basically storing our data in another place so that it can be accessed through various devices. The NIST categorizes cloud computing into three service models: software as a service (SaaS), infrastructure as a service (IaaS) and platform as a service (PaaS).

Fig. 4 : Using iCloud

A key common word here is "service" among the three models, so one of the key issues to consider when negotiating and managing your contract with a cloud provider that will be required to meet your needs. It is important for the contract to include service-level agreements (SLAs) stating specific parameters and minimum levels for each element of the service provided. The SLAs must be enforceable and state specific remedies that apply when they are not met. Aspects of cloud computing services where SLAs may be pertinent include: service availability, performance and response time, error correction time and latency. Such definitions in standard cloud provider contracts often provide a very narrow way of measuring SLA parameters.

iCloud can automatically download new music purchases to all your devices, Which means you can buy a song from iTunes on your iPad at home, and find it waiting for you on your iPhone during your morning commute, all without having to sync. You can access your purchase history from the iTunes Store on your iPhone, iPad, iPod touch, Mac, PC, or Apple TV. And since you already own the songs, albums, or TV shows in your purchase history, you can tap to download them to any of your devices.iCloud manages your Photo Stream efficiently so you don't run out of storage space on your iPhone, iPad, or iPod touch.

IV. APP STORE

The Apple App Store is a digital application distribution platform for iOS developed and maintained by Apple Inc. The service allows users to browse and download applications from the iTunes Store that were developed with the iOS SDK or Mac SDK and published through Apple Inc.

Depending on the application, they are available either for free or at a cost. The applications can be downloaded directly to a target device, or downloaded onto a personal computer (PC) or Macintosh via iTunes. 30% of revenue from the store goes to Apple, and 70% go to the producer of the app.

500 million apps downloaded.
And counting.
There are more than 15,000 apps on the App Store, and so far iPhone users have downloaded an incredible 500 million, in every category from games to business.

Fig. 5 : Apple App Store deeds

The App Store opened on July 10, 2008 via an update to iTunes. On July 11, the iPhone 3G was launched and came pre-loaded with iOS 2.0.1 with App Store support; new iOS 2.0.1 firmware for iPhone and iPod Touch was also made available via iTunes. As of June 6, 2011, there are at least 425,000 third-party apps officially available on the App Store.

As of January 18, 2011, the App Store had over 9.9 billion downloads, which was announced via the company's "10 Billion App Countdown". At 10:26 AM GMT on Saturday, January 22, 2011, the 10 billionth app was downloaded from Apple App Store.

At early July 2011, 200 million iOS users have downloaded over 15 billion apps from its App Store.

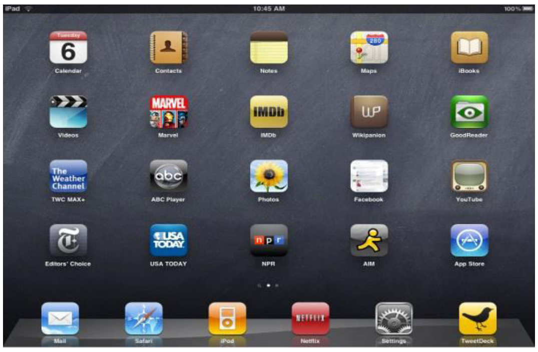


Fig. 6: iOS Screen layout

Above given is the horizontal screen layout of an ipad where the apps are arranged in such a way that the normal users can be able to easily access it and this concept generally takes a good breakthrough for better app marketing.

The term app has become a popular buzzword; in January 2011, app was awarded the honor of being 2010's "Word of the Year" by the American Dialect Society. Apple does not hold a trademark on, or claim exclusive rights to the term app, which has been used as shorthand for "application" since at least 2002, for example Google Apps (first introduced in 2006).

On October 20, 2010, Apple announced the Mac App Store which was eventually launched on January 7, 2011. It is similar to the one for iOS devices, only it has applications designed for Mac computers. The Mac App Store is only accessible by using Mac OS X Snow Leopard or Mac OS X Lion.

The App Store is accessible from the iPhone, iPod Touch and iPad via an iOS application by the same name. It is also the only way to directly download native applications onto an iOS device without jail breaking the device.

Web applications can be installed on these devices, bypassing the App Store entirely, but they tend to have less functionality. The store is also accessible through iTunes, and then on any operating system for which iTunes is provided (Mac OS X and Windows).

V. CLOUD BASED OCR

iCloud automatically backs it up daily over Wi-Fi when your device is connected to a power source. Once you plug it in, everything is backed up quickly and efficiently. That's because Backup is like everything else in iCloud: convenient and completely effortless.

When you set up a new iOS device or need to restore the information on one you already have, iCloud Backup does the heavy lifting. Just connect your device to Wi-Fi and enter your Apple ID and password. Your personal data — along with your purchased music, TV shows, apps, and books from iTunes — will appear on your device.

The iCloud updates them with your most recent appointments — saving you time for all the other things you have going on. You can also share calendars with other iCloud users. A datebook your whole family can add to. Or a team schedule that every player can access. As soon as someone adds or edits an event, iCloud updates it wirelessly on everyone's devices.

Fig. 7: Cloud storage implementation

It's clear that the expansion of the cloud will be both as exciting as it is scary — just like every other computing advancement has been. But since we are moving on slowing into the era of cloud, we hereby design an OCR which connects with the application in the iPhone, iPad to the cloud storage server so that the user can just start using the application to handle the characters. Also this clearly means a very light application by which we can easily start using with less energy consumption which will fall under the category of "Green IT".

On-line character recognition is sometimes confused with Optical Character Recognition (see Handwriting recognition). OCR is an instance of off-line character recognition, where the system recognizes the fixed static shape of the character, while on-line character recognition instead recognizes the dynamic motion during handwriting. For example, on-line recognition, such as that used for gestures in the Pen point OS or the Tablet PC can tell whether a horizontal mark was drawn right-to-left, or left-to-right.

On-line character recognition is also referred to by other terms such as dynamic character recognition, real-time character recognition, and Intelligent Character Recognition or ICR.

It is necessary to understand that OCR technology is a basic technology also used in advanced scanning applications. Due to this, an advanced scanning solution can be unique and patented and not easily copied despite being based on this basic OCR technology.

VI. CONCEPT MIXING – OCR, CLOUD, APPLE

On-line systems for recognizing hand-printed text on the fly have become well known as commercial products in recent years (see Tablet PC history). Among these are the input devices for personal digital assistants such as those running Palm OS. The Apple Newton pioneered this product.

OCR systems require calibration to read a specific font; early versions needed to be programmed with images of each character, and worked on one font at a time. "Intelligent" systems with a high degree of recognition accuracy for most fonts are now common. Some systems are capable of reproducing formatted output that closely approximates the original scanned page including images, columns and other non-textual components.

As it is, the cloud isn't a one-stop solution for storing you data, just as anyone who keeps everything on one storage drive and one drive only is rolling the dice.

iOS 5.1 and its features were announced on June 6 at the WWDC 2011 keynote address. The update was released at 6pm GMT on October 12, 2011. iOS 5 introduced the iCloud service and the Notification Center, as well as improvements to native apps such as Camera. The operating system also features new applications, such as the "Reminders" app and "Newsstand", a special home screen folder and App Store category that contain newspaper and magazine apps.

VII. CONSTRAINTS IN APPLE

Some iTunes Products, including but not limited to Content rentals, may be downloaded only once and cannot be replaced if lost for any reason. It is our responsibility not to lose, destroy, or damage iTunes Products once downloaded, and we may wish to back them up.

The delivery of iTunes Products does not transfer to you any commercial or promotional use rights in the iTunes Products. Any burning or exporting capabilities are solely an accommodation to you and shall not constitute a grant, waiver, or other limitation of any rights of the copyright owners in any content embodied in any iTunes Product.

Apple has the right, but not the obligation, to monitor any materials submitted by us or otherwise available on the iTunes Service, to investigate any reported or apparent violation of this Agreement, and to take any action that Apple in its sole discretion deems appropriate, including, without limitation, termination hereunder or under Apple's Copyright Policy

As an Account holder of the iTunes Service in good standing, you may be provided with limited access to download certain album cover art for music stored in the iTunes Library of your iTunes application. Such access is provided as an accommodation only, and Apple does not warrant, and will not have any liability or responsibility for, such album cover art or we use it.

VIII. CONCLUSION & FUTURE WORK

In this work we have presented a simple but effective solution to use the OCR application in Cloud installed in Apple mobile devices. So hereby we generated an application so that it can easily use the data server in the cloud to implement the users with particular services.

Apple though supports iCloud storage through which it is possible to trace out the character in iPhone and iPad, still further advancements are done in the application design to make it more effective

Fig. 8 : POWR in Cloud [Virtualization]

Finally, this must be transformed for our future work where the Music OCR construction work is done thereby handling it efficiently for the vision impaired users. Also the user interface is to be designed for the application in such a way to support, and last one is all about moving to make the support of POWR app based on cloud.

REFERENCES

[1] http://deepdyve.com/2012/09/install-ios-registeredusers-106-login-amd-pc-tosh/OSX86/minios

[2] "Gartner highlights key predictions for it organizations and users in2010 and beyond." http://www.gartner.com/it/page.jsp?id=1278413

[3] "Flurry: Time spent on mobile apps has surpassed web browsing, http://techcrunch.com/ 2011/ 06/20/flurry-time-spent-on-mobile-appshas-surpassed-web-browsing/, last accessed August 18, 2011.

[4] A. Smith, "Smartphone adoption and usage,"http://www.pewinternet.org/Reports/2011/ Smartphones.aspx

[5] "Lost and found: The challenges of finding your lost or stolen phone," http:// blog.mylookout.com/2011/07/lost-and-found-the-challenges-of-finding-your-lost-or-stolen-phone

[6] V. Zakorzhevsky, "Monthly malware statistics, march2011,"http://www.securelist.com/en/analysis/204792170/Monthly Malware

[7] C. Eric, "The motivations of recent android malware."http://www.symantec.com/content/en/us/enterprise/media/securityesponse/whitepapers

[8] M. Ongtang, S. McLaughlin, W. Enck, and P. McDaniel, "Semanticallyrich application-centric security in android," in Proceedings of the 2009

[9] http://finereader.abbyy.com/about_ocr/ whatis_ocr/

[10] www.aimglobal.org/technologies/ othertechnologies/ocr.pdf

[11] www.simpleocr.com/OCR_Software/ Optical_Character_Recognition/

[12] http://developer.apple.com/library/ios/ documentation/iphone/conceptual/iphoneosprogrammingguide/iphoneappprogrammingguide.pdf

◈◈◈

Mitigating Packet Dropping Misbehaviour in Multicast MANET Environment

K. Praneeth Reddy & Padam Kumar

Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee
Roorkee-247667, India
E-mail : praneeth.rk53@gmail.com, padamfec@iitr.ernet.in

***Abstract* -** Mobile Ad hoc Networks (MANETs) are taking a prominent position in both military and commercial applications. Multicast network are finding even more pragmatic applicability. In MANETs, the packets are routed in such a way that the nodes usually cooperate and forward each other's packets. However, some nodes may not do so in hostile environment by disrupting normal communication. We thus face an acute problem in poor performance because of factors like low battery power, high speed of nodes, poor signal strength, etc. The characteristics features of these factors make the prevention techniques based on trust factor of nodes unreliable. The problem is caused due to breakage of links, formed by the routing protocol, when nodes start acting selfishly either intentionally or unintentionally. We aim to tackle this packet dropping misbehavior caused by factors apart from the most common network layer attacks. Hence, to thwart the unwanted effects of the network layer attacks, we propose an efficient algorithm to improve the overall network performance. We affirm the efficiency of our proposed algorithm with simulation based results. Simulation is done on the mesh based multicast routing protocol, On Demand Multicast Routing Protocol (ODMRP), using PSPP, MATLAB and QualNet simulators.

***Keywords* –** *MANET, Packet drop, Low battery, Poor signal strenght, Node mobility, Network layer attacks, Reliable routing.*

I. INTRODUCTION

A mobile ad hoc network is self-configuring system of mobile nodes that communicate with each other without any fixed infrastructure by establishing a network on the fly [1]. Due to its ability to rapidly deploy, MANET can be used in a number of applications such as emergency scenarios, relief operations, public meeting, battlefield communication, etc.

Compared to tree-based protocols, mesh-based protocols are more robust and suitable for systems with frequently changing topology such as MANETs. In our study, we used the On-Demand Multicast Routing Protocol (ODMRP) [2], a mesh-based routing protocol, due to its simple implementation and high packet delivery ratio.

Security [3] [4] has been an active research topic in wired networks, but in MANETs, its unique characteristics put forth entirely new nontrivial challenges to security design which makes things complex. Some of these challenges are open network architecture, highly dynamic network topology, and stringent resource constraints and shared wireless medium [5]. These challenges also cause packets to be dropped: A behaviour that is similar to the network layer attacks. The network layer attacks can be tackled by using a trust mechanism but packet drop caused by others factors cannot be handled by the trust based routing protocols.

Security is the primary challenge to ad hoc wireless networks because of its lack of centralized infrastructure, stringent resource constraints, dynamic topology changes, high node mobility, poor signal strength, etc. The security issue in MANET for group communication [6] is even more challenging because of the involvement of multiple senders and multiple receivers [7] [8]. Not many researchers have focused on addressing security issues of multicast routing algorithm against packet drop is a single unified manner.

II. RELATED WORK

A. ODMRP

In ODMRP, mesh is established by using the concept of forwarding group, which is a set of nodes responsible for forwarding multicast data between sender and receiver in shortest delay path. An ODMRP source periodically updates routing tables by flooding network with route refreshment packets, Join Query. Upon receiving a Join Query, an intermediate node stores the ID of the upstream node and then rebroadcasts it. When Join Query packet reaches a multicast receiver, it replies with a Join Reply packet, which contains the multicast source ID, and the corresponding next node ID from which it received the Join Query packet. Join Reply packet is then relayed back towards the multicast source via the reverse path traversed by the Join Query packet. Route discovery phase process in ODMRP broadcasts route request message to find the destination node which makes it vulnerable to network layer attacks [9].

B. Principal Component Analysis (PCA)

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into set of values of uncorrelated variables called principal components. So, a principal component can be defined as a linear combination of optimally-weighted observed variables [10].

The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has as high a variance as possible; that is, accounts for as much of the variability in the data as possible, and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed.

Thus, in PCA one wishes to extract from a set of p variables a reduced set of m components or factors that accounts for most of the variance in the p variables. In other words, we wish to reduce a set of p variables to a set of m underlying super ordinate dimensions. These underlying factors are inferred from the correlations among the p variables.

Multi-Criteria Decision Making (MCDM) is a discipline aimed at supporting decision makers faced with making numerous and sometimes conflicting evaluations. MCDA aims at highlighting these conflicts and deriving a way to come to a compromise in a transparent process. Unlike methods that assume the availability of measurements, measurements in MCDA are derived or interpreted subjectively as indicators of the strength of various preferences. Preferences differ from decision maker to decision maker, so the outcome depends on who is making the decision and what their goals and preferences are. We now discuss the following most widely used MCDM approaches which are discussed in brief in the sections that follow:

- Simple Additive Weighting (SAW)
- Analytical Hierarchy Process (AHP)
- Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)

C. SAW [11]

The overall score is computed by multiplying the comparable utility for each attribute by the importance weight assigned to the attribute and then summing these products over all the attributes. Uni-dimensional utility functions U_i have the form:

$$U_i(x_{ij}) = 100 \times \left[\frac{x_{ij} - c_i}{b_i - c_i}\right]^{r_i}$$

Where,

U_i is the utility function for the attribute i

x_{ij} is the score of attribute i for alternative j

r_i is the risk aversion factor (utility is risk neutral if R=1, risk averse if 0<R<1, risk seeking if R>1)

b_i and c_i are the values of the best and worst outcome respectively, for attribute i.

The multidimensional utility function $U(x_j)$ is the final utility or value of alternative x_j.

$$U(x_j) = \sum_{i=1}^{n} W_i \times U_i(x_{ij})$$

The alternatives can be ranked according to $U(x_j)$.

Drawbacks of SAW: There is no interaction among the attributes, since the preferential independence axiom is required. And also there exists difficulty for the assignment of weights.

D. AHP

The AHP [11] is a structured technique for organizing and analyzing complex decisions. It provides a comprehensive and rational framework for structuring a decision problem, for representing and quantifying its elements, for relating those elements to overall goals, and for evaluating alternative solutions.

Here, both qualitative and quantitative criteria can be compared using informed judgments to derive weights and priorities. First, develop a goal hierarchy by defining a single problem as multiple criteria and again each criterion is divided into multiple criteria. To determine the relative importance of the criteria we can use judgments. Using pair wise comparisons, the relative importance of one criterion over another can be expressed. Then check for consistency of the comparisons and subsequently aggregation of the comparisons. To get the relative ranking of priorities from a pair wise comparison of matrix we find the eigenvectors. It has been proved that Eigen vector solution is the best approach [11].

Drawbacks of AHP: A serious challenge to the AHP is the phenomena of rank reversal (adding an irrelevant alternative may cause a reversal in the ranking at the top). Also, the AHP method fails to satisfy the minimal properties of merging-criteria consistency, which forces the chosen alternative to respond in a coherent way to the addition or deletion of criteria.

E. TOPSIS

In TOPSIS [12], if there are *n* node alternatives to be considered in the selection process, they can be represented in the form of a matrix, NW_i.

Step1: The value for each of the attribute in the matrix is normalized.

Step 2: The normalized decision matrix is formed by normalizing each column to get NW_{norm}.

Step 3: For this purpose, each of the attribute is assigned weight “W_i”, such that $\sum W_i = 1$.

Step 4: The weighted normalized matrix is formed by the product of each column with their respective weights to get $(NW_{W\text{-}norm})_i$.

Step 5: The best and worst value for each variable is found.

Step 6: The measure of separation from best case (S_{BEST}) and worst case (S_{WORST}) is given by values obtained in steps 4 and 5.

Step 7: The ratio of S_{WORST} and the sum of S_{WORST} and S_{BEST} gives the preference level P.

Step 8: The access node with the highest “P” value is selected.

Drawbacks of TOPSIS: No technique has been defined to determine the weights. The complexity of the whole algorithm increases exponentially as the number of variables increase because of the calculation of the Euclidean distance. Also, there is a problem of rank reversal at the top when the least ranked node is removed.

F. Rushing attack

In an on demand routing protocol, whenever source nodes flood the network with the Route Request packets in order to discover the new routes to the destination, each intermediate forwarding node processes the first Route Request Packet from a particular node to suppress the duplicate forwarding. The malicious node then ignores the duplicate packets. An attacker can quickly forward these packets by skipping some of the routing or MAC layer process. The attacker in turn gains the routes for the data transmission. All most all the on-demand routing protocols are prone to the rushing attacks [13].

G. Black hole attack

In black hole attack, an attacker first introduces itself in the forwarding group by implementing rushing attack, and then it simply drops all the packets it receive resulting a poor packet delivery ratio instead of forwarding the data packet to the proper destination [13].

III. THE ALGORITHM

We first list down all the factors that may contribute to the problem of packet dropping and subsequently determine the weights of each of these factors indicating their relative importance. Since measurements in MCDM are derived or interpreted subjectively as indicators of strength of various preferences, this technique involves taking responses by varying a certain factor while the others are not changed in the simulation setup. We aimed at talking numerous and conflicting evaluations into consideration. The rating of our interpretations of the selected factors is on a scale of 1 to 5. From the collected data we can decide the factors which contribute most to the detection of packet dropping based on the technique called principal component analysis (PCA). The key here is that every factor we have given as input contributes in some way or the other. The idea is to find the factors that are least contributing to the detection of black hole and eliminate them from the rest of the process.

From a thorough study of the behaviour pattern of the various factors present in the simulation, we could arrive at the input data needed for PCA and after subjecting the data to PCA we got the variables that contribute much to the detection of black holes i.e. variable which have values greater than 0.7 in the rotated component matrix are relevant and they are further processed. The values of these in the rotated component matrix are multiplied by the respective component Eigen values and then normalized to get the weight matrix required for our algorithm for detection of packet dropping misbehaviour.

The following are the Decision Making Criteria:

Trust: Trust is the variable which says that the node with higher trust level can be trusted.

Signal Strength: It is the measure of signal strength available to the mobile node from the transmitting node of the network.

Battery power: A mobile node having a low battery power starts behaving selfishly by not forwarding the packets.

Mobility: The mobility of a node also affects the routing efficiency significantly mainly due to link breakage. Hence, high mobility causes performance degradation.

Node Density: Higher node densities lead to better packet delivery ratio than an environment with lower node density.

Throughput: The throughput of a receiver defined as the ratio of the number of bits received over the time difference between the first and the last received packets.

Delay: The packet delay gives the measure in milliseconds of the average delay for the packet on the wireless link.

Packet Loss: It is the measure of average packet loss between the mobile node and the previous node or next hop.

Network Load: The load of a network node can also help to speculate the presence of packet drop within the network.

Our algorithm will be loosely based on the TOPSIS model, optimized for improving the efficiency and also defining a method to determine the weights of the criteria present. Now use the above obtained results of PCA in the detection algorithm which is as follows:

Step 1: Obtain the weight vector as described previously and thus get the diagonal weight matrix. Since Eigen vector corresponding to the highest Eigen value is normalized, the resultant weight vector is considered to be optimum.

$$Weight \quad Matrix = \begin{bmatrix} ev_1 & 0 & . & . & 0 & 0 \\ 0 & ev_2 & . & . & 0 & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & ev_{n-1} & 0 \\ 0 & 0 & . & . & 0 & ev_n \end{bmatrix}$$

Step 2: From the nodes within the reach of the mobile terminal obtain the decision matrix. We hereby make an assumption that the nodes in the overlay region broadcast their attribute values to the mobile terminal networks.

$$\mathrm{D} = \begin{bmatrix} A_1 & B_1 & . & . & M_1 \\ A_2 & B_2 & . & . & M_2 \\ . & . & . & . & . \\ . & . & . & . & . \\ A_n & B_n & . & . & M_n \end{bmatrix}$$

Step 3: Now normalize the decision matrix. This has to be done as units of all the attributes under consideration vary.

$$\left(\mathrm{A}_{\mathrm{norm}}\right)_{\mathrm{i}} = \frac{\mathrm{A}_{\mathrm{i}}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{A}_{\mathrm{i}}]^2}}$$

Thus the matrix obtained would be:

$$D_{norm} = \begin{bmatrix} (A_{norm})_1 & (B_{norm})_1 & . & . & (M_{norm})_1 \\ (A_{norm})_2 & (B_{norm})_2 & . & . & (M_{norm})_2 \\ . & . & . & . & . \\ . & . & . & . & . \\ (A_{norm})_n & (B_{norm})_n & . & . & (M_{norm})_n \end{bmatrix}$$

Step 4: Obtain the weighted normalize decision matrix. This can be obtained as a matrix multiplication of weight matrix and the decision matrix. This has been illustrated as follows:

$$WN = D_{norm} \times W$$

And hence the weighted normalized decision matrix will be

$$D_{norm} = \begin{bmatrix} (A_{norm})_1 & (B_{norm})_1 & . & . & (M_{norm})_1 \\ (A_{norm})_2 & (B_{norm})_2 & . & . & (M_{norm})_2 \\ . & . & . & . & . \\ . & . & . & . & . \\ (A_{norm})_n & (B_{norm})_n & . & . & (M_{norm})_n \end{bmatrix}$$

Step 5: Obtain the median vector by finding the median of each column in the normalized weighted decision matrix.

$$M = \begin{bmatrix} M_A & M_B & . & . & . & M_M \end{bmatrix}$$

Step 6: Now, obtain the delta matrix as defined in the following way:

$$\Delta WN = \begin{bmatrix} (A_{wn})_1 - M_A & (B_{wn})_1 - M_B & . \; . \; . & (M_{wn})_1 - M_M \\ (A_{wn})_2 - M_A & (B_{wn})_2 - M_B & . \; . \; . & (M_{wn})_2 - M_M \\ . & . & . \; . \; . & . \\ . & . & . \; . \; . & . \\ . & . & . \; . \; . & . \\ (A_{wn})_n - M_A & (B_{wn})_n - M_B & . \; . \; . & (M_{wn})_n - M_M \end{bmatrix}$$

This is also represented as:

$$\Delta = \begin{bmatrix} (\Delta_A)_1 & (\Delta_B)_1 & . \; . \; . & (\Delta_M)_1 \\ (\Delta_A)_2 & (\Delta_B)_2 & . \; . \; . & (\Delta_M)_2 \\ . & . & . \; . \; . & . \\ . & . & . \; . \; . & . \\ (\Delta_A)_n & (\Delta_B)_n & . \; . \; . & (\Delta_M)_n \end{bmatrix}$$

Note that there are some attributes for which we give a higher preference if attribute value is higher and thus in all such cases, the median value will be subtracted from the attribute values, as we prefer the ideal value to be lower for these attributes. However it is the other way round, where higher preference is given if attribute value is lower, for the remaining attributes where we subtract the attribute values from the median value.

Step 7: Obtained the resultant matrix for consideration of ranks by summing up the values of the row elements. The delta matrix and the resultant matrix are as shown.

$$R = \begin{bmatrix} (\Delta_A)_1 + (\Delta_{\mathbf{B}})_1 + \ldots + (\Delta_M)_1 \\ (\Delta_A)_2 + (\Delta_{\mathbf{B}})_2 + \ldots + (\Delta_M)_2 \\ . \\ . \\ (\Delta_A)_n + (\Delta_{\mathbf{B}})_n + \ldots + (\Delta_M)_n \end{bmatrix}$$

Step 8: The values of resultant matrix arranged in decreasing order give the resultant ranks of the nodes. As we consider that the resulting node which is far above from the combined median will be close to the ideal situation and the one which has the least value will be far below the median and hence will be close to the worse situation. So, all those nodes whose value is above a threshold value will be considered for routing while the other nodes will be eliminated. These threshold values are in turn used for our routing algorithm to prevent the packet dropping misbehaviour. This modified routing protocol is called the Enhanced ODMRP (E-ODMRP).

IV. SIMULATION

The simulation is done on the QualNet network simulator, MATLAB and PSPP.

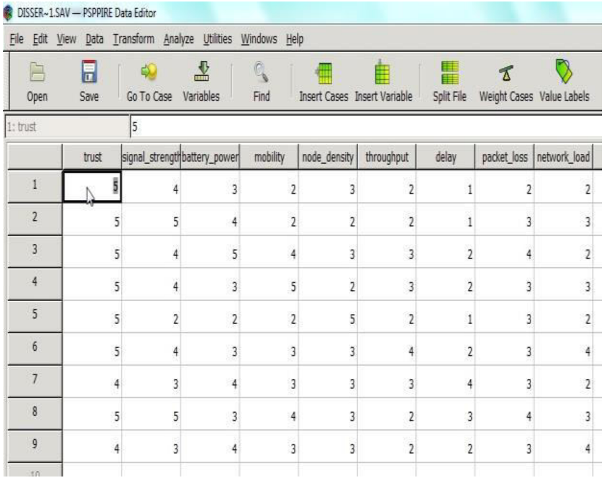
Simulation of Blackhole attack in QualNet

The black hole attack has to be simulated in two phases. In the first phase the rushing attack is carried out to become multicast forwarding nodes as shown in rushing attack. This is done by specifying the malicious nodes at the start of the simulation.

So, we first simulate rushing attacks by making use of the processing delay at every honest node. The honest nodes delay every Join Query for a certain amount of time before broadcasting it. Whereas, all those nodes that are designated as rushing attackers have their Join Query processing delay set to zero thus acquiring the route. Once the Join Query packet is forwarded by an attacker and in turn receives the Join Reply in the path of the attacker, the rushing attack is considered to be successful. But we should remember that that rushing attack is not the only method that gains access to a route in a demand driven routing protocol. Hence, rushing attack is only a passive attack.

The attacker may now establish other attacks such as dropping data packets thus corrupting or illegally accessing confidential data. In second phase, the attacker consumes the packets i.e., drops all packets and doesn't forwards them.

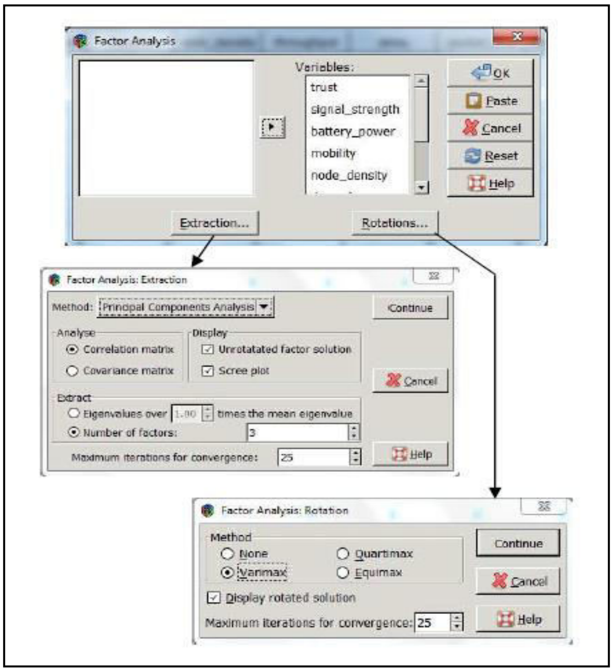
Simulation in SPSS to determine weights of factors


DISSER~1.SAV — PSPPIRE Data Editor
File Edit View Data Transform Analyze Utilities Windows Help
Open
Save
Go To Case
Variables
Find
Insert Cases
Insert Variable
Split File
Weight Cases
Value Labels

1: trust 5

	trust	signal_strength	battery_power	mobility	node_density	throughput	delay	packet_loss	network_load
1	5	4	3	2	3	2	1	2	2
2	5	5	4	2	2	2	1	3	3
3	5	4	5	4	3	3	2	4	2
4	5	4	3	5	2	3	2	3	3
5	5	2	2	2	5	2	1	3	2
6	5	4	3	3	3	4	2	3	4
7	4	3	4	3	3	3	4	3	2
8	5	5	3	4	3	2	3	4	3
9	4	3	4	3	3	2	2	3	4

Fig. 1 : Screenshot of PSPP with data entered


Factor Analysis

Variables: trust, signal_strength, battery_power, mobility, node_density

Factor Analysis: Extraction

Method: Principal Components Analysis

Analyse: ● Correlation matrix ○ Covariance matrix

Display: ☑ Unrotatated factor solution ☑ Scree plot

Extract: ○ Eigenvalues over 1.00 times the mean eigenvalue ● Number of factors: 3

Maximum iterations for convergence: 25

Factor Analysis: Rotation

Method: ○ None ● Varimax ○ Quartimax ○ Equimax

☑ Display rotated solution

Maximum iterations for convergence: 25

Fig. 2: Principal Component Analysis in PSPP

After the rating of parameters is done, as mentioned above, the data is filled in excel sheet and imported to PSPP. The screenshot is shown in Figure 1. First in the *Menu*, in Figure 1, choose *Analyze > Factor Analysis*. Then, as shown in the Figure 2, choose the appropriate *Variables*, *Extraction* and *Rotations*. After PCA, the most contributing factors are those which have above 0.7 in the rotated component matrix. These values are multiplied with the corresponding component Eigen values to give the Eigen vector. This Eigen vector is arranged in the form of a square matrix and is then normalized to give weight matrix required for the proposed MCDM algorithm.

Simulation in MATLAB for proposed detention algorithm

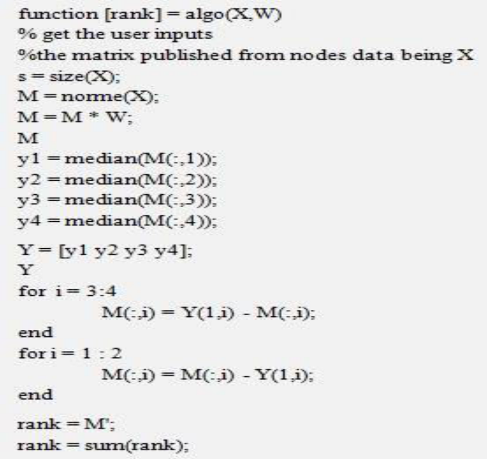

```
function [rank] = algo(X,W)
% get the user inputs
%the matrix published from nodes data being X
s = size(X);
M = norme(X);
M = M * W;
M
y1 = median(M(:,1));
y2 = median(M(:,2));
y3 = median(M(:,3));
y4 = median(M(:,4));
Y = [y1 y2 y3 y4];
Y
for i = 3:4
        M(:,i) = Y(1,i) - M(:,i);
end
for i = 1 : 2
        M(:,i) = M(:,i) - Y(1,i);
end
rank = M';
rank = sum(rank);
```

Fig. 3: Code for Proposed MCDM algorithm

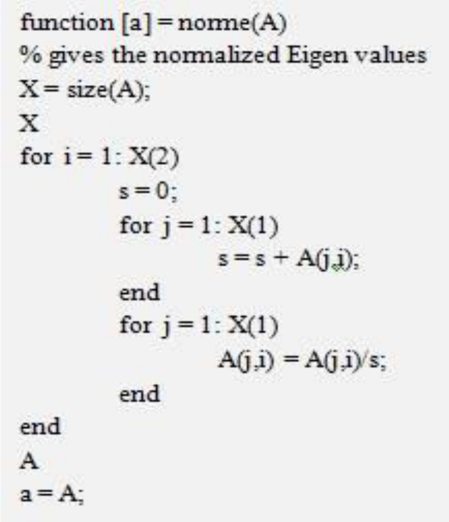

```
function [a] = norme(A)
% gives the normalized Eigen values
X = size(A);
X
for i = 1: X(2)
        s = 0;
        for j = 1: X(1)
                s = s + A(j,i);
        end
        for j = 1: X(1)
                A(j,i) = A(j,i)/s;
        end
end
A
a = A;
```

Fig. 4: Code for Normalization of the Eigen values

The detection algorithm is now simulated in MATLAB, the code of which is shown in Figures 3 and 4. Now, the ranks are given by every node to all its neighbours and the node with the least rank shows high similarity with properties of packet drop behaviour. So, the routing is done only by those nodes whose similarity with black hole node is less.

Table 1 shows the simulation parameters used during the simulation in the QualNet simulator. By malicious nodes we mean that such nodes show black hole like property or packet dropping behaviour.

Parameters	Values Assigned
Routing Protocol	ODMRP
Route refreshment interval	20 sec.
Radio type	802.11b
Channel Capacity	2 Mbits/s
Packet size	512 bytes
Number of packets	300
Number of nodes	24
Mobility model of nodes	Random waypoint
Speed of nodes	1 m/s
Queuing policy	First-in-First-out
Area	1000 m * 1000 m
Simulation time	600 seconds
Number of Malicious Nodes	0 – 6

Table 1: Simulation Parameters

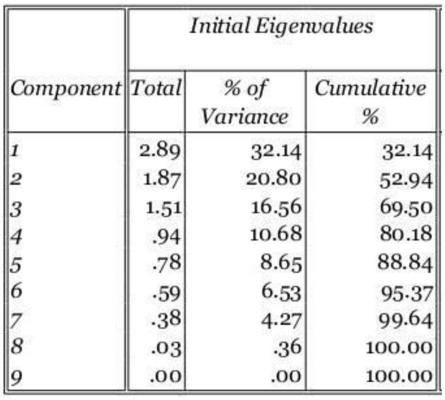
V. EXPERIMENTAL RESULTS

The various performance metrics, which are measured as a function of the number of malicious nodes, used are:

1. *Average Throughput:* The throughput of a receiver is defined as the ratio of the number of bits received

over the time difference between the first and the last received packets. The average throughput is the average of the per-receiver throughputs taken over all the receivers.

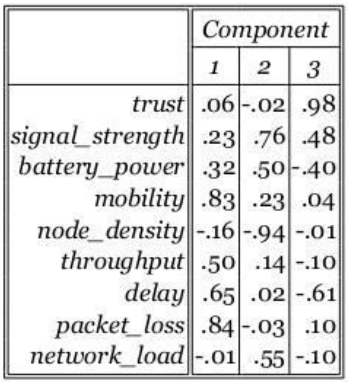
2. *Average Packet Delivery Ratio:* The packet delivery ration (PDR) is defined as the ratio of number of packets reaching the destination to the number of packets sent from the source. The average PDR is the average of the PDR taken over all the receivers.



Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.89	32.14	32.14
2	1.87	20.80	52.94
3	1.51	16.56	69.50
4	.94	10.68	80.18
5	.78	8.65	88.84
6	.59	6.53	95.37
7	.38	4.27	99.64
8	.03	.36	100.00
9	.00	.00	100.00

Fig. 5: Screenshot of Eigen values along with Variance

After subjecting the data to PCA, we get the Eigen values of the corresponding components as shown in Figure 5. From the figure, it is very clear that components 1, 2 and 3 are the most contributing components. So, we choose the number of components, which can represent the original factors approximately correct to be three. Now, the rotated component matrix is as shown in Figure 6. In order to determine those factors that contribute significantly to the detection of packet dropping, we need to select only those factors responsible which have values greater than 0.7 which are *Trust, Signal Strength, Mobility* and *Packet Loss.*



	Component		
	1	2	3
trust	.06	-.02	.98
signal_strength	.23	.76	.48
battery_power	.32	.50	-.40
mobility	.83	.23	.04
node_density	-.16	-.94	-.01
throughput	.50	.14	-.10
delay	.65	.02	-.61
packet_loss	.84	-.03	.10
network_load	-.01	.55	-.10

Fig. 6: Screenshot of Rotated Component Matrix

The highest values of these factors in the rotated component matrix are multiplied with their corresponding component Eigen values to give the following matrix:

$$Eigenvalues = \begin{bmatrix} 0.98\times1.51 \\ 0.76\times1.87 \\ 0.83\times2.89 \\ 0.84\times289 \end{bmatrix} = \begin{bmatrix} 1.480 \\ 1.421 \\ 2.399 \\ 2.428 \end{bmatrix}$$

Now to get the weight matrix, the above matrix is to be normalized and ordered as a diagonal matrix, so that it is easy for further operations, as shown below.

$$WeightMatrix = \begin{matrix} 0.192 & 0 & 0 & 0 \\ 0 & 0.184 & 0 & 0 \\ 0 & 0 & 0.310 & 0 \\ 0 & 0 & 0 & 0.314 \end{matrix}$$

We now take three test cases which can lead to the packet dropping misbehaviour, viz, black hole attack; low battery power; and poor signal strength of nodes. We then compare the performance of ODMRP with the Enhanced ODMRP (E-ODMRP).

The results for the black hole attack along with the proposed solution are shown in Figures 7 and 8. The value of trust is evaluated as shown in [14]. We can see that as number of attackers increase, the average throughput and average PDR decreases as we would expect. But the performance of E-ODRMP has significantly improved over ODMRP.

When a network is under black hole attack, it has been observed that the loss in average throughput is around 47.65% and the decrease in the average PDR is 46.31% when there are six attackers, i.e., 25% of all nodes are attackers, as compared to when there are no attackers.

And when the trust of nodes is defined the loss in average throughput is just 17.69% and the loss in average PDR is just 13.85%. Delay here in E-ODMRP increases significantly because of the overhead incurred in securing a safe routing path.

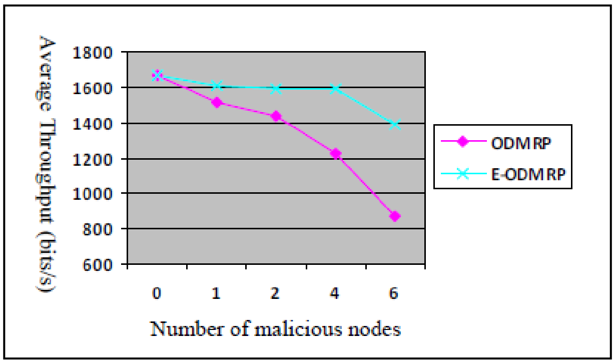


Fig. 7: Average Throughput

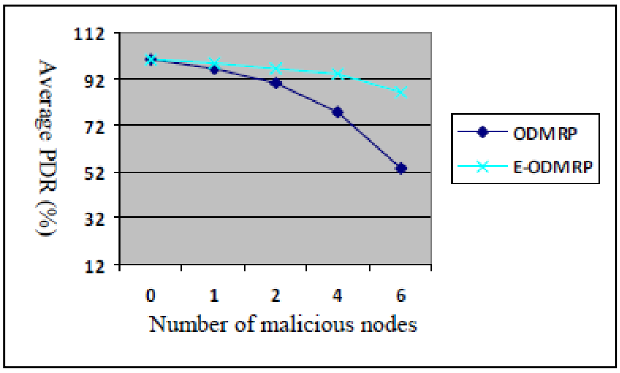


Fig. 8: Average PDR

The results when nodes have low battery power are shown in Figures 9 and 10. We see that, as number of attackers increase, average throughput and average PDR of ODMRP decreases significantly, but E-ODMRP successfully mitigates this negative packet drop effect. In ODMRP, the loss in average throughput is around 6.11% and decrease in the average PDR is 24.9% when there are six attackers as compared to when there are no attackers. For E-ODMRP, the loss in throughput is just 10.53% and decrease in the average PDR is just 9.18%. The decrease in throughput of E-ODMRP is mainly because of either network congestion or routing overhead incurred after implementation of the detection technique in ODMRP.

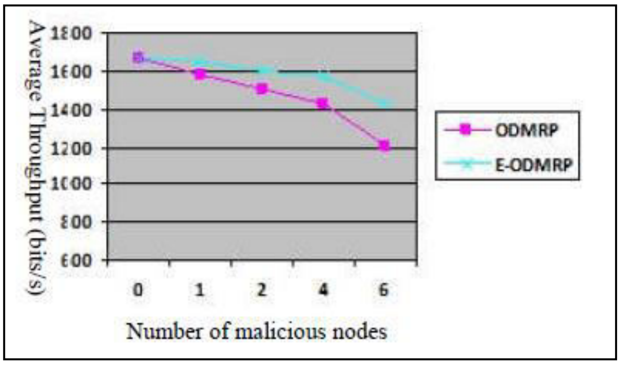


Fig. 9: Average Throughput

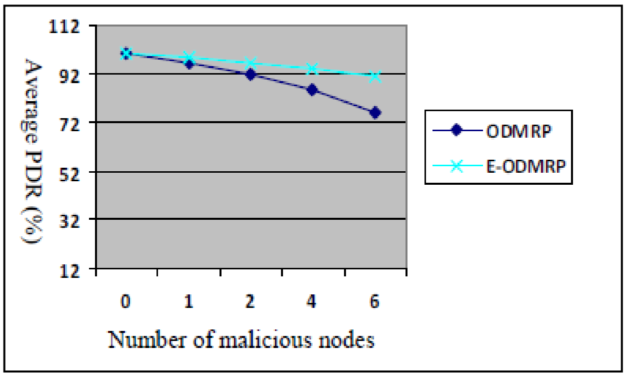


Fig.10: Average PDR

For the case when nodes receive poor signal strength from their neighbours, the results are shown in Figures 11 and 12. We see that, similar to the previous case, as number of attackers increase, the average throughput and average PDR of ODMRP decreases significantly unlike as in E-ODMRP. In ODMRP, the loss in average throughput is around 27.44% and the decrease in the average PDR is 25.8% when there are six attackers as compared to when there are no attackers. And for E-ODMRP, loss in throughput is just 174.32% and decrease in the average PDR is just 11.32%. Here again, decrease in the throughput of E-ODMRP is because of either network congestion or routing overhead incurred after implementation of the proposed detection technique in ODMRP.

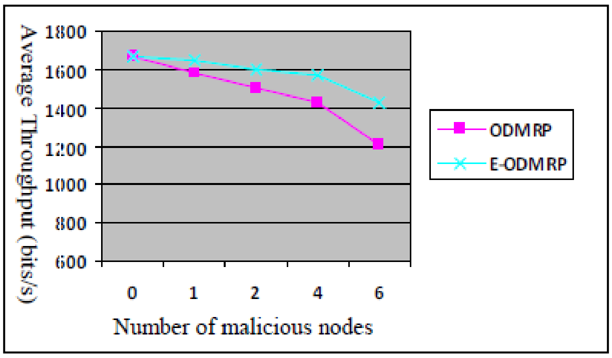


Fig. 11: Average Throughput

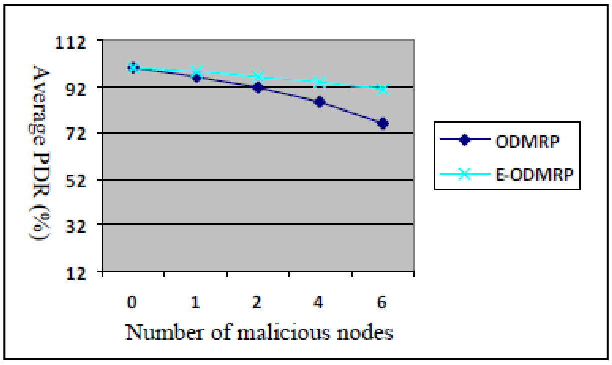


Fig. 12: Average PDR

VI. CONCLUSION

We aimed to determine a method to identify nodes with packet dropping misbehaviour in the multicast MANET environment and thus proposed a system in which anomalies in behavior is defined quantitatively by observing data exchange activity. Since, we have weights apportioned for different parameters based on their relative importance, the solution will tackle all kinds of packet dropping behaviour. Hence, this method is able to detect many kinds of packet drop attacks. The solution involves some computation but the trade off is for effective routing. Also it does not involve huge exchange of data among the nodes, for the process of detection, thus saving the network bandwidth. The detection process is on-the-fly, so even if a node starts behaving unusually at unexpected time the proposed detection algorithm is efficient to handle such an anomaly. Also, there is no need for a central authority for effective routing. Therefore, the detection process is very suitable for multicast routing in MANETs.

REFERENCES

[1] Sunita Sahu and Sushir K. Shadily, “A Comprehensive survey on Intrusion Detection in MANETs,” in International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, July-December 2010, pp. 305-310.

[2] Sung J. Lee, William Su, and Mario Gerla, “On-Demand Multicast Routing Protocol in Multihop Wireless Mobile Networks,” in Mobile Networks and Applications, Kluwer Academic Publishers, December 2002, pp. 441–453.

[3] N. Shanti and L. Ganeshan, "Security in Multicast Mobile Ad Hoc Networks," Proceedings of IJCSNS International journal of Computer Sience and Security, pp.326-330, July, 2008.

[4] D. Djenouri, L. Khelladi, N. Badache, "A Survey of Security Issues in Mobile Ad Hoc and Sensor Networks," IEEE Communication Surveys & Tutorials, vol.7, 4th Quarter ‘05.

[5] Hao Yang, H. Luo, F. Ye, S. Lu, and L. Zhang, "Security in Mobile Ad hoc Networks: Challaenges and Solutions, "IEEE Wireless Communications, 38-47, February 2004.

[6] N. Shanti, L. Ganeshan "Security in Multicast Mobile Ad Hoc Networks" Proceedings of International journal of Computer Science and Network Security, pp. 326-330, July 2008.

[7] M. Gerla, G. Pei, S. J. Lee, C. C. Chiang "On Demand Multicast Routing Protocol for Mobile Ad Hoc Networks" at http://tools.ietf.org/html/draft-ietf-manet-odmrp-00

[8] J. Garcia-Luna-Aceves, E. Madruga, "A Multicast Routing protocols for Ad Hoc Networks"; Proc. Of Infocom '99, pp. 784-792, 1999.

[9] S. Yi, P. Naldurg and R. Kravets, “A Security-Aware Routing Protocol for Wireless Ad Hoc Networks,” in 6th World Multi-Conference on Systemics, Cybernetics and Informatics, Florida, USA, 2002.

[10] Keith F. Widaman, “Common factor analysis versus principal component analysis”, in Multivariate Behavioral Research, Vol 28(3), 1993, pp. 263-311.

[11] F. Naumann, “Data Fusion and Data Quality”, in Proceedings of the New Techniques and Technologies for Statistics Seminar, 1998 (NTTS '98), Sorrent, Italy.

[12] F. Bari and V. Leung. “Multi-Attribute Network Selection by Iterative TOPSIS for Heterogeneous Wireless Access” in Proc. IEEE CCNC’07 Las Vegas, NV, Jan. 2007.

[13] Payal N. Raj and Prashant B. Swadas, “DPRAODV: A Dynamic Learning System Against Blackhole Attack in AODV based MANET,” in IJCSI International Journal of Computer Science Issues, 2009, pp.54-59.

[14] Asad Amir Pirzada, Chris McDonald "Establishing Trust in Pure Ad-hoc Networks" Proceedings of the 27th Australasian Computer Science Conference, Pages 47-54, 2004.

◈◈◈

A Methodology for Test Source Reuse in Virtual System Prototype (VSP)

Bhargava C R[1], Pawankumar B[2], Narayanan S[3], Kariyappa B S[4] & Kamalakar R[5]

[1,2,3&4] Dept of Electronics and Communication, R.V. College of Engineering, Bangalore, India
[5]IFIN ATV C-MODEL, Infineon Technologies India Pvt. Ltd, Bangalore, India
E-mail : bhargavacr@gmail.com[1], bellary.pawan@gmail.com[2], narayanans@rvce.edu.in[3],
kariyappabs@rvce.edu.in[4], Kamalakar.Rachakonda@infineon.com[5]

***Abstract* -** Over the past few years, the Integrated circuits design and verification has become increasingly complex. Industry, to overcome this problem has shifted to Electronic System Level (ESL) design flow. The ESL design at higher level of abstraction is called Virtual System Prototype (VSP). Each Intellectual Property (IP) in VSP should be verified Block level (Standalone verification) and in system level. This paper deals with reusability of the test cases for an IP in module level verification and system level verification. The aim is to reduce the test effort for same test in different verification environment. A common test source is developed for ADC IP and the functionality is verified in the standalone and system level verification environments by reusing the test cases.

***Keywords* -** *Test source, Virtual System Prototype, Standalone Verification, System level Verification, Reuse.*

I. INTRODUCTION

Since the early 1950's the complexity of Integrated circuits have been growing in accordance with the Moore's law. This has exerted a lot of pressure on design engineers, since in semiconductor industry there is a tight time to market constraint. Now, with more than a billion transistors on the chip, the design of such a circuit is called System on Chip (SoC). SoC is integration of different functionality IPs that is pre-verified with the complex high speed bus. To catch up with the time to market constraint, therefore semiconductor companies are moving for ESL design flow, which is higher level of design abstraction called VSP.

VSP is fast software simulation model of a system, which increases productivity [1]. VSP typically can be modeled cycle accurately [4] that runs same compiled and linked target code as real hardware, thus providing the early predictions of the system in architecture level to study the design feasibility. The behavior of the peripherals can also be modeled for multi-core embedded system. Once the architectural model is built, it becomes the executable specification that drives the concurrent development of the detailed Hardware (HW) and Software (SW) implementations. The key advantage is that the model can be easily refined for the HW/SW implementations and also reused easily. With VSP, the verification is often done concurrently with the design development. This saves verification time. ITRS roadmap [2] concludes the reuse rate of the IP cores in SoC and verification should be increased to achieve a 10x design productivity over next 10 years. Many complex designs today are verified with millions of test vectors before tape out and these cover only a small fraction of the potential state-space. With the increase in the chip complexity and reduced design cycle time there is an extensive need for reuse [8]. This encourages early time to market of the product. The paper [1] states one of the key challenges as test source reusability.

SystemC Transaction Level Modeling (TLM)

VSPs are modelled as systemC TLM. The Open SystemC Initiative(OSCI) TLM committee explicitly recognizes the existence of a variety of use cases in TLM, such as SW development, SW performance analysis, architectural analysis, and HW verification. There are several TLM standards [5], for instance, loosely-timed**-** coding style is appropriate for software development, Approximately-time- coding style is appropriate for architectural exploration and performance analysis and many other level of accuracy are mentioned.

VSP based on Transaction Level Modeling (TLM) has become a de-facto standard in today's complex SoC designs [6]. A *Transaction* is defined as the Communication between concurrent processes using function calls. In Register Transfer Level (RTL) the communication happens by the signal protocol, here while communicating between two processes all the signals communicates separately [3]. In Transaction level all the signal values will be communicated in a single function call (transaction) as shown in the fig. 1. So VSP using TLM enables higher simulation speed.

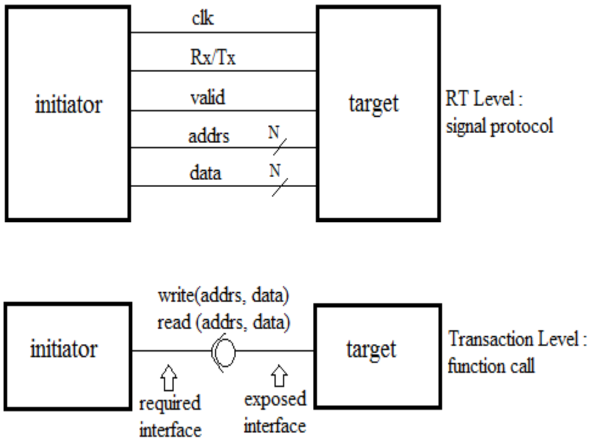


Fig. 1: Communication in TLM and RTL

This paper deals with the methodology to reuse the test cases between the VSP standalone and system level environments. The organization of the paper is as follows; first the register access mechanism in both system level and standalone is explained. Section II discuses on methodology implementation, in which guidelines for common register access , interrupt handling and external signal handling for a Design under test (DUT) are explained. Section III discusses about methodology validation by verifying an ADC IP. The console and VCD output of a sample test case, showing the reuse of test source in both environments.

Register access Mechanism in System level and standalone environments

Each IP contains so many registers with different bit fields and used for different functionalities. The bit fields can be read only, write only, read and writable etc depending on the functionality assigned for that bit field. Each register will have key parameters like *register name, register offset address, register reset value* and *register bit field names*. For example, consider an 8-bit register shown in fig. 2. The name assigned for this register is REG; its reset value is 0x01 and has offset address value of 0x00F0. It also has bit fields BIT1, BIT2 and BIT3. The verification of an IP is mainly done by writing to registers in the IP.

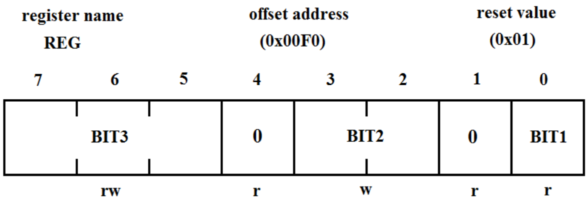


Fig. 2: A typical example register

In system level, every module/IP registers occupies fixed addresses in the memory address space of the system. Each register address will be mapped to its absolute address. An example is

*addr1 = value1; //write access to mapped address – addr1 value2 = *addr2; //read access from mapped address – addr2

Therefore registers and bit fields are accessed (reading or writing) by their names e.g. in fig. 2 the value 0xE0 can be written to register REG by writing REG.U = 0xE0. Here U is a structure/union defined to access entire register value. The bit field BIT2 can be explicitly written by writing REG.B.BIT2 = 0x3. Here BIT2 is member of bit field structure B, which is defined in REG structure/union. When accessing a register in a module, the register absolute address decodes into module/IP offset address and register offset address then transactions will be called.

In standalone IP verification, the memory address space is not available like in system level; therefore the registers are accessed by using their offset address. Suppose value 0xE0 is to be written to the REG register, send write transaction by sending register offset address 0x00F0 and value to be written 0xE0 as the parameters. On successful write the value 0xE0 will be written to register REG. Similarly for reading, send read transaction by sending register offset address and variable to which data should be read as parameters. On successful read the data in the register copied to the variable.

II. METHODOLOGY

A. *Guidelines for common register access*

In order to reuse the test source of an IP while verifying it in system level and standalone environments, same register access mechanism must be used. To achieve this, a separate test bench register stub file for the IP (DUT) is developed in standalone environment to have same register accessing like in system level environment. The steps followed for implementation are as below

- *Define a structure for accessing entire register value:*

This structure say U_TYPE is used when accessing entire value of a register. This structure contains member variables *value* to hold data of register and *register offset*. The constructor of the structure initializes member *register offset* with the passed offset. The structure also contains member functions read and write for sending read/write transactions to a register offset. In system level writing to a register uses '=' operator to write a value into the register and for reading just uses register name in printf/cout statements. To achieve this in standalone, operator overload concept is used. i.e. '=' operator overloaded for writing and no operator is overloaded for reading. So when we use '=' operator with this type structure, the Right Hand Side (RHS) value will be written to register by calling write transaction function. Similarly for read operation no operator is used in overload function. When name of a register without any operator is used read transaction function will be called.

- *Define a structure for accessing bit fields of register:*

This structure say BIT_TYPE is used when accessing bit field values of a register. This structure contains member variables *bit offset*, *bit width* and *bit value*. The constructor of the structure initializes members *bit offset* and *bit width* with the passed values. While reading a specific bit content the entire register value is read into a variable. This value is right shifted by *bit offset* times and value from Least Significant Bit (LSB) to *bit width* is returned by masking rest of the bit fields. In write function the value to be written is left shifted by *bit offset* times and other bit fields are masked except from *bit offset* to *bit offset* + *bit width*. This value is written into register by sending write transaction. Like in U_TYPE structure here also operator overloading is done for read/write operations.

- *Define bit field name structure for each register in the IP*

A separate bit field name structure for each register of the module/IP should be defined. In this structure all the bit field names of the register are declared of BIT_TYPE. The bit offset and bit width for each bit field is initialized in the constructor.

- *Define register name structure for each register in the IP:*

This structure contains U_TYPE U and BIT_TYPE B member declarations. When register name fallowed by .U is used entire register will be accessed and when register name followed by .B.X is used bit field X of the register is accessed. The constructor of this structure takes offset address of register as parameter and initializes it to member *register offset* in U_TYPE structure.

- *Declare all register names with their offset address:*

Once the structure types for all registers are defined, declare each register of the module/IP with their names and offset address according to the specification of that IP. All registers should be made global variables so that they can be accessed directly in any test case file.

B. Interrupt Handling

Every module/IP contains many service request/interrupt lines. In any verification environment when an interrupt occurs corresponding registered Interrupt Service Routine (ISR) or Interrupt handler function should be executed. Before calling a Handler function its address must be registered.

- *Interrupt registering:*

Consider a system/microcontroller with total of 256 interrupts from all the modules then Interrupt Vector Table (IVT) as shown in fig. 3 should have size of 256 to hold all interrupt vector addresses. In system level interrupts from all modules connected to Interrupt Controller, which handles the interrupts. Each ISR address should be registered in the IVT before calling the ISR. In standalone environment interrupt registering is achieved by developing appropriate functions and IVT in test bench base file. The functions are used to register an ISR for a particular priority. This base file will be inherited while standalone verification of each IP.

While verifying a module, the handler function and interrupt priority for a particular interrupt will be the same in both environments. The registering mechanism is different. In system level interrupt handler addresses for each interrupt of module are registered in IVT of the system memory space using API function and in standalone, already developed interrupt registering function in the test bench is used for registering a handler with interrupt priority.

- *Interrupt execution:*

In system level the Interrupt Controller executes the interrupt with the highest priority when multiple interrupts request servicing at the same time. To handle the interrupt execution in standalone environment a SystemC process is used, which is made sensitive to interrupt lines of the DUT. Whenever an interrupt is raised from DUT, this process will be triggered which evaluates the interrupts from highest priority. After polling the interrupts the registered ISR of the raised interrupt line with highest priority is called by passing its priority. Hence while interrupt verification of a DUT same interrupt handler will be executed in both the environments.

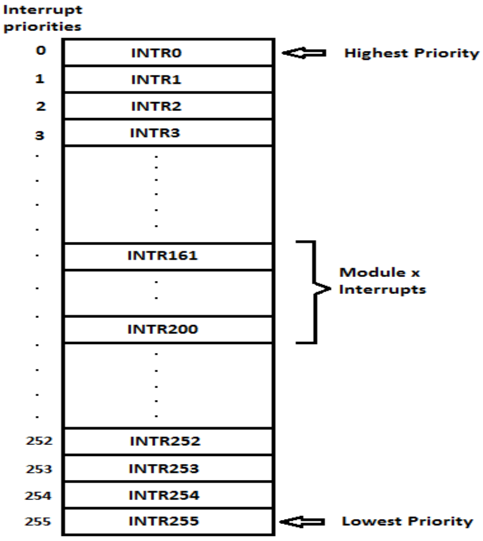


Fig. 3: Interrupt vector table

C. *Handling external signals*

Every DUT contains some input ports which are used for different functionalities. In system level the external input ports of DUT are connected to other modules, so to trigger a DUT line, the module to which the DUT line is connected must be driven. In standalone all input ports of the DUT are connected to test bench output ports by using signals. Here all ports of the DUT are directly accessible from the test bench ports. To trigger a DUT line, we can directly drive the connected test bench port.

III. METHODOLOGY VALIDATION

The methodology is validated by verifying an ADC IP at the system level and standalone environments. The test cases are developed once and ported across two environments. The test sources for verifying various functionalities of the IP are reused in both environments. It is found that at system level the execution of the test case is slower than in the standalone platform. So the test flow should be such that it supports in both the system and standalone environments.

An output of a test case to test one of the functionalities in the IP is shown in the figure. Same response is obtained for the test case while verifying at the standalone and the system level environments. The interrupt functionality for the DUT is enabled in the test file, the interrupt lines are triggered when the results are updated and executes their registered interrupt handler function. The test case is run for 4 results and hence four interrupts are executed, which is evident from the VCD output waveform fig. 4.a and 4.b. The test case console output is also shown in fig. 5.a and 5.b which shows same response in both the environments. Same response is obtained while verifying the different functionalities in the ADC IP.

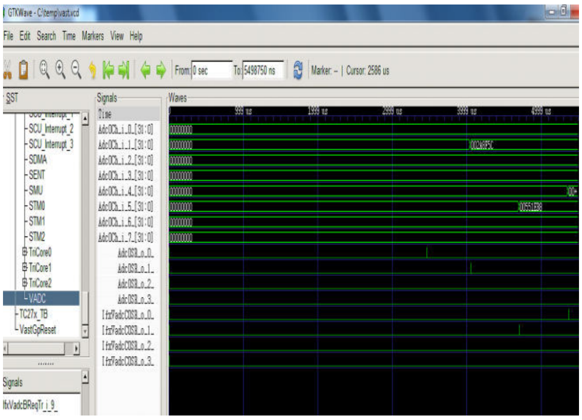


Fig 4.a. System level VCD output

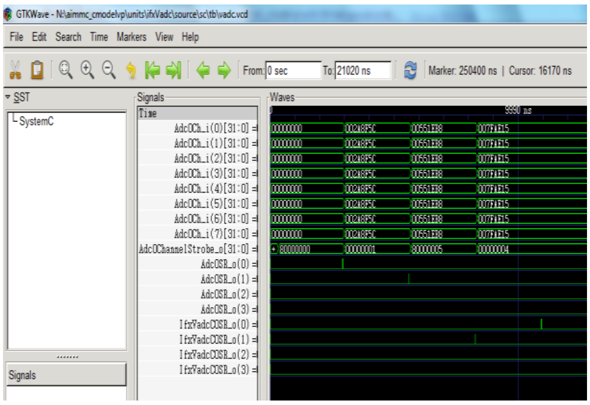
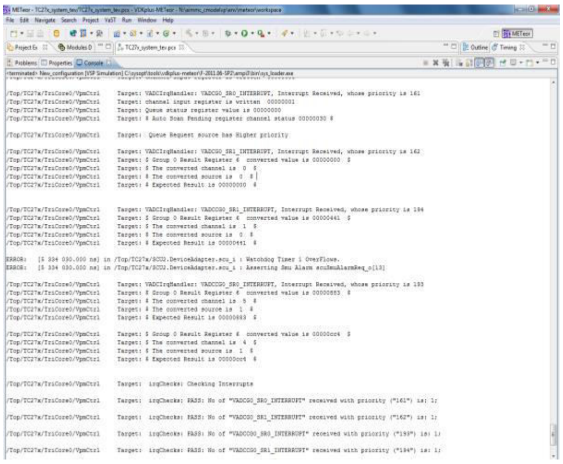
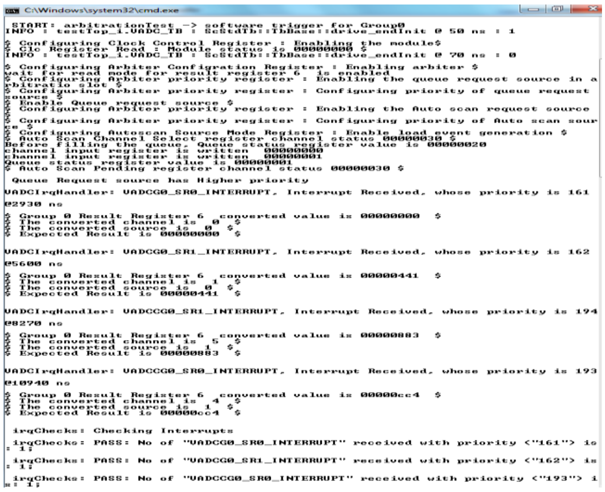


Fig 4.b. Standalone VCD output


METeor - TC27x_system_tev/TC27x_system_tev.pcx - VDKplus-METeor - N:\aimmc_cmodel\vp\am\meteor\workspace
File Edit Navigate Search Project VaST Run Window Help
Project Ex
Modules 0
TC27x_system_tev.pcx
Outline
Timing
Problems
Properties
Console

```
<terminated> New_configuration [VSP Simulation] C:\sysopt\tools\vdkplus-meteor\F-2011.06-SP2\amspd\bin\sys_loader.exe

/Top/TC27x/TriCore0/VpmCtrl     Target: VADCIrqHandler: VADCG0_SR0_INTERRUPT, Interrupt Received, whose priority is 161
/Top/TC27x/TriCore0/VpmCtrl     Target: channel input register is written  00000001
/Top/TC27x/TriCore0/VpmCtrl     Target: Queue status register value is 00000000
/Top/TC27x/TriCore0/VpmCtrl     Target: # Auto Scan Pending register channel status 00000030 #

/Top/TC27x/TriCore0/VpmCtrl     Target:  Queue Request source has Higher priority

/Top/TC27x/TriCore0/VpmCtrl     Target: VADCIrqHandler: VADCG0_SR1_INTERRUPT, Interrupt Received, whose priority is 162
/Top/TC27x/TriCore0/VpmCtrl     Target: $ Group 0 Result Register 6  converted value is 00000000  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ The converted channel is  0  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ The converted source is  0  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ Expected Result is 00000000  $


/Top/TC27x/TriCore0/VpmCtrl     Target: VADCIrqHandler: VADCCG0_SR1_INTERRUPT, Interrupt Received, whose priority is 194
/Top/TC27x/TriCore0/VpmCtrl     Target: $ Group 0 Result Register 6  converted value is 00000441  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ The converted channel is  1  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ The converted source is  0  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ Expected Result is 00000441  $

ERROR:   [5 334 030.000 ns] in /Top/TC27x/SCU2.DeviceAdapter.scu_i : Watchdog Timer 1 OverFlows.
ERROR:   [5 334 030.000 ns] in /Top/TC27x/SCU2.DeviceAdapter.scu_i : Asserting Smu Alarm scuSmuAlarmReq_o[13]

/Top/TC27x/TriCore0/VpmCtrl     Target: VADCIrqHandler: VADCCG0_SR0_INTERRUPT, Interrupt Received, whose priority is 193
/Top/TC27x/TriCore0/VpmCtrl     Target: $ Group 0 Result Register 6  converted value is 00000883  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ The converted channel is  5  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ The converted source is  1  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ Expected Result is 00000883  $

/Top/TC27x/TriCore0/VpmCtrl     Target: $ Group 0 Result Register 6  converted value is 00000cc4  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ The converted channel is  4  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ The converted source is  1  $
/Top/TC27x/TriCore0/VpmCtrl     Target: $ Expected Result is 00000cc4  $


/Top/TC27x/TriCore0/VpmCtrl     Target:  irqChecks: Checking Interrupts

/Top/TC27x/TriCore0/VpmCtrl     Target:  irqChecks: PASS: No of "VADCG0_SR0_INTERRUPT" received with priority ("161") is: 1;

/Top/TC27x/TriCore0/VpmCtrl     Target:  irqChecks: PASS: No of "VADCG0_SR1_INTERRUPT" received with priority ("162") is: 1;

/Top/TC27x/TriCore0/VpmCtrl     Target:  irqChecks: PASS: No of "VADCCG0_SR0_INTERRUPT" received with priority ("193") is: 1;

/Top/TC27x/TriCore0/VpmCtrl     Target:  irqChecks: PASS: No of "VADCCG0_SR1_INTERRUPT" received with priority ("194") is: 1;
```

Fig 5.a. System level console output



```
 START: arbitrationTest -> software trigger for Group0
INFO : testTop_i.VADC_TB : ScStdTb::TbBase::drive_endInit @ 50 ns : 1

$ Configuring Clock Control Register : Enabling the module$
$ Clc Register Read : Module status is 00000000 $
INFO : testTop_i.VADC_TB : ScStdTb::TbBase::drive_endInit @ 70 ns : 0

$ Configuring Arbiter Configration Register : Enabling arbiter $
wait for read mode for result register 6  is enabled
$ Configuring Arbiter priority register : Enabling the queue request source in a
rbitratio slot $
$ Configuring Arbiter priority register : Configuring priority of queue request
source $
$ Enable Queue request source $
$ Configuring Arbiter priority register : Enabling the Auto scan request source
$
$ Configuring Arbiter priority register : Configuring priority of Auto scan sour
ce $
$ Configuring Autoscan Source Mode Register : Enable load event generation $
$ Auto Scan Channel Select register channel status 00000030 $
Before filling the queue, Queue status register value is 00000020
channel input register is written  00000000
channel input register is written  00000001
Queue status register value is 00000001
$ Auto Scan Pending register channel status 00000030 $

 Queue Request source has Higher priority

VADCIrqHandler: VADCG0_SR0_INTERRUPT, Interrupt Received, whose priority is 161

@2930 ns

$ Group 0 Result Register 6  converted value is 00000000  $
$ The converted channel is  0  $
$ The converted source is  0  $
$ Expected Result is 00000000  $


VADCIrqHandler: VADCG0_SR1_INTERRUPT, Interrupt Received, whose priority is 162

@5600 ns

$ Group 0 Result Register 6  converted value is 00000441  $
$ The converted channel is  1  $
$ The converted source is  0  $
$ Expected Result is 00000441  $


VADCIrqHandler: VADCCG0_SR1_INTERRUPT, Interrupt Received, whose priority is 194

@8270 ns

$ Group 0 Result Register 6  converted value is 00000883  $
$ The converted channel is  5  $
$ The converted source is  1  $
$ Expected Result is 00000883  $


VADCIrqHandler: VADCCG0_SR0_INTERRUPT, Interrupt Received, whose priority is 193

@10940 ns

$ Group 0 Result Register 6  converted value is 00000cc4  $
$ The converted channel is  4  $
$ The converted source is  1  $
$ Expected Result is 00000cc4  $

 irqChecks: Checking Interrupts

 irqChecks: PASS: No of "VADCG0_SR0_INTERRUPT" received with priority <"161"> is
: 1;

 irqChecks: PASS: No of "VADCG0_SR1_INTERRUPT" received with priority <"162"> is
: 1;

 irqChecks: PASS: No of "VADCCG0_SR0_INTERRUPT" received with priority <"193"> i
s: 1;
```

Fig 5.b. Standalone console output

IV. CONCLUSION

While verifying an IP, library of test cases can be developed and the test cases can be invoked at the system level and standalone verification environments. Test source reuse reduces the test effort, which reduces the verification time. The verification engineers can spend more time on quality verification rather than developing the same test cases for different environment. There-by assisting for early time to market of the product.

V. FUTURE WORK

Test source reuse can be extended for RTL verification and Post Silicon Validation. The Methodology needs to be tested for this. This will drastically reduce the chip design life cycle.

ACKNOWLEDGEMENT

We would like to thank Prasanth Sasidharan (IFIN ATV C-MODEL), Prasanna Kesavan (IFIN ATV C-MODEL) for their technical support.

REFERENCES

[1] Avss, P.; Prasant, S.; Jain, R.; “Virtual prototyping increases productivity - A case study” Proc. Int’l Sypm. VLSI Design, Automation and Test, 2009(VLSI-DAT’09), pp 96-101. Doi-10.1109/VDAT.2009.5158104

[2] “ITRS 2011 Roadmap (System Drivers) “

[3] Stehr G., Eckmuller J., “Transaction level modeling in practice: motivation and introduction”, proc., Computer-aided design (iccad), 2010 IEEE/acm International conference, pp 324-331

[4] Maman Abdurohman et al., “Transaction Level Modeling for Early Verification on Embedded System Design” 2009 Eight IEEE/ACIS International Conference on Computer and Information Science, DOI 10.1109/ICIS.2009.41

[5] OSCI Standard, OSCI TLM-2.0 Language Reference Manual. Open SystemC Initiative, 2009.

[6] Wolfgang Ecker, Volkan Esen, Michael Velten.” TLM+ Modeling of Embedded HW/SW Systems” Infineon Technologies AG, Germany. EDAA-2010 /978-3-9810801-6-2/DATE10

[7] http://www.itrs.net/

[8] http://www.design-reuse.com

◈◈◈

Management of Network Elements using NETCONF

S.Sindhu[1], G.Sadashivappa[2] & Vignesh C.R[3]

[1&2]Dept of Telecommunication Engineering, R.V.College of Engineering, Bangalore-59, India
[3]CISCO
E-mail : ssindhu1221@gmail.com[1], g_sadashivappa@yahoo.com[2]

Abstract - As the Internet continues to grow, the tasks of operations and management of IP networks and systems are becoming more difficult. SNMP cannot meet current management requirements for complex networks, especially those configuration management needs.In 2006, the IETF released its latest effort, NETCONF, a brand new network management protocol, which is based on the XML encoding method. NETCONF protocol is a next generation configuration protocol developed by IETF organization.

Keywords -*NETCONF;SNMP;Network management; XML*

I. INTRODUCTION

The rapid growth of telecommunication and internet has witnessed the emergence of network devices which added to the complexity of network management in terms of sizes and services. Organizations are increasingly adopting new rich-media business and collaboration technologies. Video loads networks—and it radically changes the demands on the network. Therefore IP networks are becoming larger and more complex as more people use the internet and more enterprises rely on the internet. Efficient and reliable management techniques are necessary to manage these networks.

The legacy approach of CLI (command Line Interface) is vendor dependent approach[4]h, where each vendor as its own command to perform network management function.CLI commands are sent to the devices for monitoring and configuration using protocols like TELNET, SSH etc. However, those systems are dependent on the syntax of the CLI command. The Syntax change of the cli commands results in the implementation modification of the policy to CLI conversion logic.

Traditionally SNMP (Simple Network Management Protocol) has been a major management protocol for the IP-network because of its simplicity. The SNMP is an application protocol that allows logically remote users to inspect or alter management variables. The SNMP utilizes the TCP/IP protocol suite. UDP is the preferred transport of SNMP for IPv4. The size of SNMP over UDP messages is usually limited by the size of the maximum transmission unit (MTU), which is not sufficient for bulk configuration data transfers. To overcome the weakness of SNMP, evolutionary approaches have been attempted in the past few years by adding new functions and security measures, but all have failed to be adopted as standards(in SNMPv2 and SNMPv3). Netconf, however, is connection oriented, requiring a persistent connection between manager and agent. This connection provides reliable and sequential data delivery. However, SNMP has been used mostly in monitoring for fault and performance management, but was hardly used for configuration management especially in system configuration and service provisioning due to its limitations[2][7].The weaknesses of SNMP lead to the investigating alternative approaches to the network management.

II. FUNCTIONAL AREAS FOR MANAGING NETWORKS

The five functional areas for managing telecommunications networks as specified by the Open System Interconnect (OSI) network management framework, known as FCAPS:[8]

1. Fault management,
2. Configuration management,
3. Accounting management,
4. Performance management, and
5. Security management.

Although there are differences between telecommunications networks and the Internet, these functional areas are still the same. The workshop held by IAB published RFC3535 [5] to guide the IETF efforts on future network management work. The recommendations and conclusions of the IAB workshop based on network operators' requirements can be summarized as follows:

1. The network management system must be easy to use for the operators who could perform the configuration of the whole network rather than individual devices.
2. The management protocol should support a standard mechanism to save and restore complete device configuration rather than individual entities.
3. Managed devices should support multiple configurations. The protocol should support the distribution of multiple configurations to devices, and then activate any configuration. In addition, rollback between configurations should be supported.
4. The management protocol should support configuration transactions across multiple devices simultaneously in order to avoid configuration inconsistency. This function significantly simplifies network configuration tasks.
5. Device configuration should be distributed in human readable format so that text processing tools and version control systems can be used to manage and process configuration data.
6. The management protocol should provide authentication, secured transport as well as robust access control that are integrated with the existing key and credential infrastructure.

III. NETWORK MANAGEMENT APPROACHES

Command-Oriented Approach

In the Command Line Interface (CLI) approach, the network administrator logins to the device, and enters commands as illustrated in Fig 1.

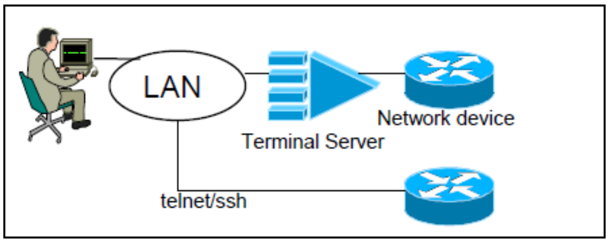


Fig. 1: Management Network for CLI

If a device supports IP, the administrator can telnet or ssh to the device. If the device does not support IP or the IP interface is not configured, the administrator uses a terminal server to access the console interface of the device. To automate the configuration procedure, it is common to compile a sequence of commands in a script file and then send the script file to the device. The following example is a telnet script to show the configuring Fast Ethernet:

```
Router#configure
Router(config)#interface  FastEthernet 0/0
Router(config if)#ip-address 10.1.1.1 255.255.255.0
Router(config-if)#no shut
Router(config if)#exit.
```

Fig 2: Fast Ethernet configuration

As shown in the figure, the CLI commands are executed in *a* hierarchical manner. CLI follows Hierarchical Dependency and Argument Dependency.

As a result of a CLI command execution, one of the three following situations can happen:

1. CLI execution error.
2. Request more input from the administrator.
3. successful execution.

When (1) happens, most of the commands scheduled to be given to the devices cannot be delivered. For case (2). every scheduled command hangs until the additional input is given by the administrator.

IV. XML TEMPLATE FOR CLI COMMANDS

As shown in the example **XML** template of Figure **3,** an XML template is the hierarchy of the XML tags.

```
<?xml version="1.0" encoding="UTF-8"?>
    <InterfaceDetail>
        <entry>
            <IPAddress>10.1.1.1</IPAddress>
            <Mask>255.255.255.0</Mask>
            <Status>no shut</Status>
        </entry>
    </ InterfaceDetail >
```

Fig. 3 : **XML** template for Figure 2

V. XML-CLI API

XML-CLI API is the interface for manipulating XML template. It provides functionalities for loading an XML template, 'materializing' it, and sending it to the network device. The materialization' is the process of converting an XML template into a sequence of CLI commands.[6]

After loaded into the memory by the XML-CLI , a template is translated into the internal data structure of a tree topology. This tree is materialized by traversing it with the arguments given by the X-CLI application programmer [2]. This process is similar to the act of passing arguments to a function which generates some specific control flow. The materialization result of the

Figure 2 is shown in the Figure 3. The arguments *'configure'*, ' FastEthernet 0/0,' '10.1.1.1' '255.255.255.0, are delivered to the XML template in sequence.

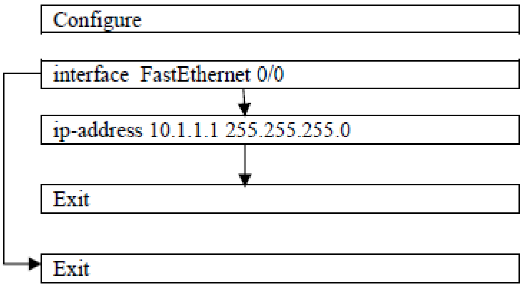


Fig. 4 : Materialized result of figure 3

The materialized commands are sent to the network device sequentially. When an error happens, the failure branch target (depicted as an directed edge in Figure 4) is taken. After the branch, only commands which have 'true' value for the attribute 'always' can be sent to the device. XML-CLI API is greatly enhanced recently for the device monitoring. The <cli> tag is extended to express the monitoring actions, and the monitoring result is parsed automatically by the X-CLI API. This feature aids the programmers who want to develop an application which monitors network device statistics using CLI.

VI. AN OVERVIEW OF NETCONF

The Netconf WG [3] was formed in May 2003.The Netconf WG is attempting to standardize a protocol suitable for the configuration managementof network devices. The Netconf WG defines the Netconf protocol and transport mappings.

Netconf uses an RPC paradigm to define a formal API for network devices. A manager encodes an RPC in XML and sends it to the agent using a secure connection-oriented session.

A Netconf session is the logical connection between a network administrator or network configuration application and a network device. A device must support at least one Netconf session, and may support more than one. It distinguishes between configuration data and state data. Configuration data is the set of read-write data, and state data is read-only data.

The NETCONF protocol [RFC4721] is an XML-based protocol used to manage the configuration of networking equipment. NETCONF is defined to be session-layer and transport independent, allowing mappings to be defined for multiple session-layer or transport protocols. NETCONF can be used within a Secure Shell (SSH) session, using the SSH connection protocol [RFC4254] over the SSH transport protocol [RFC4253]. This mapping will allow NETCONF to be executed from a secure shell session by a user or application. NETCONF can be conceptually layered into four layers, as the following graph shows:

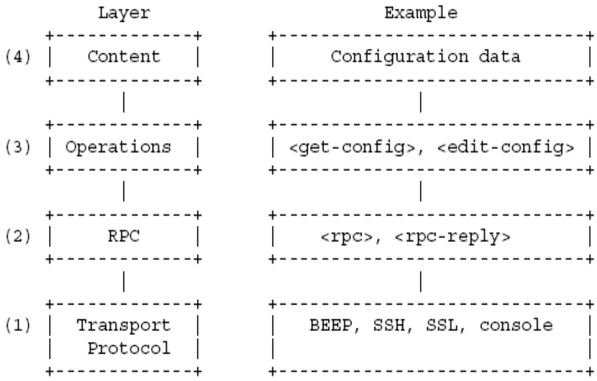

```
           Layer                        Example
      +-------------+        +-----------------------------+
  (4) |   Content   |        |     Configuration data      |
      +-------------+        +-----------------------------+
             |                              |
      +-------------+        +-----------------------------+
  (3) | Operations  |        | <get-config>, <edit-config> |
      +-------------+        +-----------------------------+
             |                              |
      +-------------+        +-----------------------------+
  (2) |     RPC     |        |    <rpc>, <rpc-reply>       |
      +-------------+        +-----------------------------+
             |                              |
      +-------------+        +-----------------------------+
  (1) |  Transport  |        |   BEEP, SSH, SSL, console   |
      |   Protocol  |        |                             |
      +-------------+        +-----------------------------+
```

Fig. 5 : Layers of Netconf

At the bottom, it is the transportation layer. NETCONF is a connection-oriented protocol which supports many kinds of transportation protocols. The RPC layer provides a transportation-independent frame mechanism to encode RPC messages, including <rpc> and <rpc-reply> messages. The client can send an RPC message through the connection to the server to request some information or make certain configurations change. The server responses to the requests of client and sends the results back following the same path. The operation layer defines several operations, such as <get>, <get-config>, <edit-config>, <lock>, <unlock>,<copy-config> etc, which can be used to do configuration management and information retrieving tasks. The top layer of NETCONF is the content layer, which is still under discussion.

The information retrieved from a running system is separated into two classes, configuration data and state data. When a device is performing configuration operations, a number of problems would arise if state data were included.

NETCONF also provides a subtree filtering mechanism to help the <get> and <get-config> operation to retrieve particular XML subtrees.

VII. NETCONF CONFIGURATION PROTOCOL

The Netconf protocol [4] uses XML for data encoding and a simple RPC-based mechanism to facilitate communication between a manager and an agent. The design goals of the Netconf protocol are as follows: Improve interoperability among the devices produced by different vendors. Provide a transport-neutral protocol based on TCP.

Provide ease of implementation to developers using existing XML related tools.Provide operations for getting and editing full or partial configurations. Support actions such as exec commands.

The operations layer includes base and additional management operations. The base operations of the Netconf protocol are defmed as follows:

<get-config>: retrieves all or parts of the specified configuration.

<edit-config>: modifies a configuration.

The <edit-config> base operation has an argument that describes the details of the configuration change. An operation attribute, which is embedded in a configuration subtree, marks the point of the hierarchy at which to perform the operation determined by the value of the operation attribute. The operation attribute has one of three values: merge, replace, and delete. It merges or replaces either all or parts of the specified configuration to the specifiedtarget configuration. It also deletes all or part of the specified configuration. The transaction of modification operations is provided optionally. The values for transaction processing are stop-on-error (default) - and ignore-error.

* <copy-config>: creates or replaces an entire configuration datastore with the contents of another complete configuration datastore.

* <delete-config>: deletes a configuration datastore. <kill-session>: terminates a Netconf session and releases all resources bound to the session.

* <lock>: allows the manager to lock the configuration of a device.

* <unlock>: releases a configuration lock previously obtained with a <lock> operation. <get-all>: retrieves configuration and state information from a device. Netconf defines configuration datastores and allows configuration operations on them. There are some special configuration datastores for the <running> configuration currently running on the network device, the <startup> configuration used at the next reboot and a <candidate> configuration that is used temporarily to make and validate changes. Only the <running> configuration datastore is present in the base model.

A set of additional functionalities that supplements the base operations are called capabilities in Netconf. Capabilities augment the base operations of the device, describing both additional operations and the content allowed inside operations. The Netconf capability permits the client to adjust its behavior to take advantage of features exposed by the device. There are capabilities such as manager, agent, writable-running, candidate, and validate. The manager capability is the manager's capability to manage the agent using the Netconf protocol. The agent capability indicates the agent's capability to be managed. The writable-running capability indicates that the device supports writes directly to the <running> configuration datastore. The candidate capability indicates that the device supports a candidate configuration datastore, which is used to hold configuration data that can he manipulated without impacting the device's current configuration. The validate capability consists of checking a candidate configuration for syntactical and semantic errors before applying the configuration to the device.To support capabilities, more operations are defined:

* <commit>: commits the candidate configuration as the device's new configuration

* <discard-changes>: reverts the candidate configuration to the current committed configuration, if the manager decides that the candidate configuration should not be committed

<validate>: validates the contents of the specified configuration

The RPC layer presents the RPC-based communication model. This layer merely uses RPC elements to define XML messages. Netconf peers use <rpc> and <rpc-reply> elements to provide a transport-independent framing mechanism for encoding RPCs. The elements in this layer are as follows:

<rpc>: expresses request messages of operations in the Netconf protocol. <rpc-reply>: presents the response message to an < rpc-request>. The <ok> element is sent in <rpc-reply> messages if no error

occurs during the processing of an <rpc-request>. Otherwise, the <rpc-error> element is delivered in <rpc-reply> messages.

VIII. CONCLUSION

This paper overview the current approaches to network management. The XML based approach is considered as an essential in supporting the increasing complex and diverse networks.XML is considered as the revolutionary approach to solve the existing network and system management. It facilitate monitoring of network elements using a standard manner and removed the tight coupling of router and switches CLIs from network management stations and created a framework by which routers and switches could be monitored and configured by various third party vendors.

REFERENCES

[1] Khalid Elbadawi,James Yu,"Improving Network Services Configuration Management",IEEE-2011

[2] James Yu, and Imad Al Ajarmeh," An Empirical Study of the NETCONF Protocol", IEEE 2010. [3] Ji Huang, Bin Zhang, Guohui Li, Xuesong Gao,Yan Li ,"Challenges to the New Network Management Protocol:Netconf",IEEE 2009.

[4] Tomoyuki Iijima, Kunihiko Toumura, Hiroyasu Kimura,Makoto Kitani,Takahisa Miyamoto, "Development of Management Interface to Configure Network Equipment", IEEE-2007

[5] Byung-Joon Lee, Taesang Choi and Taesoo Jeong,,"X-CLI : CLI-BASED MANAGEMENT ARCHITECTURE USING XML",IEEE-2006

[6] Mi-lung Choi, Hyoun-Mi Choi, and James W. Hong ,"XM L- Based Configuration Management for IP Network Devices", IEEE 2004

[7] Mi-Jung Choi, James W. Hong, and Hong-Taek Ju "XML-Based Network Management for IP Networks", IEEE 2003.

[8] J. Schoenwaelder, "Overview of the 2002 IAB Network Management Workshop", IETF, RFC3535, May 2003.

❖❖❖

Challenges and Best Practices in Integration Testing For Service Oriented Web Applications

Imran Akhtar Khan & Roopa Singh

Department of Computer Engg & IT, Shri Jagdishprasad Jhabarwal Tibrewala University
Jhunjhunu, Rajasthan
E-mail : imran4bc@gmail.com, roopas1983@gmail.com

Abstract - Service-oriented architecture makes IT applications into composite applications, which are no longer monolithic. Instead, composite applications are composed of many services often developed and deployed independently by separate development teams on different schedules. Since the deployment of these is comparatively easy than testing of the same as everything is at the back end of the system. Here comes the challenge for the testing team to ensure a high quality throughout the Software Testing Life Cycle. This paper represents the challenges and problems that testing team experience when performing integration testing of service oriented web applications. Also, the best practices are discussed to overcome the challenges

Keywords - *STLC (Software Testing Life Cycle) SOA (Service Oriented Architecture), ESB (Enterprise Service Bus), QA (Quality Assurance), Integration Testing, QoS (Quality of Service)*

I. INTRODUCTION

We have seen that service-oriented architecture brings many benefits, but it also brings unique integration testing challenges for the Quality Assurance (QA) team. So integration testing team should be well prepared and adopt best practices to build an approach that will help them effectively address those challenges. Lessons learned from several large-scale SOA implementation projects have been leveraged to define key elements that can make a QA team successful while performing integration testing of service oriented web applications.

The collaborative test team, once put in place, should rely on the well-known mantra “test early, test often.” It is far more effective to find and fix defects close to where they were introduced than it is to find them later on, when it becomes harder to spot the cause of problems—and when fixing them has a much greater impact on the application code and design. In a service-oriented architecture, where there is less control over the entire solution being built, and where services will be reused in different applications and in many different ways, it is recommended to test each service individually first. For example, will services that are reused still perform well with an additional load? Validating the functions and the performance of services being built, as well as the ones being reused, would reduce the number of issues found during the integration phase.

II. SOA OVERVIEW

What exactly is SOA? Why are we using SOA? What are the benefits of using SOA? All these are discussed below.

A. *What is SOA?*

Service-oriented architecture (SOA) is an evolution of distributed computing based on the request/response design paradigm for synchronous and asynchronous applications. An application's business logic or individual functions are modularized and presented as services for consumer/client applications. What's key to these services is their loosely coupled nature; i.e., the service interface is independent of the implementation. Application developers or system integrators can build applications by composing one or more services without knowing the services' underlying implementations. For example, a service can be implemented either in .Net or J2EE, and the application consuming the service can be on a different platform or language

B. *Why SOA?*

The reality in IT enterprises is that infrastructure is heterogeneous across operating systems, applications, system software, and application infrastructure. Some existing applications are used to run current business processes, so starting from scratch to build new infrastructure isn't an option. Enterprises should quickly respond to business changes with agility; leverage existing investments in applications and application infrastructure to address newer business requirements; support new channels of interactions with customers, partners, and suppliers; and feature an architecture that supports organic business. SOA with its loosely coupled nature allows enterprises to plug in new services or upgrade existing services in a granular fashion to address the new business requirements, provides the option to make the services consumable across different channels, and exposes the existing enterprise and legacy applications as services, thereby safeguarding existing IT infrastructure investments.

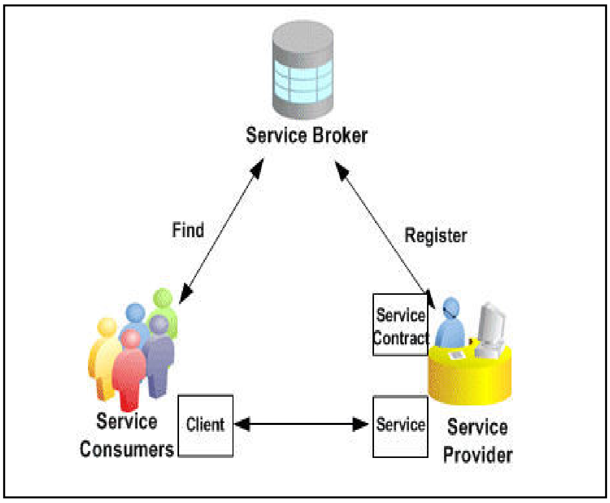


Fig. 1: High level diagram for sample service architecture

C. *Benefits of SOA*

Of the many benefits that SOA extols - agility, reuse, and integration are generally considered to be the most significant.

Agility – SOA assists with responding more quickly and cost-effectively to changing business needs which, in turn, helps the business respond to market conditions faster.

Reuse – SOA promotes more effective reuse at the macro (business unit of work) level rather than micro (code module or class level).

Integration – SOA simplifies interconnection to, and between, existing IT assets; be it integration with partners and customers, or simply integration of internal applications.

III. CHALLENGES FOR INTEGRATION TESTING OF SOA

In SOA environments, systems and applications depend on one another to complete a business process or service as explained in **Fig2**.

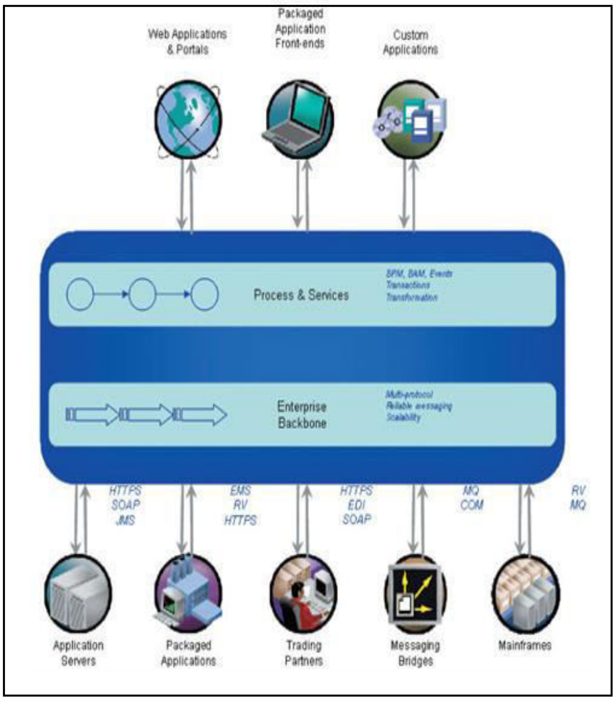


Fig. 2: Dependency of applications to provide services

But the numerous points of exchange create numerous opportunities for defects. To find and correct these defects, SOA and integration testing must address five core challenges of the service oriented applications:

- Defects in the SOA space are extremely difficult to diagnosis because the data in messages is buried in transport protocols that are inaccessible to the typical tester and system administrator. As a result, these defects usually aren't seen until the full system can be tested at the very end of the project, after potential problems have multiplied throughout the system and are more costly to fix. [17]

- XML content and messages that are transmitted back and forth in an SOA environment often encapsulate the actual substance of the message, plus the function that it's supposed to perform. These messages are often written in SOAP (Simple Object Access Protocol), and may contain many customized formats and fields. An error can occur anywhere in the large number of fields, creating an enormous set of permutations and error points and making it extremely hard to have an effective testing solution around the Web Services. [17]

- Web Services and SOA demand very explicit and predictable inputs and outputs. This largely hinges on the accuracy of the application programming. For example, in Microsoft Excel each cell must be formatted to accept the proper data. Cells for dollar figures must accept dollar signs; cells for decimals must accept decimal points, etc. Unless this is explicitly expressed at the API level, errors in message communication can abound. [17]

- SOA testing model isn't just about unit testing. When businesses build an SOA or integration initiative, they're pulling and modifying data from dozens of different systems. Many systems provide only a confirmation message that a process, such as an update, has occurred. There's no guarantee that the data that came from System x, that's now in System y, is accurate or has been put in the right place. Again this opens the door for error if the update is affecting multiple systems and isn't executing correctly. [10]
- Increased business process agility is the overwhelming advantage of moving to SOA. As SOA becomes more prevalent, SOA governance standards will become more prevalent too. However, because SOAs change frequently, it will be difficult to develop and adhere to governance standards and provide high-quality services without adopting a test-driven SOA approach. [10]
- Late binding is also a big challenge for integration testing team members. In presence of late binding, in fact, it is not possible to establish which particular service is invoked in correspondence of a call site. If our system needs to invoke a temperature service, at run time it may choose the cheapest, the fastest or maybe the most precise one. However, this poses serious testing concerns. Ideally, the caller should be tested against any possible service called. In the practice, heuristics should be identified to reduce the testing effort.[9]
- Another recurring problem with integration testing is that missing of services in the moment of integration testing and not all the components needed to test are available. Whether the missing application is an internal module, a vendor application, or a 3rd party service, waiting creates substantial timing and coordination delays in integration projects. [12]

IV. BEST PRACTICES FOR INTEGRATION TESTING

Integration Testing could be performed in many ways. Based on the experience, best practices are being discussed:

- **Increased coverage:** SOA and integration environments are made up of many components:- Web Services, enterprise service buses, legacy assets, databases, files, and numerous transport protocols that move messages and orchestrate services. What's needed is a testing tool to address these components and the layers of complexity because they contain a tremendous amount of logic that falls outside the domain of traditional application testing, such as the dollar sign and decimal point example mentioned earlier. That example illustrates how deeply embedded logic can cripple a composite application that has been built from SOA or integration components. A traditional application testing tool wouldn't find these defects, but SOA and integration testing tools will. [12]
- **Test automation**: SOA's assembly-based approach strongly favors the use of automation and test-driven development techniques. That means developers can automatically test at each stage of an application's development, or with every sprint release.

  This way the application is "evolved" as opposed to built and tested. The only way to constantly test is through an automated testing environment that supports agile practices and tools. Traditional testing tools don't have this automated capability.
- **Process visibility**: SOA testing helps QA teams and developers inspect Web Services, business process, and messages across transport protocols at many different levels. Developers can look at an actual message that moves from system A to system B, and correlate that message to a larger business process, such as making a deposit at the bank, and make sure that the larger process ultimately commits and executes as planned. [12]

  Unlike traditional testing, SOA testing lets developers see each step of a process and the end result. This way SOA testing makes it easy to design test scenarios for complex business processes.
- **Reuse**: As SOA and integration environments grow larger and larger, and test case volume grows, communications between development, operations, and QA personnel can get skewed. To administer such large integration efforts and validate quality, businesses need a tool that lets developers, system administrators, and QA teams share test cases between them.

  SOA testing tools can do this by storing each test case and libraries of test suites, and allowing different application development personnel to access them. For instance, if a QA employee finds an error, he can quickly reference the specific test case and consult the developer to address the issue.
- **Strategy:** Proper integration strategy should be developed in the very beginning of the SOA projects, during the planning phase. Without an effective strategy for integration testing, the testing

of critical logic is usually left until the end of a project. The result is generally substantial re-work that results in overruns and delays. An integration testing strategy can eliminate these problems making SOA projects more predictable to deliver. [12] The problem with unavailable components should be taken into consideration in the integration testing strategy.

- **Tools:** Development of test stubs for replacement of missing components affects project time and requires skilled test team. It remains against management decision if proper tools should be bought or developed in-house. The ability to simulate unavailable systems, services or components is a must to keeping SOA testing on track. [12]. As proposed in [8], integration problems can be checked also at run-time using automated monitoring mechanisms.

V. ESB APPROACH FOR INTEGRATION TESTING

Another way to address the problems during integration is creation of suitable testing environment, containing components which can act as proxy between the communicating services. These components can read and manipulate the communication messages between the services.

The rationale of the proposed testing environment is test-enabled Enterprise Service Bus (ESB), an automated testing framework and methodology, and a mix of tool-supported technologies to enhance the testing of services.

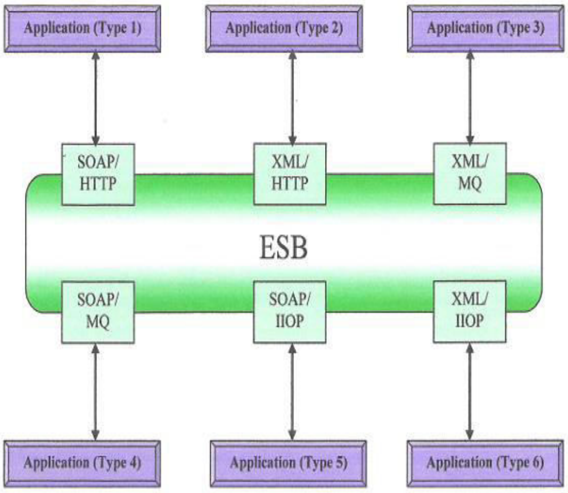


Fig 3: Test Enabled ESB

Test-Enabled ESB, as can be seen in above figure, enables intercepting and controlling the communications between the ESB and the services and applications. An intercepted message can be delivered with a delay, hence changing the apparent QoS and causing temporal contention on resources. The delivered message may be modified, which enables both fault injection and test generation. The message can be recorded, which enables debugging, visualization, many forms of analysis, and other forms of black box test generation [8].

Test-enabled ESB is an enhancement of a regular ESB that provides hooks to be used by testing services. The hooks enable “white box” testing capabilities. The interaction between the services and the bus can be modified and delayed using the APIs provided by the hooks, thus simulating different behavior of the services during integration into SOA.

VI. CONCLUSION

Services are intrinsically distributed and can run on different platforms and can be written in different languages. Despite the numerous tools on the market, which provide mainly unit testing of the services, there is less research on how distributed services, spread around several different machines can be tested and here challenges start for integration testing team.

Sometimes services can be chained with dependencies on other 3rd party services that can change without notice. Hopefully in a SOA environment the interfaces that newer services expose should be the same as these in the older version, but it is necessary to provide means to automatically check whether the services continue to behave according to the user’s assumptions.

This paper reviews challenges for integration testing team members during STLC for service oriented web applications. The benefit of this paper is that it provides analysis on the problems that exists in service oriented architecture. This can be helpful for integration testing team for their first SOA implementation project and there is need of proper solutions and methodology for future research for SOA integration testing projects.

REFERENCES

[1] http://www.oasisopen.org/committees/tc_cat.php ?cat=soa

[2] http://www.adobe.com/enterprise/pdfs/Services_ Oriented_Architecture_from_Adobe.pdf

[3] Thomas Erl, (2004), Service-Oriented Architecture: A Field Guide to Integrating XML and Web Services,3, Prentice Hall.

[4] http://www.ibm.com/developerworks/websphere/library/techarticles/0604_issw/rrlsoa.html [5] Unisannio (2005), State of the Art - Service Engineering (Testing)

[5] Service-Orencted-Architecture http://www.mitre.org/work/tech_papers/tech_papers_09/09_0168/09_0168.pdf

[6] AppLabs (2006), Testing SOA Applications

[7] Gerardo Canfora and Massimiliano Di Penta, (WS-MaTe 2006), SOA: Testing and Self-Checking

[8] Serge Lucio (2005), Don't Wait to Test SOA Applications, http://soa.sys-con.com/node/98059

[9] Steven Devijver: Challenges in testing SOA applications: an introduction, http://soa.dzone.com/news/challenges-testing-soa-applications-introduction

[10] Mamoon Yunus and Rizwan Mallal, Crosscheck Networks (2007), Watch your SOA Testing Blind Spots

[11] Eric Pulier and Hugh Taylor (2006), Solutions to SOA Security

[12] Lori Gipp (2005), Getting the Most Out of Integration Testing

[13] http://www.mercury.com/

[14] www.ibm.com/

[15] http://www.infoworld.com/article/07/05/11/19TCwebservicetest_3.html

[16] G. Canfora and M. Di Penta. Testing Services and Service-Centric Systems: Challenges and Opportunities. IT Professional, vol. 8, no. 2, 2006, pp. 10-17

[17] Bertolino, A., Marchetti, E., Polini, A.: Integration of "components" to test software components. ENTCS 82 (2003)

[18] Kyle Gabhart (2007), SOA World Product Review, SYS-CON Media

[19] Alexandre R.J. Fran¸cois (2005), Architecting Distributed Asynchronous Software Systems, University of Southern California

[20] http://www.telelogic.com/products/doors/index.cfm

[21] www.javaworld.com

[22] http://docs/oracle.com

◈◈◈

Association Rule Mining for Gene Expression Data

O. V. Kale & B. F. Momin

Department of Computer Science & Engineering, Walchand College of Engineering, Sangli
E-mail: Ompriya.2007@gmail.com, bfmomin@yahoo.com

***Abstract* -** Microarray technology has created a revolution in the field of biological research. Association rules can not only group the similarly expressed genes but also discern relationships among genes. We propose a new row-enumeration rule mining method to mine high confidence rules from microarray data. It is a support-free algorithm that directly uses the confidence measure to effectively prune the search space. Experiments on Leukemia microarray data set show that proposed algorithm outperforms support-based rule mining with respect to scalability and rule extraction.

***Keywords* -** *Gene Expression Data, Data Mining, High Confident Association Rules, Bioinformatics.*

I. INTRODUCTION

One main objective of molecular biology is to develop a deeper understanding of how genes are functionally related and, more specifically, to explain how cells control and regulate the expression of their genes and other cellular functions. Deciphering gene relationships has the potential to assist biomedical research in identifying the underlying cause of a disease and developing specific gene-targeting treatments.

Association rule mining method [2] for mining high confident association rules, which describe interesting gene relationships from microarray data sets. The DNA microarray allows parallel genome-wide gene expression measurements of thousands of genes at a given time, under a given set of conditions, for a cell/tissue of interest. Here concentration is on analyzing perturbation microarrays as they are specifically designed to understand the relationships between genes. Perturbation experiments are based on the rationale that, if a gene or cell is no longer able to function normally, the expression levels of other genes that are functionally related may be altered.

II. GENE EXPRESSION DATA

The gene expression data in microarray are presented in M×N matrix where M is the number of microarray experiments and N being the number of genes. The number of experiments M can range from dozens to thousands. On the other hand, the number of genes N can range from hundred to tens of thousands. In some context, M can be referred to as number of transactions or item sets where each gene represents an item. To add to the complexity of representation, each gene is measured in terms of absolute values. However, biologists are more interested in how gene expression changes under different environments in each respective experiment. Thus, these absolute values are discretized according to some predetermined thresholds and grouped under three different levels, namely unchanged, up regulated and down regulated.

Analysis of these massive genomic data has two important goals: First goal is try to determine how the expression of any particular gene might affect the expression of other genes. Second goal of expression data analysis is try to determine what genes are expressed as a result of certain cellular conditions, e.g. what genes are expressed in diseased cells that are not expressed in healthy cells. In this paper, an attempt has been made to review the novel concepts and techniques proposed for mining association rule from the genomic data have been reviewed.

III. PROBLEM DEFINITION

A formal statement of the AR mining problem [2], [3] is as follows: Let the data set D = $\{t_1, t_2 \dots . t_n\}$ be a set of n Microarray experiments and let $I = \{i_1, i_2 \dots . i_m\}$ be the set of all genes (m). Each microarray experiment t consists of a set of genes I from *I*. The aim is to mine all ARs (implications) of the form $I_1 \Rightarrow I_2$ which describe strong relationships between the genes based on the microarray experiments in D. I_1 is referred to as the antecedent itemset and I_2 as the consequent itemset. The strength of an AR is measured by support and confidence and the goal is to identify rules that have a support and confidence greater than the user-specified thresholds minimum support (minsup) and minimum confidence (minconf), respectively.

Definition 1 (Support) : Let I ⊆ *I* be a set of items from D. The support of an itemset I in D, denoted by σ (I), is the proportion of transactions that contain I

$$\sigma(I) = \frac{No\ of\ transactions\ containing\ I}{No\ of\ transactions} \quad (1)$$

The support of an AR I_1 => I_2 is σ (I_1 U I_2). If σ (I) >= minsup, then I is a frequent itemset.

Definition 2 (Confidence): The confidence of an AR I_1 => I_2 is denoted by conf (I_1 => I_2) refers to the strength of the association and is given by

$$\frac{\sigma(I_1 \cup I_2)}{\sigma(I_1)}$$

IV. ASSOCIATION RULE MINING

A. *Preprocessing Of Data*

A gene expression profile can be seen as a single transaction, and each gene, transcript or protein can be thought as an item. The gene expression data can be considered to be a matrix, denoted as G in real expression numbers, which is shown in Table I. The columns denote different samples or conditions. The rows denote genes. In applying association rules to gene expression data, traditional technique would be to first convert each gene expression data into one of three items, down-regulated, up-regulated, or normal expression, which can be denoted as -1, 1 and 0, respectively, as shown in Table II. This is performed by binning the 2 log of the expression level into the three classes [2] with bounds ≤ -r, ≥ r, or in between, where r is a threshold defined by user.

TABLE I

AN EXAMPLE OF MICROARRAY

	Cln3 Exp1	Cln3 Exp2	Clb2 Exp2	Clb2 Exp1	Alpha 0min
YAL001C	0.15		-0.22	0.07	-0.15
YAL002W	-0.07	-0.76	-0.12	-0.25	-0.11
YAL003W	-1.22	-0.27	-0.1	0.23	-0.14
YAL004W	-0.09	1.2	0.16	-0.14	-0.02
YAL005C	-0.6	1.01	0.24	0.65	-0.05

TABLE II

CONVERTED MICROARRAY

	Cln3 Exp1	Cln3 Exp2	Clb2 Exp2	Clb2 Exp1	Alpha 0min
YAL001C	1	0	0	1	0
YAL002W	0	0	0	0	0
YAL003W	0	0	0	1	0
YAL004W	0	1	1	0	0
YAL005C	0	1	1	1	0

B. *Association Rule Extraction*

In this section, we introduce our row-enumeration approach to mining high confident association rules efficiently. This approach addresses the two main shortcomings of AR mining: support pruning and itemset explosion. The main challenge is that no support pruning can take place to reduce the search space. A naive approach would be to grow the entire enumeration tree with no support pruning [3] until no more itemsets can be formed. This would be equivalent to generating all closed itemsets, including those that cannot produce confident rules.

Recently, support-based row-enumeration methods have emerged to facilitate the mining of microarray data. These include FARMER [7], TOPKRGS [11], CARPENTER [4], CHARM [7], CLOSET [10] and RERII [3]. These algorithms effectively prevent itemset explosion by only expanding closed itemsets and enumerating the rows (transactions) rather than the items.

C. *Grow Entire Enumeration Tree with no Support Pruning*

When applying AR mining to microarray data, each microarray experiment is considered to be a single transaction. Consider a sample transaction set as shown in Table III. We will concentrate on algorithm RERII [3] to provide a strong foundation and motivation for our approach. In RERII [3], each node X in Figure 1 will be represented with a three-element group X = {itemlist, sup, childlist}, where itemlist is the closed pattern corresponding to node X, sup is the number of rows at the node and childlist is the list of child nodes of X. For example, the root of the tree can be represented with {{}, 0, {1, 2, 3, 4, 5, 6, 7, 8}} and the node "12" can be represented with {{1, 2}, 2, {3, 4, 5, 6, 8}}.

Given a node X in the row enumeration tree, we will perform an intersection of the itemlist of node X with the itemlist of all its sibling nodes after X. Each intersection will result in a new node whose itemlist is the intersection, whose sup is X.sup + 1 and whose childlist will be available at next level intersection. And each new node will be intersected with its afterward siblings. In this way, the row enumeration tree will be recursively expanded in a depth-first way. The search space (without support pruning) for the transactions in

Table 3 is represented as a row-enumeration tree in Fig. 1a.

When applying AR mining to microarray data, each microarray experiment is considered to be a single transaction. Consider a sample transaction set as shown in table 3. We will concentrate on algorithm RERII [3] to provide a strong foundation and motivation for our approach. In RERII [3], each node X in Fig. 1 will be represented with a three-element group X = {itemlist, sup, childlist}, where itemlist is the closed pattern corresponding to node X, sup is the number of rows at the node and childlist is the list of child nodes of X. For example, the root of the tree can be represented with {{}, 0, {1, 2, 3, 4, 5, 6, 7, 8}} and the node "12" can be represented with {{1, 2}, 2, {3, 4, 5, 6, 8}}.

Given a node X in the row enumeration tree, we will perform an intersection of the itemlist of node X with the itemlist of all its sibling nodes after X. Each intersection will result in a new node whose itemlist is the intersection, whose sup is X.sup + 1 and whose childlist will be available at next level intersection. And each new node will be intersected with it's afterward siblings. In this way, the row enumeration tree will be recursively expanded in a depth-first way. The search space (without support pruning) for the transactions in Table III is represented as a row-enumeration tree in Fig. 1a.

D. Confidence Pruning

This pruning will remove nodes that cannot generate confident I-spanning rules. This pruning is based on an observation of the row enumeration tree's structure. For each node in the tree, we can predict the maximum support [4] and confidence its corresponding itemset can exhibit based on its location within the tree. It is based on the following definitions.

Definition 3 (Maximum support) : Given a node n with k sibling nodes, the maximum support of the itemset at n, represented as σ $_{\max}$ (n) or any of n's potential child nodes is

$$\sigma_{\max}(n) = n \cdot initial\_\sup port + k$$

TABLE III

TRANSACTION SET

Transaction	Items
1	A B C D E G
2	A C D E G
3	C D E F G H I
4	B C D E G
5	A C E G I
6	A D I
7	D I J
8	A B C D G

Definition 4 (Minimum feature): The item i_1 in the itemset I is the minimum feature if

$$\sigma(i_1) \le \sigma(i_2) \;\Big|\; \forall i_2 \in I$$

Definition 5 (I-spanning rule): Given an itemset I, a rule r is an I-spanning rule if

$$antecedent(r) \cup consequent(r) = I \quad and$$
$$|antecedent(r)| = 1$$

Definition 6 (Maximum confidence): Given a node n with minimum feature i, the maximum confidence of any spanning rule of the itemset at n is

$$Conf_{\max}(n) = \frac{\sigma_{\max}(n)}{\sigma(i)}$$

If Conf $_{\max}$ (n) < minconf, then n can be pruned as any further enumeration below the node will only generate less or equally confident child rules. This is because the maximum support of any child node is bounded above by σ $_{\max}$ (n) and the support of its minimum feature can only be greater than or equal to the minimum feature of n. Thus, the child node is bounded above by Conf $_{\max}$ (n). If the current parent node is not pruned by this approach, it is expanded to form a subtree of child nodes following the approach of RERII [4]. Tree after confidence pruning is shown in Fig. 2.

Rules generated by this approach are shown in Table IV.

TABLE IV

Association rules (minsup>3 and minconf>4/5)

Association Rules	Confidence	Support
C=>DEG	4/6	4
E=>CDG	4/5	4
G=>CDE	4/4	4
A=>CG	4/5	4
C=>AG	4/6	4
G=>AC	4/6	4
A=>D	4/5	4
B=>CDEG	2/3	2

B=>CDG	3/3	3
I=>D	3/4	3
J=>DI	1/1	1
F=>CDEGHI	1/1	1
H=>CDEFGI	1/1	1

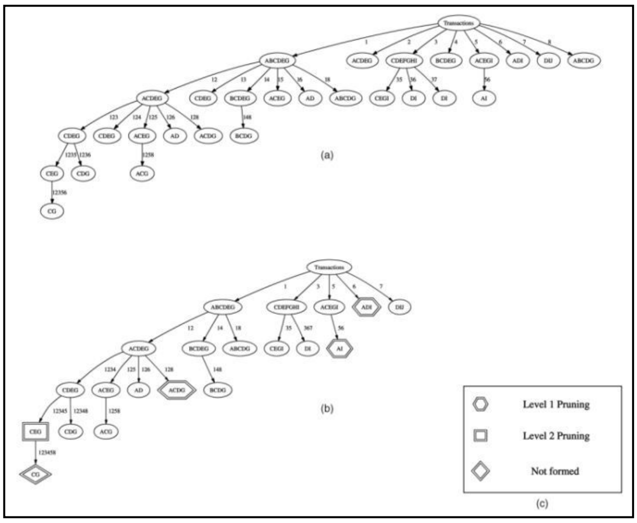
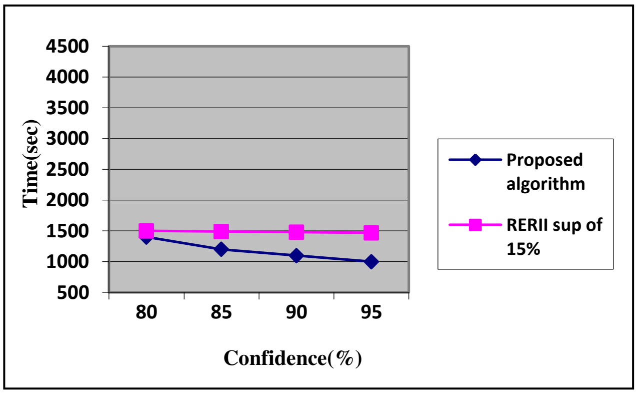


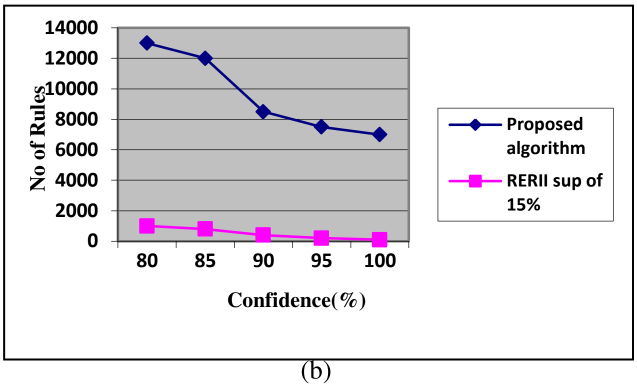
Fig.1. (a) Complete row-enumeration tree (b) Pruned row-enumeration tree. (c) Key.

V. RESULT ANALYSIS

Our experiments are performed on real-life dataset, which is the clinical data on ALL-AML leukemia (ALL). In this dataset, there are 78 tissue samples and each sample is described by the activity level of 12600 genes. Fig. 2 shows the experimental results on this datasets.



(a)



(b)

Fig. 1. Performance on the data set leukemia of RERII with 15% supports and proposed algorithm as confidence is increased (a) Scalability. (b) Number of rules discovered.

REFERENCES

[1] Tara McIntosh and Sanjay Chawla "High-Confidence Rule Mining for Microarray Analysis" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 4, NO. 4, OCTOBER-DECEMBER 2007.

[2] C. Creighton and S. Hanash, "Mining Gene Expression Databases for Association Rules," Bioinformatics, vol. 19, no. 1, pp. 79-86, 2003.

[3] G. Cong, K.-L. Tan, A. Tung, and F. Pan, "Mining Frequent Closed Patterns in Microarray Data," Proc. Fourth IEEE Int'l Conf. Data Mining (ICDM), vol. 4, pp. 363-366, 2004.

[4] F. Pan, G. Cong, K. Tung, J. Yang, and M. Zaki, "CARPENTER: Finding Closed Patterns in Long Biological Datasets," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 637-642, 2003.

[5] Rakesh Agrawal Tomasz Imielinski Arun Swami, "Mining Association Rules between Sets of Items in Large Databases" IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120.

[6] Gao Cong, Anthony K. H. Tung, Jiong Yang, "FARMER: Finding Interesting Rule Groups in Microarray Datasets" Dept. of Computer Science Natl. University of Singapore.

[7] Mohammed J. Zaki and Ching-Jui Hsiao, "CHARM: An Efficient Algorithm for Closed Association Rule Mining" Computer Science Department Rensselaer Polytechnic Institute, Troy NY 12180.

[8] Tim BeiBbarth and Terence P. Speed, “GOstat: find statistically overrepresented Gene Ontologies within a group of genes” Walter and Eliza Hall Institute of medical Research, 1G Royal Parade, Parkville, Vic 3050,Australia.

[9] G. Cong, K.-L. Tan, A.K. Tung, and X. Xu, “Mining TOP-K Covering Rule Groups for Gene Expression Data,” Proc. ACM SIGMOD Int’l Conf. Management of Data, pp. 670-681, 2005.

[10] J. Pei, J. Han, and R. Mao, “CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets,” Proc. ACM SIGMOD Int’l Workshop Data Mining and Knowledge Discovery (DMKD), pp. 21-30, 2000.

❑❑❑

Performance Analysis And Tuning Tool For IBM DB2

Ankit Chouksey

Department of Information Science and Engineering, M. S. Ramaiah Institute of Technology, Bangalore, India
E-mail :ankitchouksey@gmail.com

***Abstract* -** In these years, performance is everything and when we speak about databases, efficiency is the most relied part of every application. Also, efficiency of every application is based on the performance of database by virtue of CPU, IO, Indexes, Locks and SQL Queries. It requires a lot of time for migration of databases, maintenance and monitoring of database, enhancement and fixing bugs, creating patches etc. Hence, we should have a tool that provides appropriate detailed information in a generalized and proper format with file types (ex. txt, csv, log etc). In addition it is helpful to the required results for performance tuning of DB2 in a simpler and efficient way. To resolve these problems and to have an effortless cum time saving performance analysis succeeded by tuning we formulated this paper to overcome the above mentioned problems. This paper explains the approach behind the formation of this tool for IBM DB2.

I. INTRODUCTION

Data migration is the process of creating a copy of an data from one platform to another without disrupting running applications. In this whole process there are many circumstances that might cause performance degradation, data corruption, compatibility issues etc. This performance analysis and tuning Tool help us in various aspects related to migration, monitoring or maintenance etc.

Proposed tool is very much important depends on domain to domain. From the time-cost point of view, it reduces the time up to a great extend, and it will reflects on total cost of development or the work it would concerned. It is important to understand that the cost of statement (SQL and XML) in the database workload is defined by frequency of extension multiplied by individual execution costs. Similarly, resource and effort are also a domain where we can see importance of tool. Where using PAT (Performance analysis and tuning) Tool developer or tester would have to put less efforts and less resources. Customization of output file and format gives a way to understand problems or analyze data in easier way. Cross platform property of PAT Tool gives freedom to run on various platforms, and portability makes this tool more powerful and independent.

Previously, for retrieving all the performance or database related information, analyst/tester/developer had to check all information manually or had to use *IBM DB2* configuration Advisor for performance tuning for some extent. Each tool have its limited functionality out of all, those were made to resolve above problems, and where PAT Tool is a single package to analyze and advise having capability to save the status and information with time stamp for compare performance variation etc.

II. RELATED WORK

Our tool is related to productivity and research into large-scale methodologies, stochastic theory, and heterogeneous models. Unlike many prior approaches, we do not stick to only manual part of tuning. Further, although tools like OMEGAMON DB2 Performance Expert, DB2 Buffer Pool Analyzer, DB2 SQL Performance Analyzer, DB2 Query Monitor, DB2 EXPLAIN etc. also motivated this approach, we harnessed it independently. Though we are the first to construct a light weighted tool for similar work in single package, much existing work has been devoted to the analysis of performance.

Haider Rizvi [1] explains basic rules of thumb and also shared the latest lab experiences on how to best use the capabilities of DB2 along with their tuning expects. Philip Nelson [2] also introduce the principles and techniques of monitoring and tuning DB2 UDB for Linux, Unix and Windows and methods of assessing and quantifying performance, how to avoid problems happening, finding the problem areas and ways to solve the problems identified.

III. PRINCIPLES

The DBI architecture is split into two groups of software: the DBI and the drivers. Actual DBI programming interface defines by DBI, which routes method calls to the appropriate drivers, and provides various support services. Specific drivers are implemented for each different type of database and actually perform the operations on the databases. Figure 3.1 illustrates this architecture.

Therefore, if you are authoring software using the DBI programming interface, the method use is defined within the DBI module. From there, the DBI module works out which driver should handle the execution of the method and passes the method to the appropriate driver for actual execution.

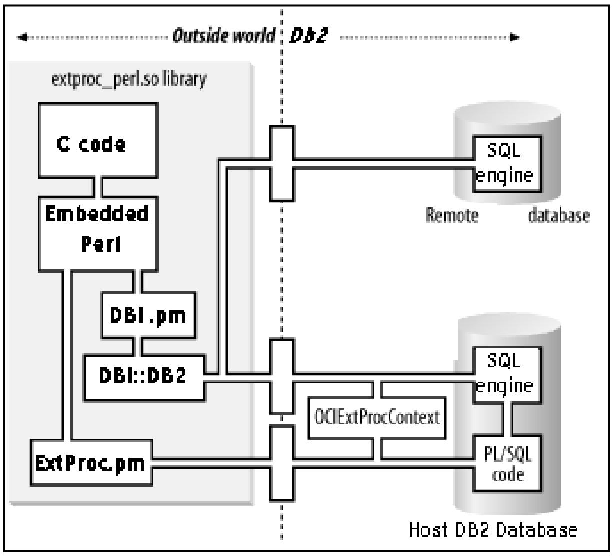


Figure 3.1 The DBI architecture

The DBI module does not perform any database work itself, nor does it even know about any types of databases whatsoever. Figure 3.2 shows the flow of data from a Perl script through to the database.

Under this architecture, it is straightforward to implement a driver for DB2 database. For this, it is required to implement the methods defined in the DBI specification. The data returned from this module is passed back into the DBI module, and from there it is returned to the Perl program. All the information that passes between the DBI and its drivers is standard Perl data types, thereby isolation of the DBI module can be preserve from any knowledge of databases. The separation of the drivers from the DBI makes the DBI a powerful programming interface that can be extended to support almost any database. Drivers currently exist for many popular databases including DB2, Oracle, Informix, mSQL, MySQL, Ingres, Sybase, Empress, SearchServer, and PostgreSQL. There are even drivers for XBase and CSV files.

Drivers are also called database drivers, or DBDs, after the namespace in which they are declared. For example, DB2 uses.

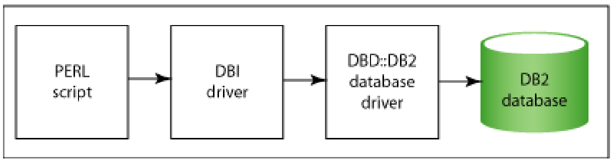


Figure 3.2 Data flow through DBI

DBD::DB2, Informix uses DBD::Informix, and so on. DBI is DataBase Independent where DBD is DataBase Dependent.

IV. FEATURES AND CAPABILITIES

PAT Tool is loaded by various customized features listed below:

- Fast through put and connections.
- System level output on current condition.
- Timestamp files gives you freedom for easy analysis
- with respect to time.
- Creates user friendly structure that would be easy to understand.
- Cross platform gives compatibility and portability.
- Advise you to increase performance at various levels.
- Ex. Table, Index, Tablespace etc.
- Customize file type as per requirement.
- Compatible with Pure Scale DB2.

Table 4.1 describes various keywords that can be use as attributes of PAT Tool. All the keywords are capable to create results in various file formats where by default file type is CSV (Comma separated values) files

Sl No.	Keywords/ Options	Description
1	connections	Display connections that return the highest volume of data to clients, ordered by rows returned.
2	cpu_time	The units of work that are consuming the highest amount of CPU time on the system.
3	query_time	Time spent on query processing in a database.
4	activities	Information about all the activities currently running on a system.

5	dms	List all the dynamic SQL statements from the database package cache ordered by the average CPU time.
6	exec_time	List queries with highest execution time.
7	cpu	List queries with highest CPU consumption.
8	IO	Most I/O intensive queries (including both bufferpool and direct read / read lob activity.
9	wait	Queries with worst relative velocity where relative velocity of the degree to which progress of a query is impacted by waits.
10	least_efficient	Queries with least efficient plans.
11	hit_ratio	Calculating bufferpool hit ratio.
12	containers	List containers and their information ordered by maximum writes.
13	table_read	List tables with read details.
14	high_prep_time	Retrieve a report on the queries with the highest percentage of time spent on preparing.
15	table_read	All input-output details of a table eg. read, delete, insert, update etc.
16	card	Find cardinality.
17	purescale_wait	Identifying Lock waits in PURE SCALE.
18	longest_sql	List longest running queries.
19	reorg	Identify tables that need to be REORG.
20	most_exe_sql	List most executed queries.
21	index_advice	List the index that should be drop or require to improve.
22	tbl_perf	Table performance analysis (Tablespace with worst CPU burner at the top).
23	tbl_perf_1	Table Performance Analysis where , TBRTX----> Table read per transaction and OLTP → Online Transaction Processing. For OLTP, when TBRRTX > 10, → Likely opportunity for improvement when TBRRTX > 100, → Definitely opportunity for improvement when TBRRTX > 1000, → Crisis.
24	db_latency	Identify database input output latency.
25	stmt_lock_wait	List all the currently running statements for the applications we found to be involved in lock waits.
26	lock_holders	List lock holders and their requesters if exist.
27	all	Display information of all the above options. (By default).
28	lock_evmon_on	Turn ON event monitoring to identify lock/deadlock details.
29	lock_evmon_off	Turn OFF event monitoring to stop monitoring and save the lock/deadlock details.

Table 4.1 List of keywords and their description.

V. ILLUSTRATIVE EXAMPLE

PAT Tool can be use by command line interface, and will give output in user friendly environment (GUI (Graphical User Interface)).

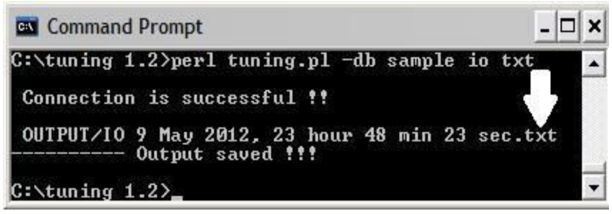
perl tuning.pl [-db/-database] [<database_name>] [[-u/-user][<username>] [-p/-password] [<password>]] [<options>]

[[<Output-file with extension>] or [<File type>]]

For example, let our objective is to identify the SQL queries whose input-output time is high (IO_TIME) or if our objective is to see those top 10 statements having highest execution time in text file of database SAMPLE. So, we will use command

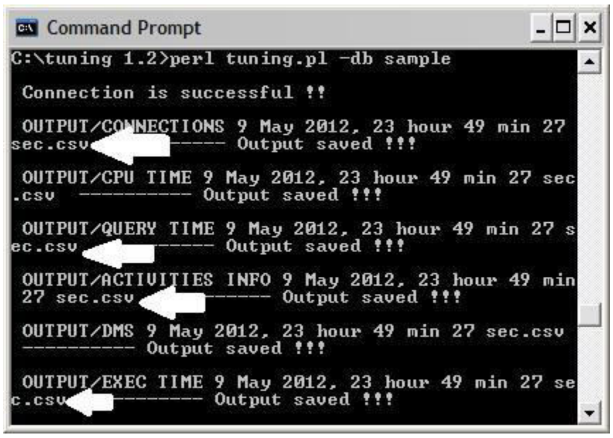
perl tuning.pl -db sample io txt

Figure 5.1 is a screen shot after run above command. Showing a output file with name IO with time stamp is saved successfully in TXT file format. Where, Figure 5.2 is a screen shot without any option, hence by default it creates all the output files in CSV file format.



```
C:\tuning 1.2>perl tuning.pl -db sample io txt

 Connection is successful !!

 OUTPUT/IO 9 May 2012, 23 hour 48 min 23 sec.txt
----------- Output saved !!!

C:\tuning 1.2>_
```

Figure 5.1 PAT Tool creates TXT files (On demand)



```
C:\tuning 1.2>perl tuning.pl -db sample

 Connection is successful !!

 OUTPUT/CONNECTIONS 9 May 2012, 23 hour 49 min 27
sec.csv  ----------- Output saved !!!

 OUTPUT/CPU TIME 9 May 2012, 23 hour 49 min 27 sec
.csv  ----------- Output saved !!!

 OUTPUT/QUERY TIME 9 May 2012, 23 hour 49 min 27 s
ec.csv  ----------- Output saved !!!

 OUTPUT/ACTIVITIES INFO 9 May 2012, 23 hour 49 min
 27 sec.csv  ----------- Output saved !!!

 OUTPUT/DMS 9 May 2012, 23 hour 49 min 27 sec.csv
 ----------- Output saved !!!

 OUTPUT/EXEC TIME 9 May 2012, 23 hour 49 min 27 se
c.csv  ----------- Output saved !!!
```

Figure 5.2 PAT Tool creates CSV files (By default)

VI. FUTURE WORK

As for future work we would suggest to add DB2 performance efficiency tips, bufferpool strategies, enhanced index advisor with option to view all indexes advised across the entire system. Capability to clear all *Advised Indexes (* A reset function to clear the list of advised indexes). That should remove all advised index entries for objects which no longer exist.

This section describes the DB2 memory areas with a focus on how they can impact performance, and explains monitoring and tuning techniques. One of the most important aspects of performance tuning is minimizing physical I/O → proper allocation of memory is essential to optimal performance. Because DB2 environments tend to change as a result of increases in numbers of users, data volumes, transaction rates etc.

Similarly, with respect to locking scenario, a function using same approach as of PAT Tool that controls the amount of memory for managing locks across all applications running concurrently may add on PAT Tool. There must have one locklist per database which defines the percentage of the locklist that an application must be using before lock escalation occurs. When inadequate memory is available for the locklist and lock escalation can occur going from a row lock to a table lock. This can result in performance degradation due to a decrease in application concurrency caused by lock waits and, potentially, deadlocks. Using above approach, we can prevent this degradation. Where, lock escalations must be monitored, and have to adjust locklist and maxlocks values in case of frequent lock escalations. The locklist value should also be increased if the *maxappls* parameter is increased.

Maxappls defines the number of applications that can run concurrently.

For ex. • Lock list → Locklist
• Maximum locks → Maxlocks
• Lock escalations → lock_escals

The following are some suggestions on how to minimize locking:

• Use frequent COMMITS during updating.

• Consider using LOCK TABLE for applications that perform large number of updates.

• Use CURSOR STABILITY.

We recommend to adopt Self Tuning Memory Manager (STMM) as enhancement work for PAT Tool. With DB2 LUW v9.1, IBM introduced automated self-tuning memory to simplify managing the package cache, sort heap, bufferpools, and locklists. The amount of memory allocated to these areas is constrained by the maximum memory allocated to the database shared memory. You can leave the automated self-tuning on. All best practices for advise domain can be added in PAT Tool.

VII.CONCLUSION

This paper has discussed about a useful light weighted tool for performance analysis and tuning in terms of I/O, SQL queries, database architecture and through an analysis of the dynamic conditions for performance degradation it proven very helpful in migrating databases from any to IBM DB2. The feature of PAT Tool greatly improves the consistency of your DB2 LUW environments, as well as reduces downtime. Remember that tuning is not a "set it and forget it" process or not a not a onetime shot: it should constantly monitor critical applications for environmental changes and adjustment of the applications accordingly.

Finally this tool is highly responsive and recommended to the parties involved in functional and regression testing as well as quality assurance team of software development companies within the context of DB2.

VIII. ACKNOWLEDGEMENT

The author would like to thank the following people for their support.

Mr. Prashant Dagar → Architect / manager DB2 Functional Verification Team, India Software labs IBM, Bangalore, India.

Mr. Krishna Raj P M → Assistant professor, MSRIT, Bangalore, India.

IX. REFERENCES

[1] Haider Rizvi (IBM Toronto Lab), Berni Schiefer (IBM Toronto Lab), "DB2 UDB Performance Tuning Fundamentals", May 2005

[2] Philip Nelson (ScotDB Ltd./ Scottish Widows PLC), "DB2 UDB for LUW V8 : Monitoring and Tuning Primer", May 2005.

[3] Fraser McArthur, DB2 Enablement Consultant, IBM Canada Ltd., "Best practices for tuning DB2 UDB v8.1 and its databases", A handbook for high performance, Apr 2004

[4] Bill Wilkins, Partner Enablement, IBM Information Management, IBM Canada, Yasir Warraich, Database Consultant, IBM Canada, "Diagnose and resolve lock problems with DB2 for Linux, UNIX, and Windows", Feb 2007 (Published Oct 2003)

[5] D. H. Brown Associates, "DB2 UDB vs. Oracle8i: Total Cost of Ownership", D. H. Brown Associates, Inc. December 2000.

[6] K. Brown, M. Mehta, M. Carey and M. Livny. "Towards Automated Performance Tuning For

Complex Workloads", Proceedings of 20th International Conference on Very Large Databases, Santiago, Chile, 1994.

[7] G. Lohman, G. Valentin, D. Zilio, M Zuliani, A Skelly, "DB2 Advisor: An optimizer smart enough to recommend its own indexes", Proceedings of the 16th IEEE Conference on Data Engineering, February 2000.

[8] B. Schiefer and G. Valentin. "DB2 Universal Database Performance Tuning", IEEE Data Engineering Bulletin 22(2), June 1999, pp. 12-19

◈◈◈

EyeBlink Controlled Human Computer Interface for Physically Challenged People

Naresh E & Praveen Kumar B P

Dept. of Information & Science, M S Ramaiah Institute of Technology, Bangalore, India
E-mail : nareshkumar.e@gmail.com, praveen.smartans@gmail.com

Abstract **-** This paper proposes an automatic Human Computer Interface (HCI) System to aid and support people with severe disabilities who have an acute need for the interfaces to computer like mouse and keyboard to communicate more interactively and naturally with the machines like Computers and other systems like an audio system, TV, or any such home appliances. The proposed system captures the voluntary eye blinks and the blink patterns are interpreted to make the communication possible between the user and the computers, while the involuntary eye blinks are ignored. The proposed system enables the communication by making use of the Human Facial features such as the eye and the nosetip which are tracked and monitored in real-time. The nosetip coordinates in the live video feed captured by the inexpensive USB camera are translated to become the coordinates of the mouse pointer on the application. The system does not make use of the offline templates. The voluntary right/left eye blinks trigger the right/left mouse clicks. The proposed system is inexpensive and affordable since it works with low cost USB cameras.

Keywords: *Human Computer Interface (HCI), SSR Filter*

I. INTRODUCTION

A great deal of research is put into building systems intended to detect the user movements and facial gestures to steer the communication between the user and the computers. People with disabilities like severe paralysis or people suffering with degenerative neurological diseases really find it difficult to interact with the computers. The motivation for the proposed system is to provide an inexpensive, interactive and more natural and a meaningful mode of communication for people afflicted with degenerative motor diseases or severe paralysis. Our system makes use of the human features like nosetip and the eyes to guide the interaction with the computers.

This goal is realized using an algorithm that distinguishes the true eye blinks from the involuntary ones. The desired facial features are detected and tracked precisely and fast enough to be applied in real-time to steer the interface and enable the communication with the computer. The proposed system can be used to enable the disabled people to interact with simple computer applications and also in playing simple games which provides a more immersive experience to the physically challenged person in a much interactive, natural and meaningful way that requires very minimal effort.

The nosetip movements are tracked and interfaced with the mouse cursor movements and the true & voluntary eye blinks are identified and interfaced with the mouse clicks. These kinds of interfaces are very useful for people who are severely paralyzed or afflicted with motor diseases and are unable to move or control any parts of their body except for their eyes to communicate. The proposed system is more interactive, provides a more natural and an inexpensive and a meaningful means of communication to communicate with the simpler computer applications and to play some simpler role playing games.

The automatic initialization phase is triggered with the analysis of the involuntary eye blinks of the current computer user, which creates an online template of the eye to be used for tracking. This phase occurs each time the current correlation score of the tracked eye falls below a defined threshold in order to allow the system to recover and regain its accuracy in detecting the blinks.

II. OVERVIEW

The overall design of the proposed system comprises of three major modules,

1. Face Detection
2. Finding the Face Candidates

3. Integrating the nosetip movements and the eyeblinks with Mouse cursor coordinates and the Mouse clicks.

A. *Face Detection*

Face detection techniques can be categorized into two approaches

(i) Feature based approach and

(ii) Image-based approach.

Feature-based techniques make use of the apparent facial features such as face geometry and the skin color. But the problem with this approach is the poor performance and reliability under variable lighting conditions. The image-based techniques take advantage of the current advances in the pattern recognition theory. Most of the image-based approaches apply basic window scanning technique to detect the face. This technique practically requires immense floating-point computations to be applied in real-time. Both the approaches applied distinctly and independently pose real-time disadvantages and problems to be applied precisely and practically.

In this paper we propose a method that combines the power and the precision of both the feature-based and image-based approach to yield a more reliable and high speed face detection system. In this method, the essences of both the feature-based and image-based approaches are used to locate the point Between-the-Eyes. The proposed method makes use of the real-time face detection algorithm Six-Segmented Rectangular filter (now called SSR filter from now onwards). The SSR filter is nothing but a rectangle divided into 6 segments and it operates by making use of the bright-dark relation around Between-the-Eyes area concept[1].

Between-the-Eyes point is selected as the face representative since it is common to most people and is easy to find for wide range of face orientation. Between-the-Eyes has dark part (eyes and eyebrows) on both sides, and has comparably bright part on upper side (forehead), and lower side (nose and cheekbone). This characteristic is stable for any facial expression.

Here an intermediate representation of the image called integral image is used to calculate the sums of pixel values in each segment of SSR filter. Firstly, SSR filter is scanned on the image and the average gray level of each segment is calculated from integral image. Then, the bright-dark relations between each segment are tested to see whether its center can be a candidate point for Between-the-Eyes. Later the stereo camera is used to find the distance information and the suitable Between-the-Eyes. Then, the candidates for Between-the-Eyes are evaluated by using a template of Between-the-Eyes matching technique. Finally the true Between-the-Eyes point can be tracked and located [1].

B. *Integral Image*

The SSR filter is computed by using intermediate representation for image called integral image. For the original image i(x, y), the integral image is defined as,

$$Ii(x,y) = \sum_{x' \le x,\, y' \le y} i(x',y') \qquad (1)$$

The integral image can be computed in one pass over the original image by the following pair of recurrences.

$$s(x,y) = s(x,y-1) + i(x,y) \qquad (2)$$

$$Ii(x,y) = Ii(x-1,y) + s(x,y) \qquad (3)$$

Where s(x ,y) is the cumulative row sum, s(x , -1) = 0, and ii(-1, y) = 0.

Using the integral image, the sum of pixels within rectangle D (r_s) can be computed at high speed with four array references as shown in Fig.1.

$$s_r = (Ii(x,y) + Ii(x-W,y-L)) - (Ii(x-W,y) + Ii(x,y-L)) \qquad (4)$$

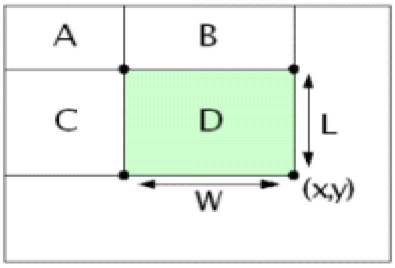


Figure 1. Integral Image

III. SSR FILTER

At the beginning, a rectangle is scanned throughout the input image. This rectangle is segmented into six segments as shown in Fig.2 (a).

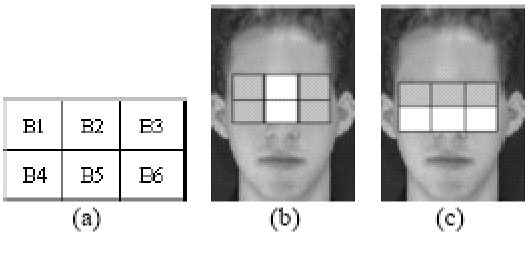


Figure 2. SSR Filter

We denote the total sum of pixel value of each segment (B1 B6) as 1 6 b b S S. The proposed SSR filter is used to detect the Between-the-Eyes based on two characteristics of face geometry [1].

(1) The nose area (n S) is brighter than the right and left eye area (er S and el S , respectively) as shown in Fig.2 (b), where

$$S_n = S_{b2} + S_{b5}$$

$$S_{er} = S_{b1} + S_{b4}$$

$$S_{el} = S_{b3} + S_{b6}$$

Then,

$$S_n > S_{er} \qquad (5)$$

$$S_n > S_{el} \qquad (6)$$

(2) The eye area (both eyes and eyebrows)(e S) is relatively darker than the cheekbone area (including nose) (c S) as shown in Fig. 2 (c), where

$$S_e = S_{b1} + S_{b2} + S_{b3}$$

$$S_c = S_{b4} + S_{b5} + S_{b6}$$

Then,

$$S_e < S_c \qquad (7)$$

The processing flow of Real-Time face detection system is shown in Fig. 4. When expressions (5), (6), and (7) are all satisfied, the center of the rectangle can be a candidate for Between-the-eyes [1].

In Fig.3 (b), Between-the-Eyes candidate area is displayed as the white areas and the non-candidate area is displayed as the black part. By performing labeling process on Fig. 3 (b), the result of using SSR filter to detect Between-the-Eyes candidates is shown in Fig. 3 (a). Because the SSR filter extracts not only the true Between-the-Eyes but also some false candidates, so we use the average Between-the-Eyes template matching technique to solve this problem. To evaluate the candidates, we define the Between-the-Eyes pattern as p_{mn} (m=0,...,31, n = 0, ..., 15) . Then right and left half of p_{mn} is re-defined again separately as p^{r}_{ij} (i=0,...,15, j = 3, ..., 15) and p^{1}_{ij} (i=0,...,15, j = 3, ..., 15), respectively, each has been converted to have average value of 128 and standard deviation of 64. Then the left mismatching value (D_l) and the right mismatching value (D_r) are calculated by using the following equation [1].

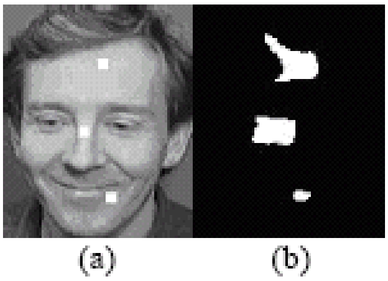


Figure 3. Between-the-Eyes candidates from SSR filter

$$D_l = \sum \frac{\left((p_{ij})_l - (t_{ij})_l\right)^2}{(v_{ij})_l} \qquad (8)$$

$$D_r = \sum \frac{\left((p_{ij})_r - (t_{ij})_r\right)^2}{(v_{ij})_r} \qquad (9)$$

The processing flow of Real-Time face detection system is shown in Fig. 4.

After extracting the templates, we pass them to the support vector machine in order to classify them. Positive classification results mean true faces, while negative ones mean false faces. Since the program will be used by one person at a time, we need to pick one of the positive results as the final detected face. To achieve that, we pick the highest positive result, but before doing so, we will multiply each positive result by the area of the cluster that its template represents [1].

C. *Finding the Nose Tip*

After locating the eyes, the final step is to find the nosetip. The nosetip should fall inside the six-segmented rectangle, so this rectangle becomes our region of interest (ROI) in finding the nosetip [13].

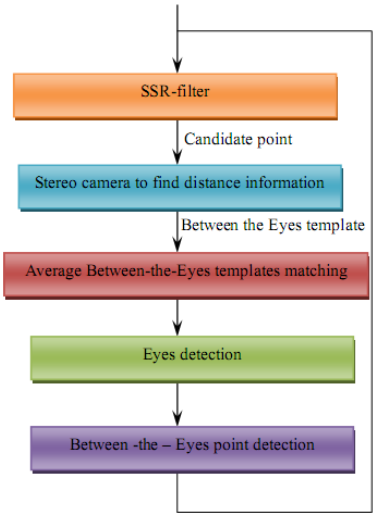

SSR-filter
Candidate point
Stereo camera to find distance information
Between the Eyes template
Average Between-the-Eyes templates matching
Eyes detection
Between -the – Eyes point detection

Figure 4. Processing Flow of Real-Time Face Detection

The nosetip has a convex shape so it collects more light than other features in the ROI because it is closer to the light source. In horizontal intensity profiles we add vertically to each line the values of the lines that precedes it in the ROI, so since that the nose bridge is brighter than the surrounding features the values should accumulate faster at the bridge location; in other words the maximum value of the horizontal profile gives us the 'x' coordinate of the nosetip.

In vertical intensity profiles we add horizontally to each column the values of the columns that precedes it in the ROI the same as in the horizontal profile, the values accumulate faster at the nosetip position so the maximum value gives us the 'y' coordinate of the nosetip. From both, the horizontal and vertical profiles we were able to locate the nosetip position. After locating the nose bridge we need to find the nosetip on that bridge. Since each NBP represents the brightest S2 sector on the line it belongs to, and that S2 sector contains the accumulated vertical sum of the intensities in that sector from the first line to the line it belongs to, we will be using this information to locate the nosetip.

Nose trills are dark areas, and the portion that they add to the accumulated sum in the horizontal profile is smaller than the contribution of other areas; in other words each NBP will add with its S2 sector a certain amount to the accumulated sum in the horizontal profile, but the NBP at the nose trills location will add a smaller amount, we will notice a local minima at the nose trills location, by locating this local minima we take the NBP that corresponds to it as the nose trills location, and the next step is to look for the nosetip above the nose trills. Since the nosetip is brighter than other features it will donate with its S2 sector to the accumulated sum more than other NBPs, which means a local maxima in the first derivate, so the location of the nosetip is the location of the NBP that corresponds to the local maxima that is above the local minima in the first derivate [13]. Tracking the nosetip will be achieved by template matching inside the ROI.

IV. HOUGH TRANSFORM

The Hough transform [12] is a technique which can be used to isolate features of a particular shape within an image. In our proposal we use this to find our eye brows. The classical Hough transform is most commonly used for the detection of regular curves such as lines, circles, ellipses, etc. The Hough transform can be used to identify the parameter of a curve which best fits a set of given edge points [12]. This edge description is commonly obtained from a feature detecting operator such as the Roberts Cross, Sobel or Canny edge detector and may be noisy, i.e. it may contain multiple edge fragments corresponding to a single whole feature. Furthermore, as the output of an edge detector defines only there features are in an image, the work of the Hough transform is to determine both what the features are (i.e. to detect the feature(s) for which it has a parametric (or other) description) and how many of them exist in the image [12]. To find the eyebrow line from the set of threshold points we apply the Hough transform. Sometimes Hough transform gives several lines so we approximate them to a final line which is the eye brow.

V. MOTION DETECTION

To detect motion in a certain region we subtract the pixels in that region from the same pixels of the previous frame, and at a given location (x,y); if the absolute value of the subtraction was larger than a certain threshold, we consider a motion at that pixel.

VI. BLINK DETECTION

We apply blink detection in the eye's ROI before finding the eye's new exact location. The blink detection process is run only if the eye is not moving, because when a person uses the mouse and wants to click, he moves the pointer to the desired location, stops, and then clicks, so basically the same for using the face, the user moves the pointer with the tip of the nose, stops, then blinks [11], [7], [8]. To detect a blink we apply motion detection in the eye's ROI; if the number of motion pixels in the ROI is larger than a certain threshold we consider that a blink was detected, because if the eye is still, and we are detecting a motion in the eye's ROI, that means that the eyelid is moving which means a blink. In order to avoid multiple blinks detection while they are a single blink (because motion pixels will appear while the eye is closing and reopening), the user can set the blink's length, so all blinks which are detected in the period of the first detected blink are omitted.

VII. CONCLUSION

The proposed system can be used to aid people with severe disabilities who have an acute need for the interfaces to computer like mouse and keyboard to communicate more interactively and naturally with the machines like Computers. The automatic initialization phase is greatly simplified in this system, with no loss of accuracy in locating the user's eyes and choosing a suitable open eye template. Another improvement in this system is, it is compatible with inexpensive USB cameras, as opposed to the high- resolution cameras. Higher frame rates and finer camera resolutions could lead to more robust eye detection that is less restrictive on the user, while increased processing power could be used to enhance the tracking algorithm to more accurately follow the user's eyes and recover more gracefully when it is lost.

ACKNOWLEGEMENT

I am happy to present this paper to my Lecturer, Naresh E. I am grateful that Naresh E gave us such an interesting idea and continuous support to do, and I have tried my best to succeed at it. I would also like to thank my family members and my friends for their support. A special thanks to the authors mentioned in the references page.

REFERENCES

[1] Oraya Sawettanusorn, Akihiro Hashimoto, and Hironori Yamauchi, *"Real-Time Face Detection System",*

[2] K. Grauman, M. Betke, J. Lombardi, J. Gips3, G.R. Bradski, *"Communication via eye blinks and eyebrow raises: video-based human-computer interfaces",* Springer-Verlag 2003.

[3] E. Hjelmas and B. K. Low, "Face Detection: A Survey," *Computer Vision and Image Understanding*, vol. 83, pp. 236-274, 2001.

[4] Pragati Garg, Naveen Aggarwal and Sanjeev Sofat, *"Vision Based Hand Gesture Recognition",* 2009.

[5] John J. Magee, Margrit Betke, James Gips, Matthew R. Scott, and Benjamin N. Waber, *"A Human–Computer Interface Using Symmetry Between Eyes to Detect Gaze Direction",* IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, Vol. 38, No. 6, November 2008

[6] Reza Hassanpour, Asadollah Shahbahrami, *"Human Computer Interaction Using Vision-Based Hand Gesture Recognition",* 2008.

[7] Michael Chau and Margrit Betke, *"Real Time Eye Tracking and Blink Detection with USB Cameras",* Boston, 2007.

[8] Arslan Qamar Malik, Jehanzeb Ahmad *"Retina Based Mouse Control(RBMC)",* World Academy of Science, Engineering and Technology, 2007.

[9] Q. Ji, H. Wechsler, A. Duchowski, and M. Flickner. Editorial: special issue: "*Eye detection and tracking detection and tracking. Computing Visual Image Computing Visual Image Understanding*", 98(1):1–3, 2005.

[10] M. Valstar, M. Pantic, Z. Ambadar, and J. Cohn. "*Spontaneous vs. posed facial behavior: Automatic analysis of brow actions"*. International Conference on Multimodal Interfaces, pages 162–170, 2006.

[11] T. N. Bhaskar, F. T. Keat, S. Ranganath, and Y.V. Venkatesh.*"Blink detection and eye tracking for eye localization"*. In Proceedings TENCON 2003, volume 2, pages 821– 824, 2003.

[12] Richard O. Duda, Peter E. Hart *"Use Of The Hough Trasformtion To Detect Lines And Curves In Pictures"* 1971.

[13] P.Viola and M.Jones, *"Rapid object Detection using a Boosted Cascade of Simple Features",* Proceedings of IEEE Conference CVRP,1, pp.511-518, 2001.

◈◈◈

Watermarking of Grayscale Text Document Images Using Histogram Swapping

Jayanth Silesh S, Anvitha Jaishankar & Rashmi N
Dept. of Information Science And Engineering, BNM Institute of Technology, Bangalore, India.
e-mail: jayanthsileshs@gmail.com, anvithaj90@gmail.com, rashmi04nagendra@gmail.com

Abstract— **Digital watermarking is widely believed to be a valid means to discourage illicit distribution of information content. The various methods for text documents are limited because of the binary nature of text documents. In this paper, a novel watermarking scheme which induces an Invisible watermark on grayscale digital text document images is presented. A histogram based algorithm is developed. The watermarked images are resistant to various geometrical attacks like rotation, flipping, translation, aspect ratio changes and resizing and others.**

Keywords- Intensity Modification, Robust watermarking, Histogram modification, Geometrical attacks, Watermarking.

I. INTRODUCTION

The last decade has witnessed the domination of digital media. The new digital reality provides users with many accommodations like high quality, manipulation of the context, creation of perfect duplicates, streaming over the internet etc. The electronic distribution of information is faster, less expensive, and requires less effort than making paper copies and transporting them. Nevertheless these technologies in combination with the World Wide Web enable the perfect copying and distribution of copyrighted material anywhere in the world with practically no cost. In addition, electronic copies are more akin to the original than paper copies. When an electronic copy is made, the original owner and the recipient have identical entities. A person with a photocopy of a journal and a person with the original bound journal may have the same information, but it looks and feels different. Illicit copies of electronic documents are likely to result in major loss of revenues. Therefore a significant problem of non authorized copying and distribution of digital text documents is raised. Also in certain cases the problem of authenticity and reliability is raised (like in medical or military implementations). Digital Watermarking is called to cope with some of these issues. Without methods which prevent or discourage illicit redistribution and reproduction of information content, copyright can be easily infringed. The primary goal of information protection is to permit proprietors of digital information (i.e., the artists, writers, distributors, packagers, market researchers, etc.) to have the same type and degree of control present in the "paper world."

There are primarily three types of text watermarking methods which have been developed previously. They are

(a) Line-Shift Coding – vertically shifts the locations of text lines to encode the document.

(b) Word-Shift Coding – horizontally shifts the locations of words within text lines to encode the document.

(c) Feature Coding – chooses certain text features and alters those features.

These three methods require the original unmarked text for decoding. The proposed method uses the histograms for watermarking and does not require the original text image for recovering the watermark.

II. PROPOSED WATERMARKING TECHNIQUE

The watermarking scheme has a good watermarked image quality that is almost invisible. It is applicable to all grayscale text document images and all image sizes. It is robust against geometrical attacks like rotation, flipping, aspect ratio changes and resizing, warping, shifting and has a good resistance to image tampering. The basic principle involved in this type of watermarking is the swapping of image intensities. The intensities in the histogram bins are selected based on a secret key or certificate. For embedding the signal, the intensities of the images are swapped. Images can be reverted back to its original, if the certificate is known.

A. Watermark embedding

Steps for embedding the watermark:
Step 1: Classify the intensities of the image into histogram bins. The intensities of the images range between 0-255 for grayscale images. Take the intensities in steps of 6. Thus intensities ranging from 0-6 lie in histogram bin 1 and intensities ranging from 7-12 lie in histogram bin 2. Thus, we get a complete of 43 histogram bins. "Figure 1", shows the histogram for grayscale text document image.

Step 2: Choose the signal to be inserted. It must a sequence of zeros and ones. The secret key or the certificate is randomly generated. It must be in the range of 1 to 43. The size of the certificate must be same as the size of the signal inserted.

Step 3: Let variable 'a' hold the first element in the certificate. The 'a'th and the 'a+1'st bin in the histogram are compared(Figure 1). If both the histograms have the same value or zero, increment 'a' by 1.
Insert the signal 1 or 0 based on the following rules.

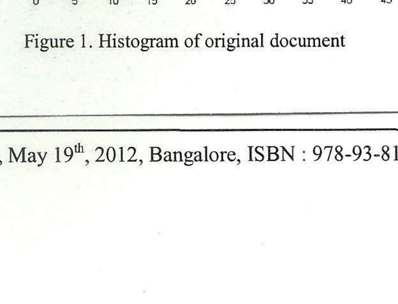


Figure 1. Histogram of original document

Rule 1 for embedding signal 0:

The condition **Hist(a)** < **Hist(a+1)** must be satisfied. If not, go to step 4 to swap the histogram values of ath & a+1st bins.

Rule 2 for embedding signal 1:

The condition **Hist(a)** > **Hist(a+1)** must be satisfied. If not, go to step 4 to swap the histogram values of ath & a+1st bins.

Step 4a:
Calculate the range of intensities by using the formula: "intensity1=(a-1)*6" & "intensity2=((a+1)-1)*6"

All intensities that come under bin 'a', are: intensity1, intensity1+1, intensity1+2, intensity1+3, intensity1+4, intensity1+5.
Similarly, for bin 'a+1' are: intensity2, intensity2+1, intensity2+2, intensity2+3, intensity2+4, intensity2+5.

Step 4b:
The complete image is scanned and all the corresponding intensities are swapped, i.e. intensity1 will be swapped with intensity2 (intensity1↔intensity2). Similarly, intensity1+1 will be swapped with intensity2+1 (intensity1+1↔intensity2+1). This has to be done for all intensities: intensity+1, intensity+2, intensity+3, intensity+4, intensity+5.

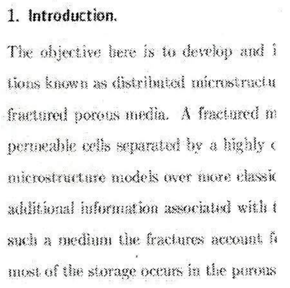
Step 3 has to be repeated until all the elements in the certificate are exhausted. "Figure 2. & Figure 3." show the original and watermarked image respectively. "Figure 3. & Figure 4." show the watermark of the original and watermarked image respectively. "Figure 6." shows the embedding flowchart.

B. *Signal extraction*

The secret key or the certificate is required to extract the signal from the watermarked image. As this is a blind watermarking technique, there is no need for the original image for recovering the signals inserted from the image but the secret key is required.

The steps for the extracting algorithm are as follows:

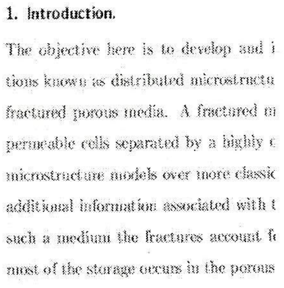
Step 1: The image is taken and the histogram is computed.



1. **Introduction.**

The objective here is to develop and i
tions known as distributed microstructu
fractured porous media. A fractured m
permeable cells separated by a highly c
microstructure models over more classic
additional information associated with t
such a medium the fractures account fo
most of the storage occurs in the porous

Figure 2. Original image



1. **Introduction.**

The objective here is to develop and i
tions known as distributed microstructu
fractured porous media. A fractured m
permeable cells separated by a highly c
microstructure models over more classic
additional information associated with t
such a medium the fractures account fo
most of the storage occurs in the porous

Figure 3. Watermarked image

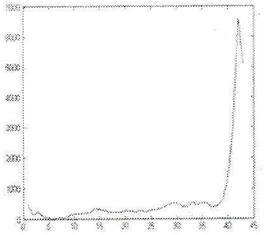


Figure 3. Histogram-Original

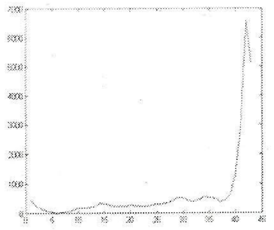


Figure 4. Histogram-Watermarked

Step 2: The first element in the secret key is taken as 'a' and the corresponding values of the histogram hist(a) and hist(a+1) are compared. If they are equal, then the values of 'a' & 'a+1' must be incremented by 1.
Each couple (a, a+1) correspond to a key. The following rules are applied.
If,

hist(a) < hist(a+1) ; then the signal is 0.
hist(a) > hist(a+1) ; then the signal is 1.

Step 2 has to be repeated until all the elements in the certificate are exhausted. The recovered signals are compared with the inserted signals to check the originality of the document.

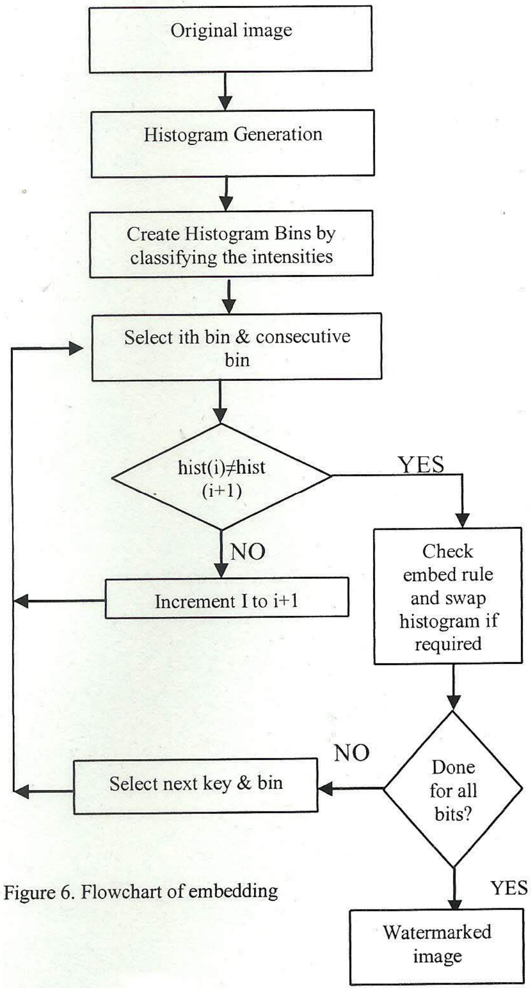


Figure 6. Flowchart of embedding

III. EXAMPLE

Consider 8 bits (1 0 1 0 1 0 1 0) to be embedded into a grayscale text document image. One histogram bin will have a range of 6 intensities. This is done to get a better quality histogram for swapping of intensities. Since 8 bits are embedded into the image secret key chosen should contain 8 elements. Consider the secret keys as: 21, 15, 6 and 40. This can be generated randomly.

Embedding signals: Since the first element in the signal is 1, follow rule 2 to embed and first key is 21, hist(21)=600 and hist(22)=650, hist(21) and hist(22) are swapped. If both the histograms have the same values or zero, then increment 'a' by 1. If they have histogram bin values hist(a)=654 & hist(a+1)=654, then, 'a' must be incremented by 1. Therefore, a=22 & a+1=23. This must be done until both 'a' and 'a+1' have different values. Repeat this process until all the elements in the secret key are exhausted. The swapping is done according to step 4 in the embedding process.

Intensity1=120, intensity2=126. Intensity1+1 is swapped with intensity2+1. Intensity1+2 is swapped with intensity 2+2 and so till all the intensities until intensity1+5 is swapped with intensity2+5.

Signal recovery: First element of the secret key(21) is considered and the histogram of the text document image is created according to step1 of the embedding process. If,

hist(21) < hist(22) ; then the signal is 0.
hist(21) > hist(22) ; then the signal is 1.

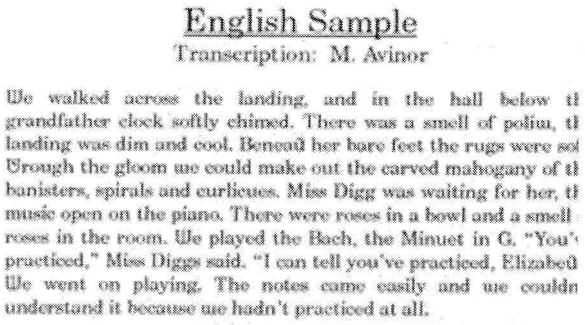
IV. EXPERIMENTAL RESULTS

The above algorithm can be applied to grayscale text document images. The algorithm will work in perfection if the histogram bins do not have too many zero values.

Experiments were conducted for more than 15 text images of different languages. "Table 1", shows the experimental results for the attacked watermarked images. Different attacks like noise, tampering and rotation of images were considered for the encoding of signals 1 0 1 0 1 0 1 0 and 0 1 0 1 0 1 0 1. "Figure7 to Figure 10" show the noised images.

"Graph 1." shows the accuracy level with the increase in the noise level. The x-axis shows the noise level in the image and the y-axis shows the accuracy percentage. The accuracy level goes on decreasing with the increase in the noise level. At a certain point the accuracy level becomes zero. This point is the threshold point. In our experiment the threshold point was found to be 0.8.

"Graph 2." shows the PSNR values for different noise levels. The PSNR values go on decreasing as the noise levels increase.
The PSNR values are the values when the watermarked image are attacked.

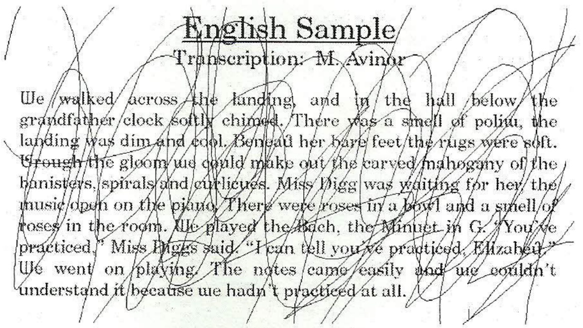


English Sample

Transcription: M. Avinor

We walked across the landing, and in the hall below tl
grandfather clock softly chimed. There was a smell of poliu, tl
landing was dim and cool. Beneað her bare feet the rugs were so
ðrough the gloom we could make out the carved mahogany of tl
banisters, spirals and curlicues. Miss Digg was waiting for her, tl
music open on the piano. There were roses in a bowl and a smell
roses in the room. We played the Bach, the Minuet in G. "You'
practiced," Miss Diggs said. "I can tell you've practiced, Elizabeð
We went on playing. The notes came easily and we couldn
understand it because we hadn't practiced at all.

Figure 7. Watermarked image

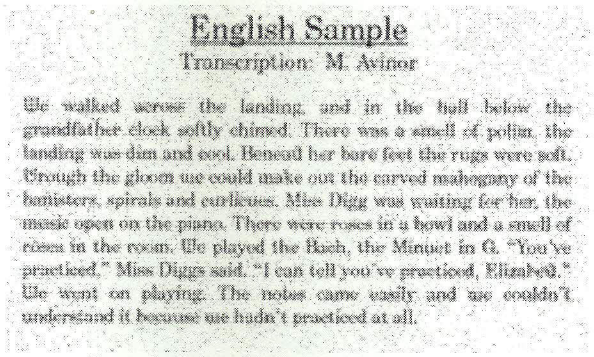


English Sample

Transcription: M. Avinor

We walked across the landing, and in the hall below the grandfather clock softly chimed. There was a smell of poliu, the landing was dim and cool. Beneað her bare feet the rugs were soft. ðrough the gloom we could make out the carved mahogany of the banisters, spirals and curlicues. Miss Digg was waiting for her, the music open on the piano. There were roses in a bowl and a smell of roses in the room. We played the Bach, the Minuet in G. "You've practiced," Miss Diggs said. "I can tell you've practiced, Elizabeð." We went on playing. The notes came easily and we couldn't understand it because we hadn't practiced at all.

Figure 8. Tampered image

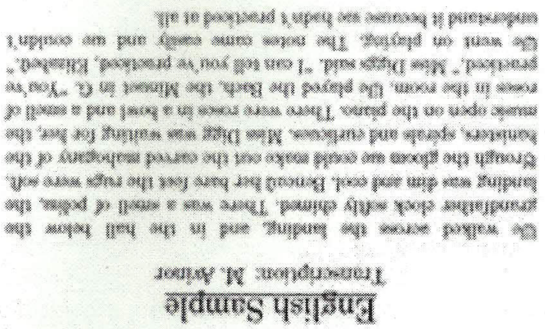


English Sample

Transcription: M. Avinor

We walked across the landing, and in the hall below the grandfather clock softly chimed. There was a smell of poliu, the landing was dim and cool. Beneað her bare feet the rugs were soft. ðrough the gloom we could make out the carved mahogany of the banisters, spirals and curlicues. Miss Digg was waiting for her, the music open on the piano. There were roses in a bowl and a smell of roses in the room. We played the Bach, the Minuet in G. "You've practiced," Miss Diggs said. "I can tell you've practiced, Elizabeð." We went on playing. The notes came easily and we couldn't understand it because we hadn't practiced at all.

Figure 9. "Salt and pepper" noised image (0.08)



English Sample

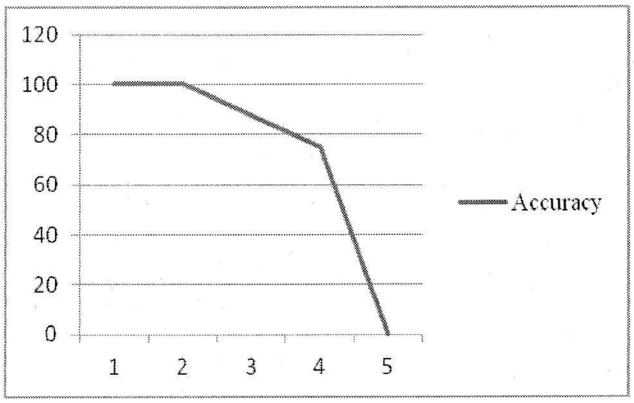
Transcription: M. Avinor

We walked across the landing, and in the hall below the grandfather clock softly chimed. There was a smell of poliu, the landing was dim and cool. Beneað her bare feet the rugs were soft. ðrough the gloom we could make out the carved mahogany of the banisters, spirals and curlicues. Miss Digg was waiting for her, the music open on the piano. There were roses in a bowl and a smell of roses in the room. We played the Bach, the Minuet in G. "You've practiced," Miss Diggs said. "I can tell you've practiced, Elizabeð." We went on playing. The notes came easily and we couldn't understand it because we hadn't practiced at all.

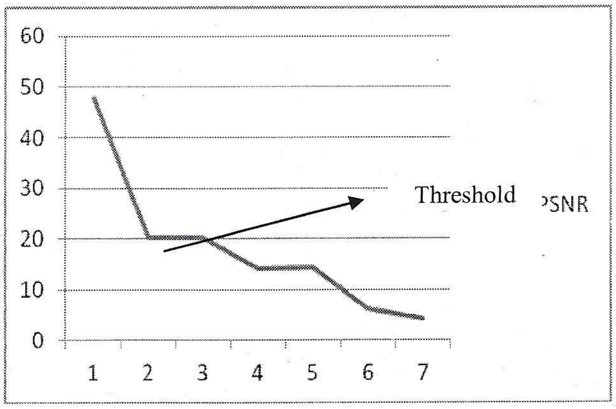
Figure 10. 180 degree rotated image

TABLE 1. TABLE SHOWING THE PSNR AND THE ACCURACY LEVELS OF THE WATERMARKED IMAGES.

Attack	Degree of Attack	Encoded Signal	Decoded Signal	PSNR	Accu-racy
Watermar-ked Image	-	1 0 1 0 1 0 1 0	1 0 1 0 1 0 1 0	47.6641	100%
Noise(Salt & Pepper)	0.02	1 0 1 0 1 0 1 0	1 0 1 0 1 0 1 0	20.1881	100%
Noise(Salt &pepper)	0.02	0 1 0 1 0 1 0 1	0 1 0 1 0 1 0 1	20.2460	100%
Noise(Salt & Pepper)	0.08	1 0 1 0 1 0 1 0	1 0 1 0 1 0 1 0	14.2154	100%
Noise(Salt & Pepper)	0.08	0 1 0 1 0 1 0 1	0 1 0 1 0 1 0 1	14.2944	100%
Noise(Salt & Pepper)	0.5	1 0 1 0 1 0 1 0	1 0 1 0 1 1 1 0	6.2872	87.5%
Noise(Salt & Pepper)	0.8	1 0 1 0 1 0 1 0	1 0 1 0 0 1 1 0	4.2480	75%
Tampering	1st Degree	1 0 1 0 1 0 1 0	1 0 1 0 1 0 1 0	21.4516	100%
Tampering	2nd Degree	1 0 1 0 1 0 1 0	1 0 1 0 1 0 1 0	14.0178	100%
Tampering	3rd Degree	1 0 1 0 1 0 1 0	1 0 1 0 1 0 1 0	8.9003	100%
Rotating	180 Degree towards right	1 0 1 0 1 0 1 0	1 0 1 0 1 0 1 0	9.7319	100%



Graph 1. Showing the accuracy levels. X-axis: noise level. Y-axis: accuracy.



Graph 2. PSNR values,when attacked on Watermarked Image. X-axis : noise level. Y-axis: PSNR value.

I. CONCLUSION

The encoded signal can be retrieved without the original document (blind watermarking. This scheme is highly resistive to any attack. Attacks such as "Slat & Pepper" can be resisted till a magnitude of 0.8 with decoding of about 75-100% accuracy.

II. COPYRIGHT FORMS AND REPRINT ORDERS

The Images used in the above paper are taken from www.google.com.

III. ACKNOWLEDGMENT

We are extremely thankful to Prof **Saritha Chakrasali** & Asst.Prof **Deepthi K Prasad** from BNM Institute of Technology, for their kind consent to be our guide and for providing their support, suggestions and unconditional guidance and their valuable time that helped us immensely in writing this paper.

IV. REFERENCES

[1]Chrysochos E., Fotopoulos V., Skodras A., Xenos M., “Reversible Image Watermarking Based on Histogram Modification”, 11th Panhellenic Conference on Informatics with international participation

[2](PCI 2007), Vol. B, pp. 93-104, 18-20 May 2007, Patras, Greece.

[3]Fotopoulos V., Skodras A.: Digital image watermarking: An overview, invited paper, EURASIP Newsletter, ISSN 1687-1421, Vol. 14, No. 4, Dec. 2003, pp. 10-19 (2003).

[4]Young-Won Kim and Il-Seok Oh, “Watermarking text document images using edge direction histograms”, Science Direct , Pattern Recognition Letters 25 (2004) 1243–1251.

[5]Joachim J. Eggers and Bernd Girod, “Quantization Watermarking”, Proceedings of SPIE Vol. 3971

[6] X.-G. Xia, C. G. Boncelet, and G. R. Arce, “A multiresolution watermark for digital images,” in Proceedings of the IEEE International Conference on Image Processing 1997 (ICIP 97), vol. 1, pp. 548–551, (Santa Barbara, CA, USA), October 1997.

is known. A suitable estimate of the covariance matrix based on the secondary data is derived and plugged into the derived detectors in place of exact covariance matrix.

The rest of the paper is organized as follows. Section 2 presents the Problem formulation and section 3 presents the GLRT detector design. The performance of the derived GLRT is analyzed in section 4 and section 5 presents conclusions.

II. PROBLEM FORMULATION

Consider a narrow band MIMO radar system with s transmitters and r receivers and assume that the antennas are with widely separated to provide uncorrelated reflection coefficients between each transmit/receive pairs of sensors, N denotes the number of pulse train for each transmit antennas, as shown in Fig.1.

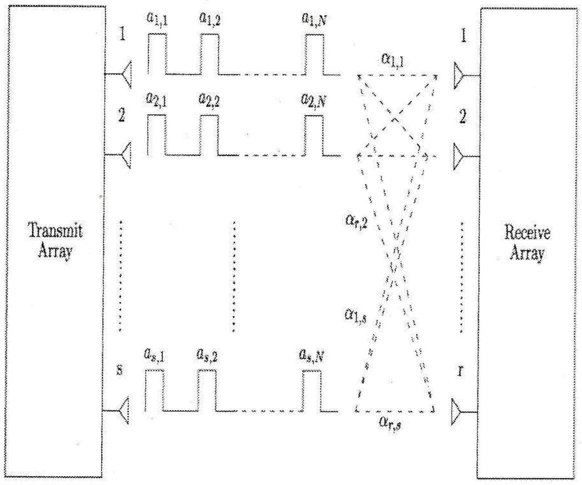


Fig. 1. Schematic representation of considered transmit/receive system.

Moreover, suppose that K (K >2N) secondary data vectors, sharing the same covariance structure of the primary data are available. Denote that r_i and, r_{ik}, i =1,.... r, k =1,..... K , are the received signal from the primary and secondary data, respectively. Then, the problem of detecting a target with MIMO radar can be formulated in terms of the following binary hypotheses test:

$$\begin{cases} H_0 & \begin{cases} r_i = n_i & i = 1, \dots\dots r \\ r_{ik} = n_{ik} & i = 1, .. r, k = 1, \dots K \end{cases} \\ H_1 & \begin{cases} r_i = A\alpha_i + n_i & i = 1, \dots . r \\ r_{ik} = n_{ik} & i = 1, \dots r, k = 1, \dots K \end{cases} \end{cases}$$

(1)

where, A_{NXs}=□$[a_1, \dots \dots a_s] \epsilon C^{NXs}$ is the transmit code matrix which defines s different code words of length N,

$a_l = [a_{l1}, \dots \dots . a_{lN}]^T$□□ i=1,….s are referred as the code word of the 1 th antennas and N is the length of the codeword;

$\alpha_i = [\alpha_{i1}, \dots . . \alpha_{is}]^T$, i=1,…..r are the complex values accounting for both the target backscattering and the channel propagation effects between the transmitters and receivers;

$r_i = [r_{i,1}, \dots . ., r_{1,N}]^T$ *denotes* the N dimensional column vectors and are the echo signals of i th receive antennas contaminated by the clutter;

The clutter vectors n_i, i=*1*,….□r are assumed as compound-Gaussian random vectors, or SIRVs, i.e.,

$$\boldsymbol{n_i} = \sqrt{\boldsymbol{\sigma_i g_i}} \qquad \text{i=1,....r}$$

The textures ($\sigma_i, i = 1, \dots r$) are non-negative random variables and the speckle components ($g_{i,}i = 1, \dots . r$) are correlated N-dimensional complex circular Gaussian vectors and independent each other. At the design stage, we model,(σ_i, i =1,…□r) as the unknown deterministic parameters. This is tantamount to assuming independent zero-mean complex circular Gaussian vectors with covariance matrix

$$\boldsymbol{R_i} = \boldsymbol{E}[\boldsymbol{n_i n_i^\dagger}] = \boldsymbol{\sigma_i R_0}$$

where $\boldsymbol{R_0} = \boldsymbol{E}[\boldsymbol{g_i g_i^\dagger}]$ is the covariance structure.

According to the Neyman-Pearson criterion, the optimum solution to the hypotheses testing problem is the likelihood ratio test, but, for the case at hand, it cannot be implemented since total ignorance of the parameters $\propto_i$ is assumed. A possible way to circumvent this drawback is to resort to the GLRT which is tantamount to replacing the unknown parameters with their maximum likelihood (ML) estimates under each hypothesis.

III. GLRT DESIGN

The GLRT of MIMO radar with an unknown covariance matrix against compound-Gaussian clutter is derived here. More specifically, first assume that the clutter covariance structure is known and derives a GLRT maximizing the likelihood function of the primary data over the remaining unknown parameters. Then a suitable estimate of the unknown covariance based on the secondary data is inserted to make the detector fully adaptive.

A straightforward way to determine the threshold T given a false-alarm rate is to use Monte-Carlo simulation. The number of simulations and computation load are usually huge because of the small value of *P*FA.

$$\frac{\underset{\alpha_1,..\alpha_r,\sigma_1,...\sigma_r}{max} f(r_1,....,r_r|H_1,\alpha_1,..\alpha_r,\sigma_1,..\sigma_r)}{\underset{\sigma_1,....\sigma_r}{max} f(r_1,....r_r|H_0,\sigma_1,...\sigma_r)} \underset{H0}{\overset{H1}{\gtrless}} T \qquad (2)$$

Where
$f(r_1,...r_r|H_1) = f(r_i,..r_r|H_1,\alpha_1,..\alpha_r\sigma_1,...\sigma_r)$ and
$f(r_1,...r_r|H_0) = f(r_i,.....r_r|H_0,\sigma_1,...\sigma_r)$ denote the pdfs of the data under *H1* and *H0*, respectively. More specifically, they are given by

$$f(r_1,.r_r|H_0) = \frac{1}{\pi^{Nr}\prod_{i=1}^{r} det(R_i)} exp\left\{-\sum_{i=1}^{r} r_i^{\dagger} R_i^{-1} r_i\right\} \qquad (3)$$

Under H0 and

$$f(r_1,...r_r|H_1) = \frac{exp\{-\sum_{i=1}^{r}(r_i - A\alpha_i)^{\dagger} R_i^{-1}(r_i - A\alpha_i)\}}{\pi^{Nr}\prod_{i=1}^{r} \det(R_i)} \qquad (4)$$

under H1,

where det(.) denotes the determinant.

To determine the maximum likelihood estimators of□□ σ1,…,σr under H0, the log-likelihood function of (3) is

$$lnf(r_1,....r_r|H_0) = -Nrln\,\pi - N\sum_{i=1}^{r} ln\sigma_i - rln\det(R_0) - \sum_{i=1}^{r}\frac{r_i^{\dagger} R_0^{-1} r_i}{\sigma_i} \qquad (5)$$

It can be shown that (5) admits the following solution

$$\hat{\sigma}_{i0} = \frac{r_i^{\dagger} R_0^{-1} r_i}{N} \qquad (6)$$

i=1,….r

As to the estimators of α1,…..αr and σ1,…,σr under H1, the log-likelihood function of (4) is

$$\ln f(r_1,......r_r|H_1) = Nrln\,\pi - N\sum_{i=1}^{r} ln\sigma_i - rln\det(R_0) - \sum_{i=1}^{r}\frac{(r_i - A\alpha_i)^{\dagger} R_0^{-1}(r_i - A\alpha_i)}{\sigma_i} \qquad (7)$$

Thus it is easy to obtain the maximum likelihood estimate of the complex amplitude α_i as

$$\hat{\alpha}_{i1} = (\mathbf{A}^{\dagger}\mathbf{R}_0^{-1}\mathbf{A})^{-1}\mathbf{A}^{\dagger}\mathbf{R}_0^{-1}\mathbf{r}_i$$

i=1,….,r (8)

$$\hat{\sigma}_{i1} = \frac{\mathbf{r}_i^{\dagger}(\mathbf{R}_0^{-1} - \mathbf{R}_0^{-1}\mathbf{A}(\mathbf{A}^{\dagger}\mathbf{R}_0^{-1}\mathbf{A})^{-1}\mathbf{A}^{\dagger}\mathbf{R}_0^{-1})\mathbf{r}_i}{\mathbf{N}}$$

i= 1,…,r

$$\prod_{i=1}^{r}\frac{\boldsymbol{r}_i^{\dagger}\boldsymbol{R}_0^{-1}\boldsymbol{r}_i}{\boldsymbol{r}_i^{\dagger}(\boldsymbol{R}_0^{-1}-(\boldsymbol{R}_0^{-1}\boldsymbol{A}(\boldsymbol{A}^{\dagger}\boldsymbol{R}_0^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\dagger}\boldsymbol{R}_0^{-1})\boldsymbol{r}_i}\underset{H0}{\overset{H1}{\gtrless}}T \qquad \textbf{(9)}$$

where the detection threshold T is a suitable modification of the original threshold in (2)

Adaptive detection:

In order to make the derived detectors fully adaptive, we replace the covariance matrix R_0 by a suitable estimate in the LHS of (9) based on the secondary data, which shares the same correlation properties with the cell under test and free of signal. To make the detectors ensure the CFAR property w.r.t texture statistics, a normalized sample covariance matrix is adopted[13], based on the secondary data collected by the receiver antennas, that is,

$$\hat{R}_{0i\,\boxed{?}}=\frac{\mathbf{N}}{\mathbf{K}}\sum_{\mathbf{k=1}}^{\mathbf{K}}\frac{\boldsymbol{n}_{i,k}\boldsymbol{n}_{i,k}^{\dagger}}{\boldsymbol{n}_{i,k}^{\dagger}\boldsymbol{n}_{i,k}} \qquad (10)$$

Substituting (10) in (9), we come up with the following adaptive detectors, i.e.,

$$\prod_{i=1}^{r}\frac{\boldsymbol{r}_i^{\dagger}\,\hat{R}_{0i}^{-1}\,\boldsymbol{r}_i}{\boldsymbol{r}_i^{\dagger}(\hat{R}_0^{-1}-\hat{R}_{0i}^{-1}\,\mathbf{A}(\mathbf{A}^{\dagger}\,\hat{R}_{0i}^{-1}\,\mathbf{A})^{-1}\mathbf{A}^{\dagger}\,\hat{R}_{0i}^{-1})\boldsymbol{r}_i}\underset{\boldsymbol{H0}}{\overset{\boldsymbol{H1}}{\gtrless}}\boldsymbol{T1} \qquad (11)$$

Where the detection threshold T1 s are a suitable modification of the original values in (9).

We highlight that, with given N , the proposed adaptive detector end up coincident with (9) as K diverges. However, for finite K , the performance of the estimate and, eventually, of the adaptive detector itself depends upon the actual values of N . It is thus necessary to quantify the loss of the proposed decision strategy with respect to its "non adaptive" counterpart under situations of exact covariance matrix. This is one of the objects of the next section.

IV. PERFORMANCE ASSESSMENT

To compare the performance of derived detector with the detector derived by A.De Maio in Gaussian noise, we simulate the GLRT detector derived by A.De Maio namely GC-GLRT, that is

$$\sum_{i=1}^{r}\boldsymbol{r}_i^{\dagger}\;\hat{\mathbf{R}}_{0\mathbf{i}}^{-1}\mathbf{A}(\mathbf{A}^{\dagger}\,\hat{\mathbf{R}}_{0\mathbf{i}}^{-1}\mathbf{A})^{-1}\mathbf{A}^{-1}\mathbf{A}^{\dagger}\,\hat{\mathbf{R}}_{0\mathbf{i}}^{-1}\boldsymbol{r}_i\underset{H0}{\overset{H1}{\gtrless}}T2 \qquad (12)$$

We assume a clutter-dominated scenario, and the clutter is sampled from K-distribution with pdf

$$f(z)=\frac{\sqrt{2v/\mu}}{\Gamma(v)}\left(\sqrt{\frac{2v}{\mu}}\,z\right)^{v}K_{v-1}\left(\sqrt{\frac{2v}{\mu}}\,z\right) \qquad (13)$$

the texture component $\sqrt{\sigma_i}$ is gamma distribution, with pdf

$$f(\sqrt{\boldsymbol{\sigma}_i})=\frac{1}{\Gamma(v)}\left(\sqrt{\frac{v}{\mu}}\right)^{v}\sqrt{\boldsymbol{\sigma}_i}^{\,v-1}e^{\frac{-v}{\mu\sqrt{\sigma_i}}}\boldsymbol{u}(\boldsymbol{\sigma}_i) \qquad (14)$$

where $\Gamma(.)$ is the Eulerian Gamma function, $\nu > 0$ is the parameter ruling the shape of the distribution, $u(.)$ denotes the unit step function, and $K_v(.)$ is the modified second kind Bessel function with order ν, which rules the clutter spikiness, namely smaller the value of ν , higher the tails of the distribution. The distribution will become Gaussian for $\nu \to \infty$.

The clutter has exponential correction structure covariance matrix R0, the (i,j) element of which is $\rho^{|i-j|}$, where ρ is the one-lag correlation coefficient and is set to 0.9 in the simulations.

Finally, the transmit code matrix A is the orthogonal space time codes, and the signal-to-clutter ratio (SCR) is defined as

$$\mathbf{SCR}=\frac{\boldsymbol{\sigma}^2}{\boldsymbol{Ns}}\,\boldsymbol{tr}[\boldsymbol{A}^{\dagger}\boldsymbol{R}_0^{-1}\boldsymbol{A}]$$

In Fig.2, we analyze the shape parameter v of the clutter that effect the detection performance. The p_d s of derived GLRT and of GC-GLRT are plotted versus SCR with $P_{fa}=10^{-4}$, N=8, r=4, s=4, ρ=0.9,K=32 for several values v. The curves show that the performance of derived GLRT is better in more spikier clutter with smaller v, however, as to GC-GLRT, the situation is reverse. It is because that the derived GLRT is devised in compound-Gaussian clutter, and the GC-GLRT is devised in Gaussian clutter, the performance is better for more matched case. More specifically, the gap in the case $P_d=0.9$ is about 8dB between derived GLRT and GC-GLRT for v=0.5, however the performance of GC-GLRT is better than derived GLRT in the case $P_d>0.9$ for v=5, and the gaps is about 0.7db. It is because that the distribution becomes nearly to Gaussian for high value of v.

The effect of the number of transmit antennas is studied in fig 3 and the P_d s are plotted versus SCR with several values of s. The curve of derived GLRT shows that the performance of s=2 is better than that of s=4 and s=6. As to the GC-GLRT, the increase in the value s can lead to a significant performance improvement.

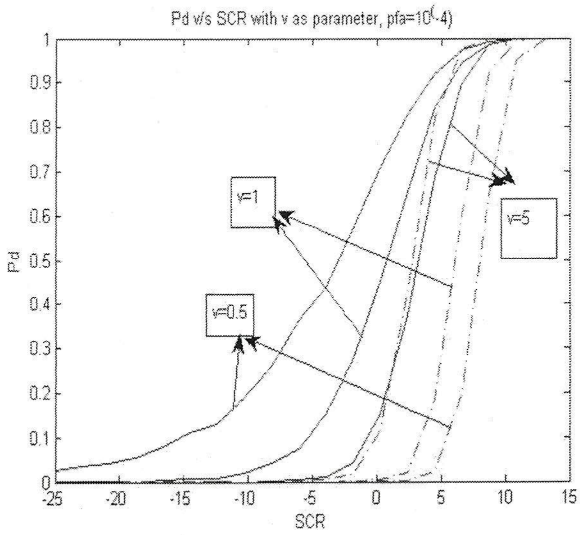


Fig 2. P_d versus SCR plots of derived GLRT(solid curves) and GC-GLRT (dashed curves) receivers, for $P_{fa}=10^{-4}$, N=8, r=4, s=4, ρ = 0.9, K=32, v as a parameter.

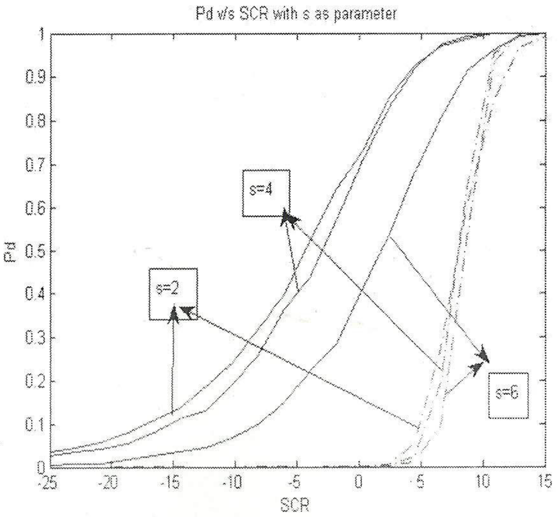


Fig 3. P_d versus SCR plots of derived GLRT (solid curves) and GC-GLRT (dashed curves) receivers, for $P_{fa}=10^{-4}$, N=8, r=4, , ρ = 0.9 , K=32, v=0.5, s as parameter.

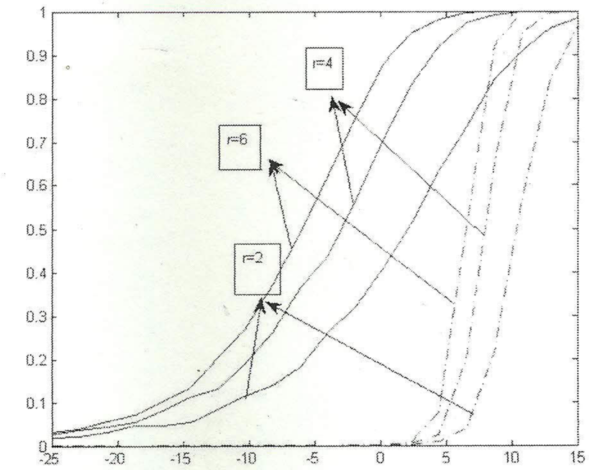


Fig 4. P_d versus SCR plots of derived GLRT (solid curves) and GC-GLRT (dashed curves) receivers, for $P_{fa}=10^{-4}$, N=8, s=4, ρ=0.9, K=32, v=0.5, r as a parameter.

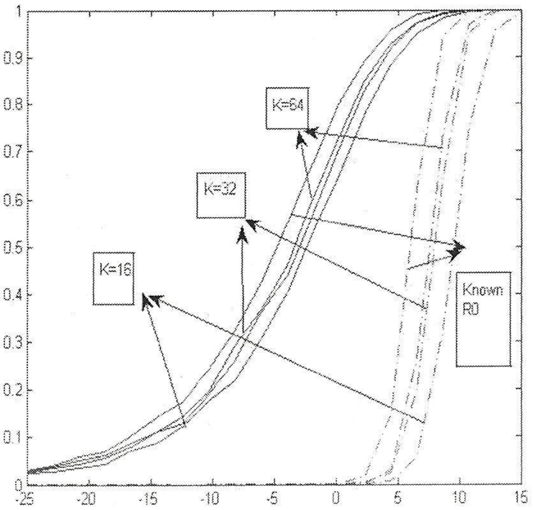


Fig 5. P_d versus SCR plots of derived GLRT (solid curves) and GC-GLRT (dashed curves) receivers, for $P_{fa}=10^{-4}$, N=8, s=4, r=4,ρ=0.9, v=0.5, K as a parameter.

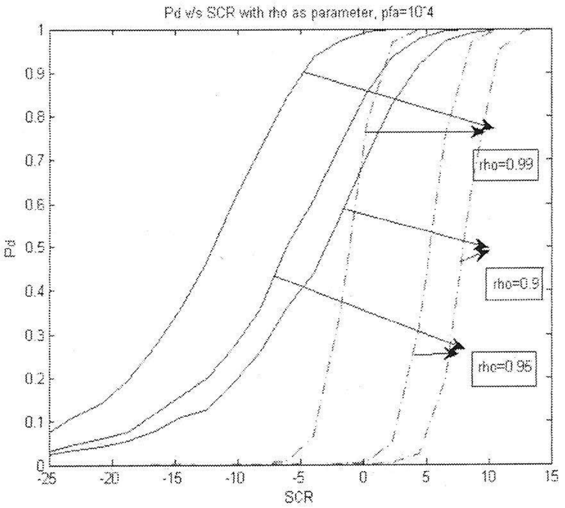
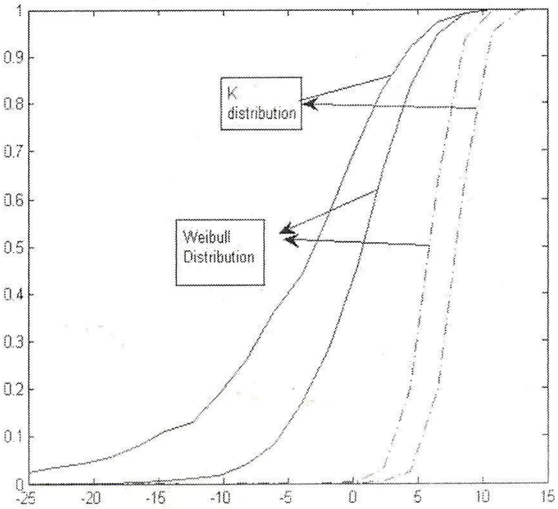


Fig 6. P_d versus SCR plots of derived GLRT (solid curves) and GC-GLRT (dashed curves) receivers, for $P_{fa}=10^{-4}$, N=8, s=4, r=4, ρ=0.9, K=32, v=0.5, rho as a parameter.



The number of receive antennas that effect the performance of detection is analyzed in fig 4, and the P_d s are plotted versus SCR with several values of r. The results show that the performance are increased steadily with increasing the number of r for both derived GLRT and GC-GLRT. Specifically, the gaps in the case P_d=0.9 are about 4dB, 6.2dB between r=2 and r=6 for derived GLRT and GC-GLRT, respectively. The gaps between the two receivers is about 6.4dB in the case r=4.

The effect of the Parameter Rho is considered in fig6 and it shows that the performance increases with increase in rho for derived detector where as in GC-GLRT, the performance is poor for the value of rho=0.9.

To make the derived GLRT fully adaptive, the estimated covariance matrix using the secondary data is inserted into which are obtained with known covariance matrix. The effect of the size K of the secondary data on the performance of derived GLRT and GC-GLRT is analyzed in fig 5. The curves show that the increase in the size K can lead to a significant performance improvement for the derived GLRT. The performance gaps in the case P_d=0.9 between K=16 and K=64 are about 0.9dB and 2.6 dB for derived GLRT and GC-GLRT, respectively. The performance with exact covariance matrix is also accessed, and the results show that the adaptive loss is acceptable.

The comparision of the performance of K-distribution with the Weibull distribution is done in fig 7 and the graph shows that the performance of K-distribution is better compared to the weibull distribution.

V. CONCLUSIONS

This paper has mainly developed the MIMO radar detection problem to compound-Gaussian case, and designed the GLRT detector.The design procedure has been adopted. It should be pointed out that the normalized sampled covariance matrix can ensure CFAR property with respect to textures, however, does not guarantee CFARness with respect to the structure of the covariance matrix.

The performance of the derived GLRT and together with GC-GLRT is studied by several numerical results. The resuts show that the derived GLRT has the better performance in spikier clutter. It has demonstrated that the loss due to the prior uncertainty as to clutter covariance result in acceptable losses, as compared to the case of exact statistics. We should point out that the performance is not increased steadily with increasing the number of transmit antennas and there should be an optimal values with given parameters.

REFERENCES

[1] A. Haimovich, R. S. Blum, L. J. Cimini, "MIMO radar with widely separated antennas," IEEE Signal Processing Magazine, pp: 116 – 129, Jan. 2008.

[2] E. Fishler, A. Haimovich, R. S. Blum, L. Cimini, D. Chizhik, and R. Valenzuela, "Spatial diversity in radars-Models and detection performance," IEEE trans. on Signal Processing, Vol. 54, No. 3, pp: 823-838, Mar. 2006.

[3] A. de Maio, M. Lops, "Design principles of MIMO radar detectors," IEEE Trans. on Aerospace and Electronic Systems, Vol. 43, No. 3, pp: 886-898, Jul. 2007.

[4]. Tarokh, V., Seshadri, N., and Calderbank, A. R. Space-time codes for high data-rate wireless communication: Performance criterion and code construction. *IEEE Transactions on Information Theory*, 44 (Mar. 1998), 744—765. 896 IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS VOL. 43, NO. 3 JULY 2007

[5]. Hochwald, B. M., and Marzetta, T. M. Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading. *IEEE Transactions on Information Theory*, 46 (Mar. 2000),543—564

[6] L. Xu, J. Li, "Iterative generalized-likelihood ratio test for MIMO radar," IEEE Trans. on Signal Processing, Vol. 55, No. 6, pp: 2375-2385, , Jun. 2007.

[7] Fishler, E., Haimovich, A., Blum, R., Cimini, L., Chizhik, D., and Valenzuela, R.
Performance of MIMO radar systems: Advantages of angular diversity. In *Proceedings of Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, Nov. 2004, 305—309.

[8] Fishler, E., Haimovich, A., Blum, R., Cimini, L., Chizhik, D., and Valenzuela, R. Spatial diversity in radars–Models and detection performance. *IEEE Transactions on Signal Processing*, 54 (Mar. 2006),823—838.

[9] E. Conte, M. Longo, "Modelling and simulation of non-Rayleigh radar clutter," IEEE Proceedings-F, Vol. 138, No. 2, pp: 121-130, , Apr. 1991.

[10] K. J. Sangston, K. and Gerlach, "Coherent detection of radar targets in a non-Gaussian background," IEEE Trans. on Aerospace and Electronic Systems, Vol. 30, No. 2, pp: 330-340, Aug. 1994.

[11] K. Gerlach, "Spatially distributed targets detection in non-Gaussian clutter," IEEE Transactions on Aerospace and Electronic Systems, Vol. 35, No. 3, pp: 926-934, , Jul. 1999.

[12] F. Gini; M.V. Greco; L. Verrazzani; "Detection problem in mixed clutter environment as a Gaussian problem by adaptive preprocessing," Electronics Letters, Vol. 31, No. 14, pp: 1189-1190, , Jul. 1995.

[13] F. Gini; M.V. Greco; "Covariance Matrix Estimation for CFAR Detection in Correlated Heavy Tailed Clutter," Signal Processing, Vol.82, No. 12, pp: 1847-1895, , Dec 2002.

[14] Guolong cui; "2 step GLRT design of MIMO radar in compound Gaussian Clutter," 2010.

[15] Sea clutter: "Scattering, the K-distribution and Radar performance" by Keith D. Ward, Robert J.A. Tough and Simon Watts.

A Probabilistic Clustering Based Algorithm For Textual Data Categorization

Keerthi Thota & DS Bhupal Naik

CSE Department, Vignan University, Guntur, India
E-mail : Keerthi.navi@gmail.com, dsbhupal@gmail.com

Abstract - Automatic classification is the task of assigning a set of objects into groups called clusters .The objects in the same cluster are more similar to each other than to those in other clusters. Data categorization is the task is to assign a document to one or more classes or categories. This may be done "manually" or algorithmically .To process the data categorization, dimensionality of the feature of the data has to be reduced. In this paper, we propose probabilistic similarity clustering algorithm to reduce the dimensionality of features. The words in the document are grouped to cluster based on similarity. The probabilistic clustering technique compute the probability with which each point belongs to each cluster and assigns the membership weight .In similar way all the words in the document are classified and based on those membership weights data is categorized. This approach is most appropriate for avoiding the arbitrariness of assigning an object to only one cluster when it may be close to several.

Keywords - *Clustering, Probabilistic clustering, Feature reduction, Text categorization.*

I. INTRODUCTION

Clustering is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters .Text categorization or topic spotting is the task of automatically sorting a set of documents into categories from a predefined set. Currently, text categorization research is pointing in several interesting directions [1]. One of them is the attempt at finding better representations for text; while the bag of words model is still the unsurpassed text representation model, researchers have not abandoned the belief that a text must be something more than a mere collection of tokens, and that the quest for models more sophisticated than the bag of words model is still worth persuing. So that feature extraction has to be done to reduce the high dimensionality. Feature extraction can be done through both feature selection [2][3] [4] and feature reduction. Feature reduction is most efficient than feature selection, but expensive. .The aim of the classical Feature extraction method is to convert the high dimensional data set into low dimensional set by Using through algebraic transformations. principal component analysis, linear discriminant analysis these are the linear transformation feature reduction [5] approaches. We propose a fuzzy similarity feature clustering algorithm, which is an incremental feature clustering approach to reduce the number of features for the text classification task. The words in the feature vector of a document set are represented as distributions, and processed one after another. Words that are similar to each other are grouped into clusters.

II. BACKGROUND WORK

Feature clustering is the efficient feature reduction approach, [8] the divisive information-theoretic feature clustering algorithm was proposed by Dhillon is an information-theoretic feature clustering approach, and is more effective than other feature clustering methods. In these feature clustering methods, each new feature is generated by combining a subset of the original words. However, difficulties are associated with these methods. A word is exactly assigned to a subset, i.e., hard-clustering, based on the similarity magnitudes between the word and the existing subsets, even if the differences among these magnitudes are small. Also, the mean and the variance of a cluster are not considered when similarity with respect to the cluster is computed. Furthermore, these methods require the number of new features be specified in advance by the user. In the previous [6] they used an information-theoretic framework that is similar to Information Bottleneck to derive a global criterion that captures the optimality of word clustering. In order to find the best word clustering, i.e., the clustering that minimizes this objective function, presented a new divisive algorithm for clustering words. This algorithm is reminiscent of the *k*-means algorithm but uses Kullbac Leibler divergences (Kullback and Leibler, 1951) instead of squared Euclidean distances. That divisive algorithm monotonically decreases the objective function value. And also minimizes "within-cluster divergence" and Leibler, 1951) instead of squared Euclidean distances. That divisive algorithm [8] monotonically decreases the objective function value. And also minimizes "within-cluster divergence" and simultaneously maximizes "between-cluster divergence". Thus it find word clusters that are markedly better than the agglomerative algorithms [6] of Baker and McCallum (1998) and Slonim and Tishby (2001). The increased quality of our word clusters translates to higher classification accuracies, especially at small feature sizes and small training sets.

III. OUR METHOD

In this paper we propose a probabilistic feature clustering algorithm, which is an incremental feature clustering approach to reduce the number of features for the textual data categorization. We consider the words in the document from the document set, which are belonging to the different classes. From those words we will construct the feature vector, which is the number of words in the document. The word pattern of each word is constructed by taking the probability of the word occurring in that document set .The word pattern of a word is considered as the sum of the number of occurrences of the word in the document set into word of the document belonging to that class by sum of occurrences of the word in the document set.

Let us consider the documents $d_1,d_2……d_n$ from the document set D which are belonging to the classes $c_1,c_2,c_3……….c_p$ of having the words from $w_1,w_2…….w_k$. The word pattern x_i can be constructed as follows

$X_i = < x_{i1},x_{i2},………x_{ip}>$

$= < p(c_1/w_i), p(c_2/w_i),…….. p(c_p/w_i)>$

Where

$$p(c_j/w_i) = \frac{\Sigma^n{}_{q=1}d_{qi}\times\delta_{qj}}{\Sigma^n{}_{q=1}d_{qi}}$$

For example consider d_1,d_2,d_3,d_4 documents from the document set which are belonging to the classes c_1,c_2, c_1 and c_3.let the occurences of the word w1 is present in the documents 1,2,3,4 is

The probability of word w1 in the class c1 is

$$P(c_1/w_2)= \frac{1*1+2*0+3*1+4*0}{1+2+3+4} = 0.4$$

$$P(c_2/w_1)= \frac{1*0+2*1+3*0+4*0}{1+2+3+4} = 0.2$$

$$P(c_3/w_1)= \frac{1*0+2*0+3*0+4*01}{1+2+3+4} = 0.4$$

The word pattern of word w1 is

x1=<0.4,0.2,0.4>

Thereafter the word pattern construction their mean and variance are calculated. the mean value calculated as the sum of the word pattern by the size of the cluster formed by the word pattern and the variance as the square root of the value x .the value x is defined as the sum of the square of the differences of the word pattern and mean value by the size of the cluster formed by the word pattern. Probabilistic clustering is an self constructing learning approach .word patterns are constructed one by one and their mean and variance are constructed. Initially no clusters will be created by the user, thereafter similarity of word patterns are calculated. Based on the similarity if it is similar to any existing cluster then it is grouped to that cluster and its mean and variance are updated. The algorithm of the probabilistic clustering works like this .First we have to consider the parameters such as the number of word patterns m, number of classes p the documents are occupied ,predefined similarity threshold value ρ, initial deviation σ_0 and initialize the number of cluster value k to zero because initially there would be no clusters. The input given to the algorithm is the constructed word pattern and the output of the algorithm is the clusters formed from the word patterns. First, there are no existing fuzzy clusters on which xi has passed the similarity test. For this case, we assume that xi is not similar enough to any existing cluster and a new cluster .the new cluster is created by incrementing the k value to 1 and now the mean and deviation are to be calculated .Now the mean value is the mean of the that word pattern and deviation is also the deviation of that word pattern. Next word pattern is taken and its similarity is compared with the similarity of the existing cluster as

$$t = arg \max_{1\le j\le k}(\mu_{G_j}(\mathbf{x}_i)).$$

If the value is similar the word pattern is grouped into that cluster .after that the mean value and deviation values are updated as

$$m_{tj} = \frac{S_t \times m_{tj} + x_{ij}}{S_t + 1},$$

$$\sigma_{tj} = \sqrt{A - B} + \sigma_0,$$
$$A = \frac{(S_t - 1)(\sigma_{tj} - \sigma_0)^2 + S_t \times m_{tj}{}^2 + x_{ij}{}^2}{S_t},$$
$$B = \frac{S_t + 1}{S_t}\left(\frac{S_t \times m_{tj} + x_{ij}}{S_t + 1}\right)^2,$$

for $1 \leq j \leq p$, and

$$S_t = S_t + 1.$$

Otherwise the word pattern is not similar then number of cluster s value k will be incremented by creating the new cluster and the mean and variance of that word pattern are the mean and deviation of that newly created cluster. This process continuous until all the word patterns are processed and the desired number of clusters are formed automatically.

IV. EATURE REDUCTION

Feature clustering is an efficient approach for feature reduction Groups all features into some clusters where features in a cluster are similar to each other Let D be the matrix consisting of all the original documents with m features and D' be the matrix consisting of the converted documents with new k features .

\New feature set w ‘ = { $w_1{}^{'}w_2{}^{'}$,…..$w_k{}^{'}$}

corresponds to a partition{w_1,W_2,…,W_k} of the original feature set W

By applying the probabilistic clustering approach to the document set, the desired number of clusters are formed .Each cluster represents the extracted feature and the thus somewhat the dimensionality reduction has been achieved that is high dimensional dataset has been reduced to the low dimensional dataset. The original document set D is converted into $D^{'}$The feature extraction form can be represented as

$D^{'}$=DT

Where

D= $[D_1,D_2……….D_k]^T$

$D^{'}$= $[D_1{}^{'},D_2{}^{'}……….D_k{}^{'}]^T$

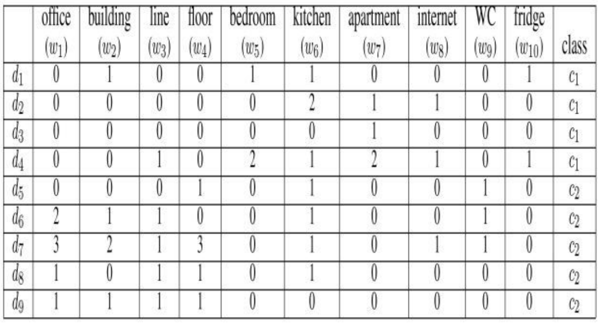
And T is the weighting matrix

The goal of feature reduction is achieved by finding an appropriate T such that k is smaller than m By applying our clustering algorithm, word patterns have been grouped into clusters, and words in the feature vector W are also clustered accordingly. For one cluster, we have one extracted feature. Since we have k clusters, we have k extracted features. The elements of T are derived based on the obtained clusters, and feature extraction will be done. We propose three weighting approaches: hard, soft, and mixed. In the hard-weighting approach, each word is only allowed to belong to a cluster, and so it only contributes to a new extracted feature.

For example we consider one sample document set D, which is shown in the below figure fig1.

FIGURE:1

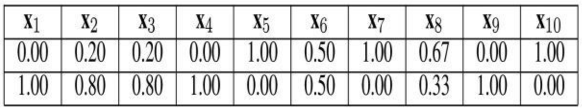
SAMPLE DOCUMENT



	office (w_1)	building (w_2)	line (w_3)	floor (w_4)	bedroom (w_5)	kitchen (w_6)	apartment (w_7)	internet (w_8)	WC (w_9)	fridge (w_{10})	class
d_1	0	1	0	0	1	1	0	0	0	1	c_1
d_2	0	0	0	0	0	2	1	1	0	0	c_1
d_3	0	0	0	0	0	0	1	0	0	0	c_1
d_4	0	0	1	0	2	1	2	1	0	1	c_1
d_5	0	0	0	1	0	1	0	0	1	0	c_2
d_6	2	1	1	0	0	1	0	0	1	0	c_2
d_7	3	2	1	3	0	1	0	1	1	0	c_2
d_8	1	0	1	1	0	1	0	0	0	0	c_2
d_9	1	1	1	1	0	0	0	0	0	0	c_2

Let D TABLE :1

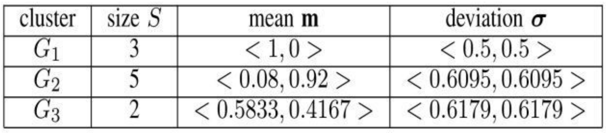
Word pattern construction



x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0.00	0.20	0.20	0.00	1.00	0.50	1.00	0.67	0.00	1.00
1.00	0.80	0.80	1.00	0.00	0.50	0.00	0.33	1.00	0.00

Mean and variance of the clusters

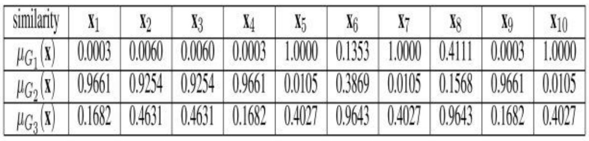
TABLE:2



cluster	size S	mean $\mathbf{m}$	deviation $\boldsymbol{\sigma}$
G_1	3	$<1,0>$	$<0.5,0.5>$
G_2	5	$<0.08,0.92>$	$<0.6095,0.6095>$
G_3	2	$<0.5833,0.4167>$	$<0.6179,0.6179>$

TABLE: 3

Similarities of word pattern



similarity	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
$\mu_{G_1}(\mathbf{x})$	0.0003	0.0060	0.0060	0.0003	1.0000	0.1353	1.0000	0.4111	0.0003	1.0000
$\mu_{G_2}(\mathbf{x})$	0.9661	0.9254	0.9254	0.9661	0.0105	0.3869	0.0105	0.1568	0.9661	0.0105
$\mu_{G_3}(\mathbf{x})$	0.1682	0.4631	0.4631	0.1682	0.4027	0.9643	0.4027	0.9643	0.1682	0.4027

be a simple document set, containing 9 documents d1, d2; . . . ; d9 of two classes c1 and c2, with 10 words “office,” “building,”; . . . ; “fridge” in the feature Vector W. The document of the kitchen is taken and word pattern for that is constructed as shown in the table 1.After that the mean and variance and the size of clusters are calculated and are shown in the table 2.

We run our self-constructing clustering algorithm, by setting σ0= 0.5 and ρ= 0.64, on the word patterns and obtain 3 clusters G1, G2, and G3, which are shown in Table 3.The fuzzy similarity of each word pattern to each cluster is shown in Table 4. The weighting matrices TH, TS, and TM obtained by hard-weighting, soft-weighting, and mixed-weighting (with γ= 0.8), respectively, are shown in Table 5. The transformed data sets D'H, D'S, and D'M obtained for different cases of weighting are shown in Table 6.Based on D'H, D'S, or D'M, a classifier with two SVMs is built. Suppose d is an unknown document, and d =<0; 1; 1; 1; 1; 1; 0; 1; 1; 1> . We first convert d to d' by .Then, the transformed document is obtained as d'H =dTH =<2; 4; 2>; d'S =dTS =<2:5591; 4:3478; 3:9964>, or d'M =dTM = <2:1118; 4:0696; 2:3993>. Then the transformed unknown document is fed to the classifier. For this example, the classifier concludes that d belongs to c2.

The above figure represents the performances of the different feature reduction methods applied on the document set .This figure points out that our approach is the best among all other approaches used for feature reduction.

By this algorithm, the derived membership Functions match closely with and describe properly the real distribution of the training data. Besides, the user need not specify the number of extracted features in advance, and trial-and-error for determining the appropriate number of extracted features can then be avoided. Experiments on real-world data sets have demonstrated that our method can run faster and obtain better extracted features than other methods.

.

TABLE:4

Weighting Matrices Of Hard,Soft And Mixed

$$\mathbf{T}_H = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \mathbf{T}_S = \begin{bmatrix} 0.0003 & 0.9661 & 0.1682 \\ 0.0060 & 0.9254 & 0.4631 \\ 0.0060 & 0.9254 & 0.4631 \\ 0.0003 & 0.9661 & 0.1682 \\ 1.0000 & 0.0105 & 0.4027 \\ 0.1353 & 0.3869 & 0.9643 \\ 1.0000 & 0.0105 & 0.4027 \\ 0.4111 & 0.1568 & 0.9643 \\ 0.0003 & 0.9661 & 0.1682 \\ 1.0000 & 0.0105 & 0.4027 \end{bmatrix}, \mathbf{T}_M = \begin{bmatrix} 0.0001 & 0.9932 & 0.0336 \\ 0.0012 & 0.9851 & 0.0926 \\ 0.0012 & 0.9851 & 0.0926 \\ 0.0001 & 0.9932 & 0.0336 \\ 1.0000 & 0.0021 & 0.0805 \\ 0.0271 & 0.0774 & 0.9929 \\ 1.0000 & 0.0021 & 0.0805 \\ 0.0822 & 0.0314 & 0.9929 \\ 0.0001 & 0.9932 & 0.0336 \\ 1.0000 & 0.0021 & 0.0805 \end{bmatrix}.$$

TABLE:5

Transformed Datasets

	(w'_1)	(w'_2)	(w'_3)
d'_1	2	1	1
d'_2	1	0	3
d'_3	1	0	0
d'_4	5	1	2
d'_5	0	2	1
d'_6	0	5	1
d'_7	0	10	2
d'_8	0	3	1
d'_9	0	4	0

$\mathbf{D}'_H$

	(w'_1)	(w'_2)	(w'_3)
d'_1	2.1413	1.3333	2.2327
d'_2	1.6818	0.9411	3.2955
d'_3	1.0000	0.0105	0.4027
d'_4	5.5524	1.5217	4.4051
d'_5	0.1360	2.3192	1.3006
d'_6	0.1483	5.1362	2.3949
d'_7	0.5667	10.0829	4.4950
d'_8	0.1420	3.2446	1.7637
d'_9	0.0126	3.7831	1.2625

$\mathbf{D}'_S$

	(w'_1)	(w'_2)	(w'_3)
d'_1	2.0283	1.0667	1.2465
d'_2	1.1364	0.1882	3.0591
d'_3	1.0000	0.0021	0.0805
d'_4	5.1105	1.1043	2.4810
d'_5	0.0272	2.0638	1.0601
d'_6	0.0297	5.0272	1.2790
d'_7	0.1133	10.0166	2.4990
d'_8	0.0284	3.0489	1.1527
d'_9	0.0025	3.9566	0.2525

$\mathbf{D}'_M$

.

V. TEXT CATEZORIZATION

Text categorization is used to automatically assign previously unseen documents to a predefined set of categories. This paper gives a short introduction into text categorization (TC), and describes the most important tasks of a text categorization system. It also focuses on Support Vector Machines (SVMs), [10][11][12]the most popular machine learning algorithm used for TC, and gives some justification why SVMs are suitable for this task.

Our text categorization approach depends on representing the text document as a projection on word clusters, then applying the weka tool also we can build the text classifiers. Since we compare our approach with SVM on the same experimental settings, we include, in this section, some details about SVM and combining SVM with distributional clustering for TC. The Support Vector Machine (SVM) is an inductive learning scheme for solving the two-class pattern recognition problem. Recently SVMs have been shown to give good results for text categorization (Joachims, 1998, Dumais et al., 1998)[13][14]. The method is defined over a vector space where the classification problem is to find the decision surface that "best" separates the data points of one class from the other. SVMs can handle with exponentially or even infinitely many features, because it does not have to represent examples in that transformed space, the only thing that needs to be computed efficiently is the similarity of two examples. Redundant features (that can be predicted from another features), and high dimension are well-handled, i.e. SVM does not need an aggressive feature selection. Text categorization systems may make mistakes. We want to compare different text classifiers to decide which one is better, that is why performance measures are for. Some of them measure the performance on one binary category; others aggregate per-category measures, to give an overall performance. Denote TP, FP, TN, FN the number of true/false positives/negatives. A Support Vector Machine (SVM) performs classification by constructing a *k*-dimensional hyper plane that optimally separates the data into exactly two categories. First, let's look at a 2-dimensional example. Assume our training data, consisting of two features, has a categorical target variable with two categories {Category*1*, Category*2*}, represented by plus and circles as shown in the figure 2. The figure 3 shows about the best features extracted from document set by using the svm.so ,the svm are very useful for the text categorization.In paper we have used the svm for the text categorization.the others can use the others text categorization methods for the textualdata categorization.

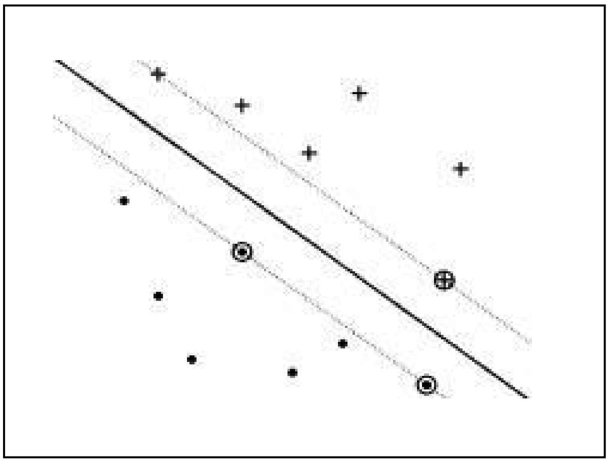


Fig. 2

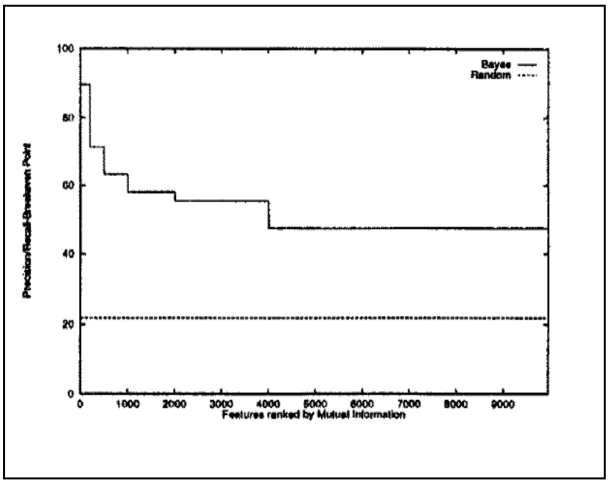


Fig. 3

TABLE 6

Sampled Execution Times (Seconds) of Different Methods on 20 Newsgroups Data

♯ of extracted features	20	58	84	120	203	280	521	1187	1453
threshold (ρ)	(0.01)	(0.02)	(0.03)	(0.06)	(0.12)	(0.19)	(0.23)	(0.32)	(0.36)
IG	19.98	19.98	19.98	19.98	19.98	19.98	19.98	19.98	19.98
DC	88.38	204.90	293.98	486.57	704.24	972.69	1973.3	3425.04	5012.79
IOC	6943.40	19397.91	28098.05	39243.00	67513.52	93010.60	———	———	———
FFC	8.61	13.95	17.68	23.44	39.36	55.30	79.79	155.99	185.24

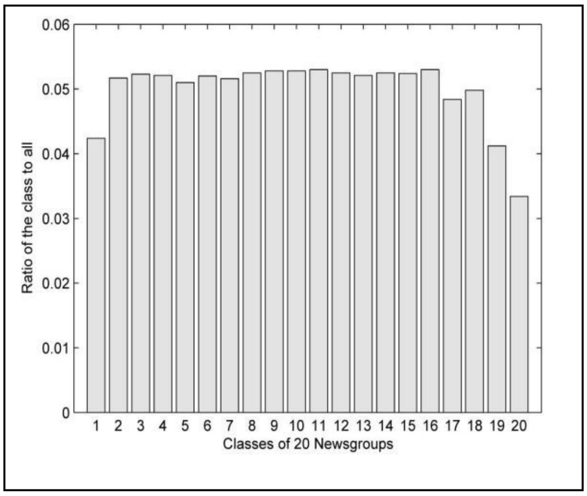


Fig. 4

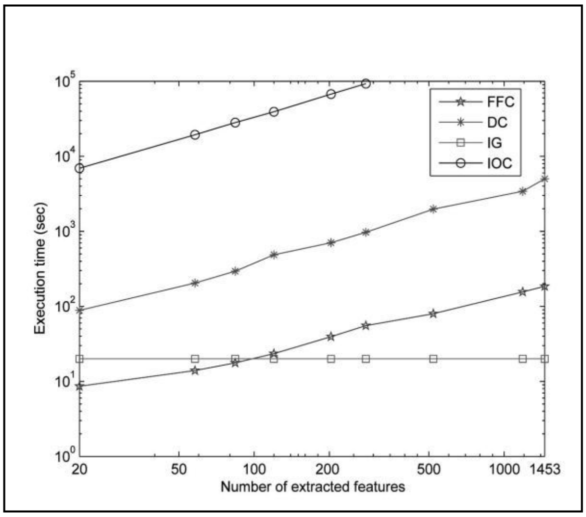


Fig. 5

EXPERIMENT

20 Newsgroups Data Set: The 20 Newsgroups collection contains about 20,000 articles taken from the Usenet newsgroups. These articles are evenly distributed over 20 classes, and each class has about 1,000 articles, as shown in Fig.4. In this figure, the x-axis indicates the class number, and the y-axis indicates the fraction of the articles of each class. We use two-thirds of the documents for training and the rest for testing. After preprocessing, we have 25,718 features, or words, for this data set. Fig. 4 shows the execution time (sec) of different feature reduction methods on the 20 Newsgroups data set. Since HFFC, S-FFC, and M-FFC have the same clustering phase, they have the same execution time, and thus we use FFC to denote them in Fig. 5. In this figure, the horizontal axis indicates the number of extracted features. To obtain different numbers of extracted features, different values of _ are used in FFC. The number of extracted features is identical to the number of clusters. For the other methods, the number of extracted features should be specified in advance by the user. Table 6 lists values of certain points in Fig. 5. Different values of _ are used in FFC and are listed in the table. Note that for IG, each word is given a weight. The words of top k weights in W are selected as the extracted features inW0. Therefore, the execution time is basically the same for any value of k. obviously, our method runs much faster than DC and IOC. For example, our method needs 8.61 seconds for 20 extracted features, while DC requires 88.38 seconds and IOC requires 6,943.40 seconds. For 84 extracted features, our method only needs 17.68 seconds, but DC and IOC require 293.98 and 28,098.05 seconds, respectively. As the number of extracted features increases, DC and IOC run significantly slow. In particular, when the number of extracted features exceeds 280, IOC spends more than 100,000 seconds without getting finished, as indicated by dashes in Table 6.

VI. CONCLUSION

Probabilistic -based clustering is one of the techniques we have developed in our machine learning research. In this paper, we apply this clustering technique to text categorization problems. We are also applying it to other problems, such as image segmentation, data sampling, fuzzy modeling, web mining, etc. The work of this paper was motivated by distributional word clustering proposed in the previous research work. In our approach we consider the document set and the words in the document set are processed and the feature vector is constructed..To that our clustering approach is applied and features are extracted and desired number of clusters are formed which are useful in the categorization of the textual data.

Our method is good for text categorization problems due to the suitability of the distributional word clustering concept. This approach is most appropriate for avoiding the arbitrariness of assigning an object to only one cluster when it may be close to several.

VII. ACKNOWLEDGEMENT

We are grateful to the authors of the previous researchers, who are helpful in improving the quality and presentation of the paper by making the understanding of the previous work done. We are to glad to Grateful to our guide who has guided us to do this work successful.

REFERENCES

[1] A.L. Blum and P. Langley, "Selection of Relevant Features and Examples in Machine Learning," Artificial Intelligence, vol. 97, nos. 1/2, pp. 245-271, 1997.

[2] E.F. Combarro, E. Montan˜ e´s, I. Dı´az, J. Ranilla, and R. Mones, "Introducing a Family of Linear Measures for Feature Selection in Text Categorization," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 9, pp. 1223-1232, Sept. 2005.

[3] R. Kohavi and G. John, "Wrappers for Feature Subset Selection," Aritficial Intelligence, vol. 97, no. 1-2, pp. 273-324, 1997.

[4] Y. Yang and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proc. 14th Int'l Conf. Machine Learning, pp. 412-420, 1997.

[5] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.

[6] L.D. Baker and A. McCallum, "Distributional Clustering of Words for Text Classification," Proc. ACM SIGIR, pp. 96-103, 1998.

[7] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "Distributional Word Clusters versus Words for Text Categorization," J. Machine Learning Research, vol. 3, pp. 1183-1208, 2003.

[8]] I.S. Dhillon, S. Mallela, and R. Kumar, "A Divisive Infomation- Theoretic Feature Clustering Algorithm for Text Classification," J. Machine Learning Research, vol. 3, pp. 1265-1287, 2003.

[9] D. Ienco and R. Meo, "Exploration and Reduction of the Feature Space by Hierarchical Clustering," Proc. SIAM Conf. Data Mining, pp. 577-587, 2008.

[10] T. Joachims, "Text Categorization with Support Vector Machine: Learning with Many Relevant Features," Technical Report LS-8-23, Univ. of Dortmund, 1998.

[11] C. Cortes and V. Vapnik, "Support-Vector Network," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[12] B. Scho¨lkopf and A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, 2001.

[13] H. Kim, P. Howland, and H. Park, "Dimension Reduction in Text Classification with Support Vector Machines," J. Machine Learning Research, vol. 6, pp. 37-53, 2005.

[14] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

◈◈◈

Hop-by-hop Congestion Control over a Wireless Multi-hop Network

Swapna Priya Lanka & Poorna Satyanarayana. B

Computer Science and Engineering Department, Jawaharlal Nehru Technological University,
Chaitanya Engineering College, Visakhapatnam, India
E-mail : swapnapriya.lanka9@gmail.com, poornasatyanarayana@gmail.com

Abstract - This paper focuses on congestion control over multihop, wireless networks. In a wireless network, an important constraint that arises is that due to the MAC (Media Access Control) layer. Many wireless MACs use a time-division strategy for channel access, where, at any point in space, the physical channel can be accessed by a single user at each instant of time. In this paper, we develop a fair hop-by-hop congestion control algorithm with the MAC constraint being imposed in the form of a channel access time constraint, using an optimization based framework. In the absence of delay, we show that this algorithm are globally stable using a Lyapunov function based approach. Next, in the presence of delay, we show that the hop-by-hop control algorithm has the property of spatial spreading. In other words, focused loads at a particular spatial location in the network get "smoothed" over space. We derive bounds on the "peak load" at a node, both with hop-by-hop control, as well as with end-to-end control, show that significant gains are to be had with the hop-by-hop scheme, and validate the analytical results with simulation.

Keywords - *Control theory, Mathematical programming/optimization.*

I. INTRODUCTION

The main objective of the project is to control over congestion in a wireless multihop networks A Joint Congestion Control and Routing Scheme to Exploit Path Diversity in the Internet focuses on congestion control over multihop, wireless networks. In a wireless network, an important constraint that arises is that due to the MAC (Media Access Control) layer. Many wireless Macs use a time-division strategy for channel access, where, at any point in space, the physical channel can be accessed by a single user at each instant of time.We develop a fair hop-by-hop congestion control algorithm with the MAC constraint being imposed in the form of a channel access time constraint, using an optimization-based framework. . Next, in the presence of delay, we show that the hop-by-hop control algorithm has the property of spatial spreading. In other words, focused loads at a particular spatial location in the network get "smoothed" over space. We derive bounds on the "peak load" at a node, both with hop-by-hop control, as well as with end-to-end control, show that significant gains are to be had with the hop-by-hop scheme.

II. SYSTEM MODEL

User Interface:

In this, we are going to create an Interface where the user can interact with the system by placing different types of hosts and types of routers. Here firstly, one host will be selected i.e., the packet sending host and another host at another end is selected i.e., packet receiving host. Now the Routers are placed in between the host. These Routers, we can place as many as we want. Now, the relation button is clicked and we have establish a relation between sender and receiver via the routers. Here we have to care that all the Routers should be covered and the relation should be unidirectional. And also we can properties for each host, i.e, the receiving host and sending host. The system should maintain organizational details. The network provides reliable and cost effective communication. The system should store and communicate data. To control the congestion occurred during the communication This paper is to control over congestion in a wireless multihop networks. Actually we are using TCP/IP protocol so internally it is connection oriented because we are using TCP protocol but on overall view it is connection less because we are using IP protocol. Hop by hop schemes provide feedback about the congestion state at a node to the hop preceding it. By this feedback only the preceding node then adapts its transmission rate. The feedback is provided based on the queue length at the congested node. If the queue length exceeds the threshold, congestion is indicated and the preceding node is indicated to decrease its preceding rate.

Control Theory:

Congestion control concerns controlling traffic entry into a telecommunications network, so as to avoid congestive collapse by attempting to avoid oversubscription of any of the processing or link capabilities of the intermediate nodes and networks and taking resource reducing steps, such as reducing the rate of sending packets.

The modern theory of congestion control was pioneered by Frank Kelly, who applied microeconomic theory and convex optimization theory to describe how individuals controlling their own rates can interact to achieve an "optimal" network-wide rate allocation Examples of "optimal" rate allocation are max-min fair allocation and Kelly's suggestion of proportional fair allocation, although many others are possible.

The mathematical expression for optimal rate allocation is as follows. Let x_i be the rate of flow i, C_l be the capacity of link l, and r_{li} be 1 if flow i uses link l and 0 otherwise. Let x, c and R be the corresponding vectors and matrix. Let $U(x)$ be an increasing, strictly convex function, called the utility, which measures how much benefit a user obtains by transmitting at rate x. The optimal rate allocation then satisfies

$$\max_{x} \sum_{i} U(x_i)$$

Such that $Rx \leq c$

The Lagrange dual of this problem decouples, so that each flow sets its own rate, based only on a "price" signaled by the network. Each link capacity imposes a constraint, which gives rise to a Lagrange multiplier, p_l.

The sum of these Lagrange multipliers, is the price to which the flow responds.

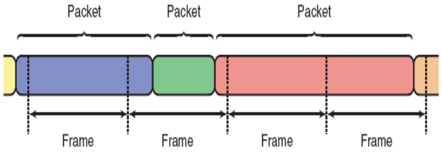
$$y_i = \sum_{l} p_l r_{li},$$

Congestion control then becomes a distributed optimization algorithm for solving the above problem. Many current congestion control algorithms can be modeled in this framework, with p_l being either the loss probability or the queuing delay at link l.

A major weakness of this model is that it assumes all flows observe the same price, while sliding window flow control causes "burstiness" which causes different flows to observe different loss or delay at a given link. ‘n’ information technology, a packet is a formatted block of data carried by a packet mode computer network. Computer communications links that do not support packets, such as traditional point-to-point telecommunications links, simply transmit data as a series of bytes, characters, or bits alone. When data is formatted into a packet, the network can transmit long messages more efficiently and reliably. In general, the term packet applies to any message formatted as a packet, while the term datagram is generally reserved for the packets of an unreliable service. A reliable service is one where the user is notified if delivery fails. An unreliable one is where the user is not notified if delivery fails.

For example, IP provides an unreliable service. TCP uses IP to provide a reliable service, whereas UDP uses IP to provide an unreliable one. All these protocols use packets, but UDP packets are generally called data grams.

When the ARPANET pioneered packet switching, it provided a reliable packet delivery procedure to its connected hosts via its 1822 interface. A host computer simply arranged the data in the correct packet format, inserted the address of the destination host computer, and sent the message across the interface to its connected IMP. Once the message was delivered to the destination host, an acknowledgement was delivered to the sending host. If the network could not deliver the message, it would send an error message back to the sending host. Meanwhile, the developers of CYCLADES and of ALOHA net demonstrated that it was possible to build an effective computer network without providing reliable packet transmission. This lesson was later embraced by the designers of Ethernet. If a network does not guarantee packet delivery, then it becomes the host's responsibility to provide reliability by detecting and retransmitting lost packets. Subsequent experience on the ARPANET indicated that the network itself could not reliably detect all packet delivery failures, and this pushed responsibility for error detection onto the sending host in any case. This led to the development of the end-to-end principle, which is one of the Internet's fundamental design assumptions.



.

Mathematical programming/optimization:

Optimization theory seeks to discover the means to find points where (real-valued) functions take on maximal or minimal values. (Vector-valued functions require multi-objective programming, and are almost always reduced to real-valued functions by weighting.) We consider both the basic theory and questions regarding the encoding of the tools developed into software. Topics in optimization also appear in Calculus of variation (typically seeking functions, curves, or other geometric objects which are optimal in some way); global analysis; and operations research (typically seeking choices of parameters to optimize some simple multivariate function). Those areas tend to emphasize the theory and application of optimization rather than the computational issues involved. Calculus of variations and optimization seek functions or geometric objects which are optimize some objective function. Certainly this includes a discussion of techniques to find the optima, such as successive approximations or linear programming. In addition, there is quite a lot of work establishing the existence of optima and characterizing them. In many cases, optimal functions or curves can be expressed as solutions to differential equations.

Common applications include seeking curves and surfaces which are minimal in some sense. However, the spaces on which the analysis is done may represent configurations of some physical system, say, so that this field also applies to optimization problems in economics or control theory for example. Operations research may be loosely described as the study of optimal resource allocation. Mathematically, this is the study of optimization. Depending on the options and constrain in the setting, this may involve linear programming, or quadratic-, convex-, integer-, or Boolean-programming.

Hop-By-Hop Transmission

First, congestion controller at the source of each session reacts based on the sum of the congestion prices at each node. Here each node passes the feedback (partial sum) price upstream. Each node adds its current congestion cost to that it received from a downstream node, and passes this information toward the upstream node. The source will ultimately receive the sum of all price information from the corresponding downstream nodes and use the information for controlling rates.

The basic idea of a hop-by-hop algorithm is that every node in the path of the session operates a congestion control algorithm

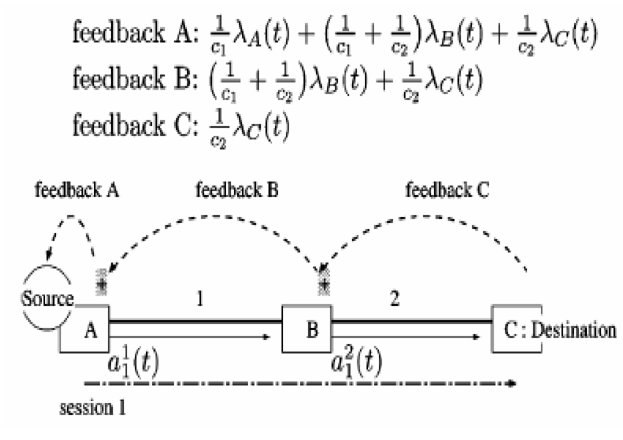

feedback A: $\frac{1}{c_1}\lambda_A(t) + (\frac{1}{c_1} + \frac{1}{c_2})\lambda_B(t) + \frac{1}{c_2}\lambda_C(t)$
feedback B: $(\frac{1}{c_1} + \frac{1}{c_2})\lambda_B(t) + \frac{1}{c_2}\lambda_C(t)$
feedback C: $\frac{1}{c_2}\lambda_C(t)$
feedback A
feedback B
feedback C
Source
A
1
B
2
C : Destination
$a_1^1(t)$
$a_1^2(t)$
session 1

Fig: Hop-by-Hop Congestion Control Algorithm

Using this "price passing" method, the source of session 1 receives aggregate congestion price from its downstream nodes and controls its transmission rate based on it.

Let us denote as the actual transmission rate at the 'i'th hop of session 'r' in the hop-by-hop control algorithm. Corresponding to each node 'i' along the path of session is a virtual transmission rate, which is described by

$$\dot{c}_r^i(t) = \kappa\left(w_r - a_r^i(t) \sum_{j\in A_k(r)} \left(\frac{1}{c_{l_i(j,r)}} + \frac{1}{c_{l_o(j,r)}}\right)\lambda_j(t)\right)$$

$$a_r^i(t) = \min\left[c_r^i(t), a_r^{i-1}(t)\right]$$

Implementation is the process of bringing the developed system into operational use and turning it over to the user. The implementation of computer based system requires that test be prepared and that the system and its elements be tested in planned and structured manner Implementation is the process of assuring that the information system is operational and then allowing users take over its operation for use and evaluation.

We can implement many activities by using this algorithm.

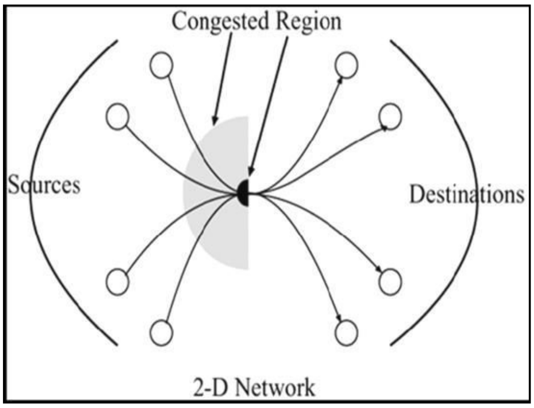
In the above algorithm, we sum over all prices downstream along session r . Thus, each node operates a (perflow) controller based on the perceived congestion due to *downstream nodes*, and determines the maximum rate it can transmit at (the virtual transmission rate). The actual rate it chooses transmits at the rate of the minimum of the *incoming* data rate from i -1 th hop node in the session's path (the previous hop node),

i.e, $a^{i-1}{}_r$(t); and the maximum possible rate ci_r(t).

We comment that at each intermediate node, the controller has knowledge of the local link rates, as well as the "rampup" constant wr for each of the sessions that is incident on the node. It can be shown that the stability analysis and later analysis are valid even if the node uses an upper bound on the ramp-up constant. Thus, from an implementation perspective, one could assume that fwrg are globally bounded by some value w; and use this value at each intermediate node. Heuristically, the convergence proofs are valid even when a bound is used because the data transmission rate into the network is ultimately governed by the source, which will use the correct value of wr: However, to keep the exposition simple, we will use the exact value of wr at each node in this paper.

III. SPATIAL SPREADING

In this section, we derive the peak occupied buffer size with the end-to-end controller as well as with the hop-by-hop controller described in Section VI. We consider the evolution of these algorithms in the presence of propagation delay. We analytically show the effect of spatial spreading by explicitly deriving the reduction in peak buffer overload under the hop-by-hop scheme. Consider the tree network in Figure 4, with N sessions and L links between each of the sources and the common destination. Such a network could model a community roof-top wireless network, with the common node being connected to a wired infrastructure. The source node for each session resides on a (different) node as shown in Figure .



We assume that each link has a round-trip delay of d; and the corresponding end-to-end delay for the session being D = Ld: We assume that the intermediate links (each accessed by only one flow) are well provisioned so that congestion occurs only at the common access point for all the flows (the bottleneck node in Figure 4). Since we consider a system with N flows, we scale the capacities of the bottleneck node with the input and output capacities of the bottleneck node being Nc_I and Nc_O respectively.

This scaling ensures that the *steady-state rate allocated to each user is invariant with the number of sessions.* Physically, this would correspond to a bandwidth scaling at the bottleneck.

IV. SIMULATION RESULTS

In this section, we present simulation results that compare the hop-by-hop algorithm with the end-to-end algorithm. Through both fluid and packet simulation, we show that there is a significant decrease in the peak load with the hop-by-hop algorithm.

V. CONCLUSION

We proposed an algorithm to transfer packets from source to destination without congestion i.e., congestion control over a hops more than one. The result obtains show that there is a significant decrease in the peak load with the hop-by-hop algorithm. We develop a fair hop-by-hop congestion control algorithm with the MAC constraint being imposed in the form of a channel access time constraint, using an optimization-based framework.

REFERENCES

[1] G. Holland and N. H. Vaidya, “Analysis of TCP performance over mobile ad hoc networks,” Proc. IEEE/ACM Mobicom, pp. 219–23 august.1999.

[2] F. P. Kelly, A. Maulloo, and D. Tan, “Rate control in communication networks: Shadow prices, proportional fairness and stability,”

[3] S. Kunniyur and R. Srikant, “End-to-end congestion control: Utility functions, random losses and ECN marks,”

◈◈◈

Record Matching for Web Databases by Domain-specific Query Probing

V.Alekhya & Ds.Bhupal Naik

CSE Department, Vignan university
E-mail : ale.vemulapalli1@gmail.com, dsbhupal@gmail.com

Abstract - Record matching refers to the task of finding entries that refer to the same entity in two or more files, is a vital process in data integration. Most of the record matching methods are supervised, which requires the user to provide training data. These methods are not applicable for web database scenario, where query results dynamically generated on-the- fly. To address the problem of record matching in the Web database scenario, we present an unsupervised, online record matching method, UDD, which effectively identifies the duplicates from query result records of multiple web databases. First, same source duplicates are eliminated by using exact matching method the "presumed" non duplicate records from the same source can be used as training examples . Starting from the non duplicate set, we use two cooperating classifiers a weight component similarity summing classifier and an SVM classifier, to iteratively identify duplicates in the query results from multiple Web databases.

Keywords - *Record Matching, duplicate detection, record linkage, data deduplication, SVM .*

I. INTRODUCTION

Today, more and more databases that dynamically generate Web pages in response to user queries are available on the Web. These Web databases compose the deep or hidden Web, which is estimated to contain a much larger amount of high quality, usually structured information and to have a faster growth rate than the static Web. Most Web databases are only accessible via a query interface through which users can submit queries. Once a query is received, the Web server will retrieve the corresponding results from the back-end database and return them to the user. To build a system that helps users integrate and, more importantly, compare the query results returned from multiple Web databases, a crucial task is to match the different sources' records that refer to the same real-world entity. The problem of identifying duplicates, that is, two (or more) records describing the same entity, has attracted much attention from many research fields, including Databases, Data Mining, Artificial Intelligence, and Natural Language Processing. Most previous work is based on predefined matching rules hand-coded by domain experts or matching rules learned offline by some learning method from a set of training examples. Such approaches work well in a traditional database environment, where all instances of the target databases can be readily accessed, as long as a set of high-quality representative records can be examined by experts or selected for the user to label.. Consequently, hand-coding or offline-learning approaches are not appropriate in web database scenarios for two reasons. First, the full data set is not available beforehand, and therefore, good representative data for training are hard to obtain. Second, and most importantly, even if good representative data are found and labeled for learning, the rules learned on the representatives of a full data set may not work well on a partial and biased part of that data set. problem Definition Our focus is on Web databases from the same domain, i.e., Web databases that provide the same type of records in response to user queries.

II. BACKGROUND WORK

Most record matching methods adopt a framework that uses two major steps([5]&[6]):

1. Identifying a similarity function : Using training examples (i.e., manually labeled duplicate and non duplicate records) and a set of predefined basis similarity measures/functions over numeric and/or string fields, a single composite similarity function over one pair of records, which is a weighted combination (often linear) of the basis functions, is identified by domain experts or learned by a learning method, such as Expectation-Maximization, decision tree, Bayesian network, or SVM ([1],[2],[3]&[4])

2. Matching records : The composite similarity function is used to calculate the similarity between the candidate pairs and highly similar pairs are matched and identified as referring to the same entity.

Problem Definition*:* Suppose there are s records in data source A and there are t records in data source B, with each record having a set of fields/attributes. Each of the t records in data source B can potentially be a duplicate of each of the s records in data source A. The goal of duplicate detection is to determine the matching status, i.e., duplicate or non duplicate, of these s Â t record pairs.

An important aspect of duplicate detection is to reduce the number of record pair comparisons. Several methods have been proposed for this purpose including standard blocking[9] sorted neighborhood method Bigram Indexing, and record clustering[1] . Even though these methods differ in how to partition the data set into blocks, they all considerably reduce the number of comparisons by only comparing records from the same block. Since any of these methods can be incorporated into UDD to reduce the number of record pair comparisons, we do not further consider this issue. While most previous record matching work is targeted at matching a single type of record, more recent work has addressed the matching of multiple types of records with rich associations \between the records. Even though the matching complexity increases rapidly with the number of record types, these works manage to capture the matching dependencies between multiple record types and utilize such dependencies to improve the matching accuracy of each single record type. Unfortunately, however, the dependencies among multiple record types are not available for many domains. Compared to these previous works, UDD is specifically designed for the Web database scenario where the records to match are of a single type with multiple string fields. These records are heavily query-dependent and are only a partial and biased portion of the entire data, which makes the existing work based on offline learning inappropriate. Moreover, our work focuses on studying and addressing the field weight assignment issue rather than on the similarity measure. In UDD, any similarity measure, or some combination of them, can be easily incorporated.

III. UDD METHOD

To overcome such problems, we propose a new record matching method Unsupervised Duplicate Detection (UDD) for the specific record matching problem of identifying duplicates among records in query results from multiple Web databases.

The key ideas of our method are: We focus on techniques for adjusting the weights of the record fields in calculating the similarity between two records.

1. Two records are considered as duplicates if they are "similar enough" on their fields. As illustrated by the previous example, we believe different fields may need to be assigned different importance weights in an adaptive and dynamic manner.

2. Due to the absence of labeled training examples, we use a sample of universal data consisting of record pairs from different data sources as an approximation for a negative training set as well as the record pairs from the same data source. We believe, and our experimental results verify, that doing so is reason- able since the proportion of duplicate records in the universal set is usually much smaller than the proportion of non duplicates.

Employing two classifiers that collaborate in an iterative manner, UDD identifies duplicates as follows: First, each field's weight is set according to its "relative distance," i.e., dissimilarity, among records from the approximated negative training set. Then, the first classifier, which utilizes the weights set in the first step, is used to match records from different data sources. Next, with the matched records being a positive set and the non duplicate records in the negative set, the second classifier further identifies new duplicates. Finally, all the identified duplicates and non- duplicates are used to adjust the field weights set in the first step and a new iteration begins by again employing the first classifier to identify new duplicates. The iteration stops when no new duplicates can be identified.

C1 —Weighted Component Similarity Summing (WCSS) Classifier

In our algorithm, classifier C1 plays a vital role. At the beginning, it is used to identify some duplicate vectors when there are no positive examples available. Then, after iteration begins, it is used again to cooperate with C2 to identify new duplicate vectors. Because no duplicate vectors are available initially, classifiers that need class information to train, such as decision tree and NaıveBayes, cannot be used. An intuitive method to identify duplicate vectors is to assume that two records are duplicates if most of their fields that are under consideration are similar. On the other hand, if all corresponding fields of the two records are dissimilar, it is unlikely that the two records are duplicates. To evaluate the similarity between two records, we combine the values of each component in the similarity vector for the two records. Different fields may have different importance when we decide whether two records are duplicates. The importance is usually data-dependent, which, in turn, depends on the query in the Web database scenario.

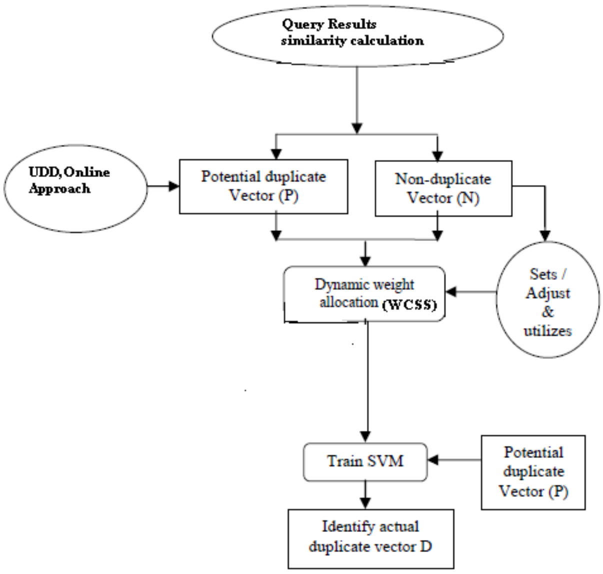
Hence, we define the similarity between records r1 and r2

$$Sim(r_1,r_2) = \sum_{i=1}^{n} w_i \cdot v_i,$$

where

$$\sum_{i=1}^{n} w_i = 1$$

System Architecture


Query Results similarity calculation
UDD, Online Approach
Potential duplicate Vector (P)
Non-duplicate Vector (N)
Dynamic weight allocation (WCSS)
Sets / Adjust & utilizes
Train SVM
Potential duplicate Vector (P)
Identify actual duplicate vector D

and wi €[0,1] is the weight for the ith similarity component which represents the importance of the ith field. The similarity Sim(r1, r2) between records r1 and r2 will be in [0,1] according to the above definition..Duplicate Identification After we assign a weight for each component, the duplicate vector detection is rather intuitive. Two records r1 and r2 are duplicates if Sim(r1 , r2) Tsim , i.e., if their similarity value is equal to or greater than a similarity threshold. In general, the similarity threshold Tsim should be close to 1 to ensure that the identified duplicates are correct. Increasing the value of Tsim will reduce the number of duplicate vectors identified by C1 while, at the same time, the identified duplicates will be more precise. C2 —Support Vector Machine Classifier After detecting a few duplicate vectors whose similarity scores are bigger than the threshold using the WCSS classifier, we have positive examples, the identified duplicate vectors in D, and negative examples, namely, the remaining no duplicate vectors in N 0 . Hence, we can train another classifier C2 and use this trained classifier to identify new duplicate vectors from the remaining potential duplicate vectors in P and the no duplicate vectors in N . A classifier suitable for the task should have the following characteristics. First, it should not be sensitive to the relative size of the positive and negative examples because the size of the negative examples is usually much bigger than the size of the positive examples. This is especially the case at the beginning of the duplicate vector detection iterations when a limited number of duplicates are detected. Another requirement is that the classifier should work well given limited training examples. Because our algorithm identifies duplicate vectors in an iterative way, any incorrect identification due to noise during the first several iterations, when the number of positive examples is limited, will greatly affect the final result.

Evaluation Metric : As in many other duplicate detection approaches, we report the overall performance using recall and precision, which are defined as follows:

$$precision = \frac{\#of\ Correctly\ Identified\ Duplicate\ Pairs}{\#of\ All\ Identified\ Duplicate\ Pairs}$$

$$recall = \frac{\#of\ Correctly\ Identified\ Duplicate\ Pairs}{\#of\ True\ Duplicate\ Pairs},$$

However, as indicated in [10], due to the usually imbalanced distribution of matches and non matches in the weight vector set, these commonly used accuracy measures are not very suitable for assessing the quality of record matching. The large number of no matches usually dominates the accuracy measure and yields results that are too optimistic. Thus, we also use the F-measure, which is the harmonic mean of precision and recall, to evaluate the classification quality [2]:

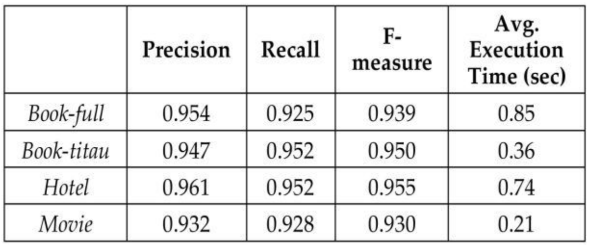
$$F - measure = \frac{2 \cdot precision \cdot recall}{(precision + recall)}.$$

IV. EXPERIMENTAL RESULTS

We ran UDD on the Cora data set and its three subsets individually when the similarity threshold Tsim =0:85.Although Cora is a noisy data set, our algorithm still performs well over it. UDD has a precision of 0.896, recall of 0.950, and F-measure of 0.923 over the Cora data set. We compared our results with other works that use all or part of the Cora data set. Bilenko and Mooney [1], in which a subset of the Cora data set is used, report an F-measure of 0.867. Cohen and Richman [2] report 0.99/0.925 for precision/recall using a subset of the Cora data set. Culotta and McCallum [7] report an F-measure of 0.908 using the full Cora data set. From this comparison, it can be seen that the performance of UDD is comparable to these methods, all of which require training examples.

Effect of the threshold Tsim. The iteration row in the below table indicates the number of iterations required for UDD to stop. It can be seen that the duplicate vector detection iterations stop very quickly. All of them stop by the fifth iteration. On the one hand, the smaller Tsim is, the more iterations are required and the higher the recall. This is because the WCSS classifier with smaller Tsim identifies more duplicates and most of them are correct duplicates. Hence, with more correct positive examples, the SVM classifier can also identify more duplicates, which, in turn, results in a higher recall. On the other hand, the smaller Tsim is, the lower the precision. This is because the WCSS classifier with smaller Tsim is more likely to identify incorrect duplicates, which may incorrectly guide the SVM classifier to identify new incorrect duplicates. In our experiments, the highest F- measures were achieved when Tsim =0:85 over all the four data sets

Performance of UDD on the Web Database Data Sets



	Precision	Recall	F-measure	Avg. Execution Time (sec)
Book-full	0.954	0.925	0.939	0.85
Book-titau	0.947	0.952	0.950	0.36
Hotel	0.961	0.952	0.955	0.74
Movie	0.932	0.928	0.930	0.21

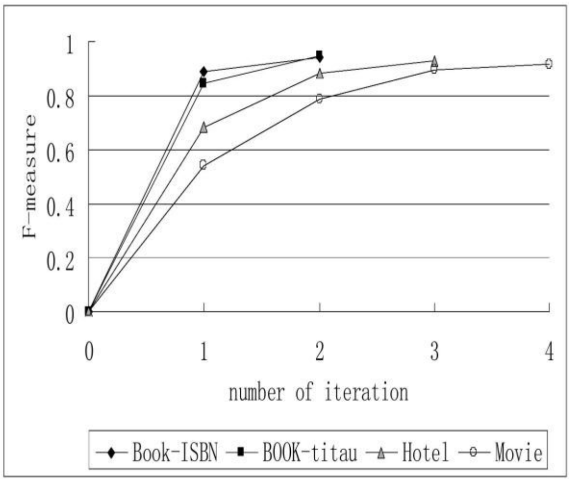
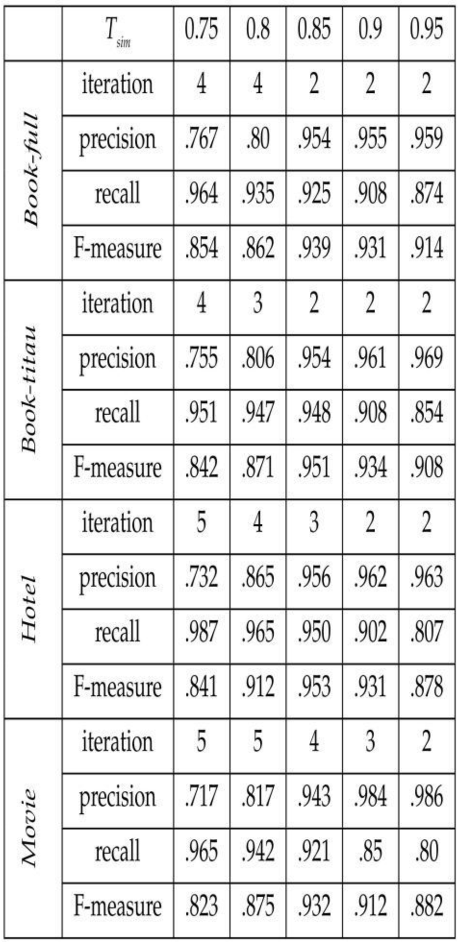


Fig. 1 : Performance of UDD at the end of each Duplicate detection iteration over the four Datasets when T_{sim}=0.85



	T_{sim}	0.75	0.8	0.85	0.9	0.95
Book-full	iteration	4	4	2	2	2
	precision	.767	.80	.954	.955	.959
	recall	.964	.935	.925	.908	.874
	F-measure	.854	.862	.939	.931	.914
Book-titau	iteration	4	3	2	2	2
	precision	.755	.806	.954	.961	.969
	recall	.951	.947	.948	.908	.854
	F-measure	.842	.871	.951	.934	.908
Hotel	iteration	5	4	3	2	2
	precision	.732	.865	.956	.962	.963
	recall	.987	.965	.950	.902	.807
	F-measure	.841	.912	.953	.931	.878
Movie	iteration	5	5	4	3	2
	precision	.717	.817	.943	.984	.986
	recall	.965	.942	.921	.85	.80
	F-measure	.823	.875	.932	.912	.882

ig. 2 : Performance of UDD with Different Tsim on the Web Database Data Sets

Effect of the number of iterations:

The above figure 1 shows the performance of UDD at the end of each duplicate detection iteration over the four Web database data sets when Tsim = 0:85. It can be seen that the iteration stops quickly for all data sets and takes at most four iterations.

Figure 2 shows UDD's performance when using five different similarity thresholds (Tsim: 0.75, 0.80, 0.85, 0.90,and 0.95) on the four Web database data sets. The iteration row in this table indicates the number of iterations required for UDD to stop. It can be seen that the duplicate vector detection iterations stop very quickly. All of them stop by the fifth iteration. On the one hand, the smaller Tsim is, the more iterations are required and the higher the recall. This is because the WCSS classifier with smaller Tsim identifies more duplicates and most of them are correct duplicates. Hence, with more correct positive examples, the SVM classifier can also identify more duplicates, which, in turn, results in a higher recall. On the other hand, the experiments, the highest F-measures were achieved when Tsim = 0:85 over all the four data sets. smaller Tsim is, the lower the precision. This is because the WCSS classifier with smaller Tsim is more likely to identify incorrect duplicates, which may incorrectly guide the SVM classifier to identify new incorrect duplicates. In our smaller compared with other fields. Thus, they gain larger weights. In turn, the vectors staying in N are even more unlikely to be identified as duplicates in the next iteration because of the larger weights on their fields with small similarity values. Consequently, a high Tsim value makes it difficult for the WCSS classifier to find new positive instances after the first two iterations.

We also observe from figure 2 that the iterations stop more quickly when the threshold Tsim is high. When Tsim is high, vectors are required to have more large similarity values on their fields in order to be identified as duplicates in the early iterations. Hence, vectors with only a certain number of fields having large similarity values and other fields having small similarity values are likely to stay in the negative example set N. Recall that, according to the nonduplicate intuition, when setting the component weights in the WCSS classifier, fields with more large similarity values in N gain smaller weights and fields with more small similarity values in N gain larger weights. The vectors that stay in N would make the fields with small similarity values on all vectors in N relatively even smaller compared with other fields. Thus, they gain larger weights. In turn, the vectors staying in N are even more unlikely to be identified as duplicates in the next iteration because of the larger weights on their fields with small similarity values. Consequently, a high Tsim value makes it difficult for the WCSS classifier to find new positive instances after the first two iterations.

Influence of duplicate records from the same data source. Recall that, in this Section , we assumed that records from the same data source are non duplicates so that we can put pairs of them into the negative example set N. However, we also pointed out that, in reality, some of these record pairs could be actual duplicates. We call the actual duplicate vectors in N false non duplicate vectors. The number of false non duplicate vectors increases as the duplicate ratio,defined in Definition 1 in this Section, increases.The false nonduplicate vectors will affect the two classifiers in our algorithm in the following ways:

1. While setting the component weights W in WCSS according to the nonduplicate intuition, components that usually have large similarity values in N will be assigned small weights. Note that the false nonduplicate vectors are actually duplicate vectors with large similarity values for their components. As a result, an inappropriate set of weights could be set for WCSS.
2. Since the SVM classifier learns by creating a hyperplane between positive and negative examples, the false nonduplicate vectors will incorrectly add positive examples in the negative space. As a result,the hyperplane could be moved toward the positive space and the number of identified duplicate vectors will be much smaller.

V. CONCLUSION

Duplicate detection is an important step in data integration and most state-of-the-art methods are based on offline learning techniques, which require training data. In the Web database scenario, where records to match are greatly query-dependent, a pertained approach is not applicable as the set of records in each query’s results is a biased subset of the full data set. To overcome this problem, we presented an unsupervised, online approach, UDD, for detecting duplicates over the query results of multiple Web databases. Two classifiers, WCSS and SVM, are used cooperatively in the convergence step of record matching to identify the duplicate pairs from all potential duplicate pairs iteratively. Experimental results show that our approach is comparable to previous work that requires training examples for identifying duplicates from the query results of multiple Web databases

ACKNOWLEDGEMENTS

This research was supported by the vignan university, guntur. Iam thankful to our guide who helpful for doing my research

REFERENCES

[1] M. Bilenko and R.J. Mooney, “Adaptive Duplicate DetectionUsing Learnable String Similarity Measures,” Proc.ACMSIGKDD,pp.39-48,2003.

[2] W.W. Cohen and J. Richman, “Learning to Match and Cluster Large High-Dimensional Datasets for Data Integration,” Proc. ACM SIGKDD,pp.475-480,2002

[3] Y. Thibaudeau, “The Discrimination Power of DependencyStructures in Record Linkage,” Survey Methodology, vol. 19,pp. 31-38, 1993.

[4] W.E. Winkler, “Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage,” Proc. Section Survey Research Methods, pp. 667-671, 1988.

[5] A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios, “Duplicate Record Detection: A Survey,” IEEE Trans. Knowledge and Data Eng.,vol. 19, no. 1, pp. 1-16, Jan. 2007.

[6] N. Koudas, S. Sarawagi, and D. Srivastava, “Record Linkage:Similarity Measures and Algorithms (Tutorial),” Proc. ACMSIGMOD, pp. 802-803, 2006.

[7] A. Culotta and A. McCallum, “A Conditional Model of Deduplication for Multi-Type Relational Data,” Technical Report IR-443, Dept. of Computer Science, Univ. of Massachusetts Amherst

[8] M.A. Hernandez and S.J. Stolfo, “The merge/Purge Problem for Large Databases,” ACM SIGMOD Record, vol. 24, no. 2, pp. 127-138, 1995.

[9] M.A. Jaro, “Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida,” J. Am. Statistical Assoc., vol. 89, no. 406, pp. 414-420, 1989.

Providing the Boundary line Controlled Request with Adaptable Transmission Rates in WDM Mesh Networks

AnushaAnnapureddy & B.Balaji

CSE Department, Vignan University, Guntur, India
E-mail : anusha508.reddy@gmail.com, balajiwgl@gmail.com

Abstract - The mixture of applications increases and supported over optical networks, to the network customers new service guarantees must be offered .The partitioning the data into multiple segments which can be processed independently the useful data to be transferred before a predefined deadline .this is a deadline driven request. To provide the request the customer chooses the bandwidth DDRs provide scheduling flexibility for the service providers. It chooses bandwidth while achieving two objectives 1.satisfying the guaranteed deadline 2.decreasing network resource utilization .by using bandwidth allocation policies improve the network performance and by using mixed integer linear program allows choosing flexible transmission rates .

Keywords - *Bandwidth-on-demand;deadline-driven request (DDr); flexible transmission rate; large data transfers; wavelength division multiplexing (WDM) network.*

I. INTRODUCTION

Now a day's telecommunications networks are trying to increase the bandwidth by their users as well as different of services they must support. Grid computing, eScience applications (bandwidth-hungry applications) these require adaptable bandwidth reservation and need strict quality-of-service (QoS) guarantees .the new application requirements is –large bandwidth, dynamism,and flexibility—can be well accommodated by optical networks using wavelength- division multiplexing. This explains about understanding problems caused due to bandwidth allocation for improving new services by telecommunication networks to provide new services like IPTV gird computing …etc which need accurate and consistent bandwidth. This can be solved by implementing wave length division multiplexing which works on optical networks. This technology uses optical cross connections and protocols like ASON/GMPLS which are used for controlling automatic and dynamic provision of light paths. In an Automatically switched optical network (ASON) the customer defines a new path by its start and end point , the bandwidth needed; the path itself is not specified by the customer.

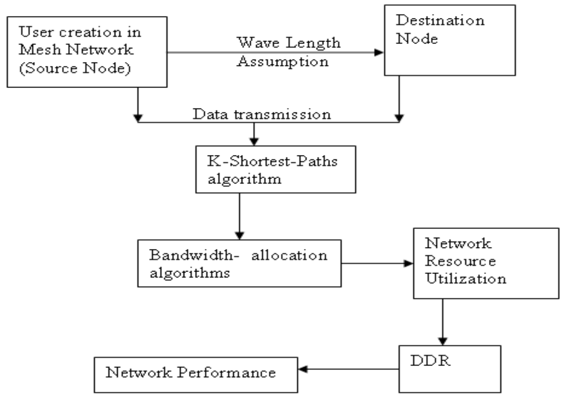
II. PROBLEM STATEMENT

Capacity measures for a network connection across the Internet are useful to many applications. Its applicability encompasses QoS guarantees, congestion control and other related areas. In this paper we define and measure the available capacity of a connection, through observations at endpoints only. Our measurements account for the variability of cross traffic that passes through the routers handling this connection. Related to the estimation of available capacity, we suggest modifications to current techniques to measure the packet service time of the "bottleneck" router of the connection. Finally, we present estimation results on wide-area network connections from our experiments to multiple sites.

III. PROPOSED SYSTEM

Hence, the service provider can choose the bandwidth (transmission rate) to provide the request. In this case, even though DDRs impose a deadline constraint, they provide scheduling flexibility for the service provider since it can choose the transmission rate while achieving two objectives: 1) satisfying the guaranteed deadline; and 2) optimizing the network's resource utilization. We investigate the problem of provisioning DDRs with flexible transmission rates in wavelength-division multiplexing (WDM) mesh networks, although this approach is generalizable to other networks also. We investigate several (fixed and adaptive to network state) bandwidth-allocation policies and study the benefit of allowing dynamic bandwidth adjustment, which is found to generally improve network performance. We show that the performance of the bandwidth-allocation algorithms depends on the DDR traffic distribution and on the node architecture and its parameters. In addition, we develop a mathematical formulation for our problem as a mixed integer linear program (MILP), which allows choosing flexible transmission rates and provides a lower for our provisioning algorithms.

IV. SYSTEM ARCHITECTURE


User creation in Mesh Network (Source Node)
Wave Length Assumption
Destination Node
Data transmission
K-Shortest-Paths algorithm
Bandwidth- allocation algorithms
Network Resource Utilization
DDR
Network Performance

Source node acts as a mesh network and wavelength is assumption takes place between the source node and destination node and data is transferring between the two nodes and transferring the data chooses the shortest path and also assigning the bandwidth allocation algorithm between the two nodes nodes gives the bandwidth how much is their and by using the network resources the data can easily reach the deadline called DDR and gives network performance. Each router and gateway within the mesh is typically connected through two or more devices ,which provides number of different path available for network communication .the route between two end devices often goes through multi “hops” communication passes through other devices to cover long distances. If an intermediary device fails is offline or busy , the message can still get through via an alternative path .mesh network look for the most efficient path available and will automatically re-route messages to avoid any failures.

V. LITERATURE SURVEY

A. Advanced Wavelength Reservation Method Based on Deadline-Aware Scheduling for Lambda Grid Networks

The increase of network technologies and high performance computing, research on grid computing is very popular using a high-performance virtual machine made by grid computing makes it possible to execute large-scale jobs. Such jobs include large-scale scientific and engineering calculations and the high-speed processing of large amounts of data this trend only reinforces the understanding that computing grid service users have different performance requirements. The fees charged to users are high if the job is to be completed in the shortest time and low otherwise. Many users are prepared to accept some delay in job completion, provided that the job is completed within a deadline, in return for a lower service fee. Different users will have different job priorities and different deadlines, so job scheduling that satisfies all job deadlines is essential. Since it is necessary to transfer all input data to job-execution nodes before job execution can commence, it is important to efficiently assign wavelengths in lambda grid networks. The conventional job scheduling approach assigns a lot of time slots to a new call in order to finish a job as fast as possible, regardless of its deadline. Hence, the probability of a short deadline call being assigned time slots is low, which raises the blocking probability of such calls.

Fig. 1 shows the basic model of a lambda grid system. Each node has a scheduler, which is called “master,” to manage the computing resources. The masters exchange information on a regular basis. This information includes the load, the computational capacity, the amount of free space of data storage, and the devices available. When a user has a job to execute, the user submits it to the local master. The local master divides the job into several sub jobs. It then schedules the sub jobs and distributes them to the remote sites via optical links. Job distribution follows the policy of job scheduling. For example, if the policy is to complete the job as rapidly as possible, the sub jobs may be distributed to one or remote sites that have high capacity. Each remote site receives the sub jobs, executes them, and returns the results to the local site. The local site combines the results into a single result and returns the result to the user. The requests for data transmission (hereafter, “calls”)requests for data transmission (hereafter, “calls”) are generated with job execution of time slots

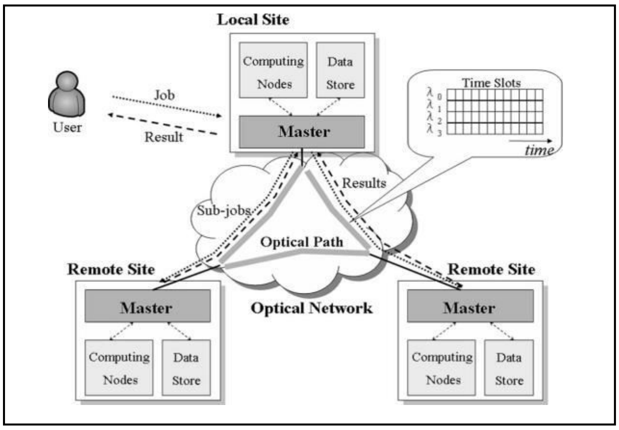


Fig. 1 : Lambda grid system based on grid computing through optical WDM paths.

Fig. 1. Lambda grid system based on grid computing through optical WDM paths Since the input data needed for job execution are geographically dispersed. The user specifies the job's deadline when accessing the local master. The computing node sends a call to the local master that includes data size, job deadline, and destination of the job-execution node. The master decides the deadline of data transmission based on the job's deadline Advanced reservation methods were introduced to the grid systems to guarantee resource availability at the time when an application was executed. Advanced multiple resource reservation methods ensure that all resources are available when needed by the application. The reservation of data transmission time slots is often needed to guarantee that job completion occurs within the deadline. For data transmission, calls that specify, among other details, data size, start time, and deadline are issued to reserve time slots. With the conventional advanced reservation method, a new request may make it necessary to reschedule the reserved times. stated that ideal resource reservation scheduling with a consideration of data size, start time, and deadline is an NP-hard problem. That is, it is not possible to design an algorithm that can always give the optimal reservation schedule. One solution is to make locally optimal decisions. The lambda grid, which employs wavelength division multiplexing(WDM) and optical paths, is an attractive candidate The WDM(wave length division multiplexing) offers large network capacity, so high speed data transfer is possible. The optical paths guarantee network availability for job-execution assurances, so data transfer is reliable. The grid environment requires that wavelength information, such as bandwidth and the utilization of wavelengths, be managed as resource information.

B. *On Traffic Grooming Choices for IP over WDM network*

Traffic grooming continues to be a rich area of research in the context of WDM optical networks. We provide an overview of the optical and electronic grooming techniques available with focus on IP as the client layer We discuss the various architectural alternatives available: peer, overlay and augmented models. We first provide a survey on the research work in the area of traffic grooming in optical circuit switched networks. We then identify problems with electronic grooming in terms of high speed router design and bring out the merits of optical grooming. Next, we describe the shared wavelength optical network technology called light-trails and compare its performance with electronic grooming networks for both the peer and overlay models. Based on our simulations on random graphs of various diameters, we identify the threshold router speeds at which light-trails can compete with the electronic grooming solution for a given network scenario. We conclude that since the present router capacities are below the threshold speed or such routers are likely to remain expensive for some time, light-trails is an appealing candidate solution. ILP based techniques work in the static grooming problem with an objective to maximize network throughput. An ILP based mathematical formulation is presented for single hop and multi-hop grooming for multi granularity connection with non bifurcation constraints. Two heuristics with one that maximizes single-hop traffic (MST) and the other that maximizes resource utilization (MRU) are presented. Simulations were performed to observe the throughput with limited number of transceivers and wavelengths and were compared with the optimal solution. The paper concludes that MRU performs better if tunable transceivers are used and MST performs better if fixed transceivers are used. Auxiliary graph based techniques Online approaches for provisioning connections of different bandwidth granularities were dealt with in. For a connection to be established between an existing light path and if that fails to use a series of light paths. If the connection has not been accommodated yet, a new direct light path is set up or a mix of old and new light paths are used in propose a simple model for routing in peer model by assigning different weights to already existing circuits and new wavelength links. The special emphasis in the paper is on the signaling and protocol implementation aspects of the grooming scheme. A generic graph model for grooming traffic in a heterogeneous Grooming network environment is presented in the algorithm takes into account the heterogeneities in the network in terms of wavelengths, transceivers, and conversion and grooming capabilities. Besides, it solves the grooming problem in a combined way instead of splitting it into multiple sub problems and solving them independently. Three different policies were introduced, edge weight assignment principles were discussed and three traffic selection schemes were analyzed. The basic model can be used for static traffic as well Network flow based techniques the study the problem of traffic grooming to reduce the number of transceivers in optical networks. This problem is shown to be equivalent to a certain traffic maximization problem. An ILP formulation is presented and a greedy heuristic that uses the min cost flow problem is described. Simulation and ILP results were compared for uniform and random traffic pattern for small networks. An algorithm for integrated routing for the peer model. It uses a graph based approach that contains both the virtual and physical links. The model identifies all the min cuts for every possible ingress-egress pair and considers a link to be critical for this pair, if this link appears in at least one of its cuts. Each link is assigned a cost based on the number of LSR pairs for which this link is considered critical. By discouraging a new flow from using these links, the amount of residual capacity in the network can be maximized at every iteration. However, the complexity of this heuristic is pairs. is high since it has to compute max flow for all node pairs. Augmented Model most significant contribution of the work in is to identify a specific type of control information that could be exchanged along the IP and optical networks for the augmented model. The paper suggests that the WDM layer pass Lij, the number of light paths that can be established between LSRs i and j, to the IP/MPLS layer. Lij could be the number of common free wavelengths available on every link of the path identified by the routing algorithm. It is approximated that the amount of capacity available between i and j is the sum of residual capacities on the existing logical topology and the amount that could ring and mesh networks. E-grooming is inherently a hard problem. This can be seen from the fact that e-grooming problem in path, star and tree topologies are NP-Complete. Since the RWA for such topologies are trivial, this result be used in the future (Lij).

By assigning a cost to the link that is inversely proportional to the total residual capacity, the algorithm achieves an order of magnitude improvement in results than provided. E-grooming has been studied for various topologies like path, star, tree, brings out the hardness of the 'grooming' aspect of the problem. The objective of the e-grooming problem is to optimize a cost function that is typically one of the following:

--- Minimize equipment requirements

VI. MATHEMATICAL MODEL

So far, we have examined RWA and bandwidth-allocation algorithms for DDRs. In order to better understand our problem we state it as a MILP, which can solve the RWA and bandwidth-allocation sub problems together. There are three variations of our MILP. The first allocates flexible bandwidth to the requests; hence, it is named Adaptive_ILP. The other two allocate fixed bandwidth to the requests and are named Min_ILP and Max_ILP.since they use the Min and Max bandwidth-allocation policies. These MILP formulations can be used as benchmarks for our heuristic provisioning approaches. Our MILP model assumes that all DDR arrivals and deadlines are known; hence they are based on static traffic. However, the solution of the MILP constitutes a valid lower bound on the performance of our provisioning approaches (which consider a dynamic traffic environment). Our MILPs can provision DDRs in a network equipped with Opaque OXCs. The three MILP formulations are computationally Complex, especially Adaptive_ILP , as it includes: 1) selection of the appropriate bandwidth for DDR , which can be translated into a flexible finish time for the transmission of data; 2) RWA and grooming; and 3) constraints for time-disjointedness of requests that share common resources. That is why we simplify the routing, by considering only K alternate routes for each DDR, an approach utilized in other works that consider time-domain scheduling

1) maximize the number of accepted requests

$$\text{Maximize: } \sum_{i=1}^{m} \sum_{k=1}^{k} p^{i,k} \qquad (1)$$

count the number of accepted request by considering which request count path for their file transfer

2) maximize total network through put

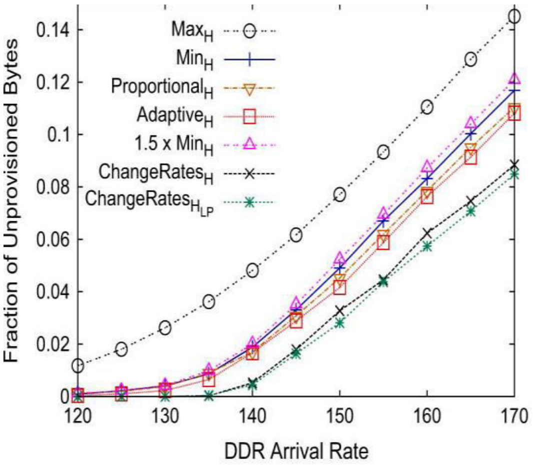
$$\text{Maximize: } \sum^{m(Fi\times}{}_{k=1}{}^{k\sum pi,k)}$$

Considers total data transferred for each DDR and provision the request that provide maximum throughput. RWA for Hybrid Architecture We can either use existing light paths or create new ones by using free physical resources. Depending on the grooming policy used, different weights are assigned to the edges in the auxiliary graph G''. Minimum weight path algorithms are then applied on G''.KSP in Algorithm 1 is the K-Shortest-Paths algorithm. The paths P obtained from applying KSP on G' are a sequence of existing light paths and/or physical links.

VII.SIMULATION RESULTS

Fig 2 Performance of the bandwidth-allocation policies forBD2 andBD3 (a) Fraction of unprovisioned bytes forBD2 Fig.2(a) shows the fraction of unprovisioned bytes for .Among the fixed allocation policies, Max_H rejects significantly more bandwidth than Min_H . As expected, the 1.5 *Min_H policy has intermediate performance between MinH and MaxH. Considering the adaptive bandwidth-allocation policies both AdaptiveH and ProportionalH perform slightly better than, which is expected because they utilize more information (i.e., the network state). Both flavors of Change Rates improve performance over the other bandwidth-allocation approaches (same as for BD1,ChangeRates_{LP} provisions slightly more bandwidth than Change Rates). Overall, for, BD2the performance can be improved if we utilize the adaptive policies over the fixed ones; further improvement is possible if Change Rates approaches are used. Fig 2(b) Fraction of unprovisioned requests for BD3 and effect of allowing

limited number of rate c changes. Fig2(b) shows the performance of the allocation policies for BD3 and performs a sensitivity analysis on the Changing Rates with Time Limitations policy, which may be preferred in practice as it involves less signaling overhead compared with the standard Changing Rates. Time T (shown in brackets) is the minimum time between two possible consecutive rate changes in the lifetime of a DDR. Fig. 2(b) shows that, for time periods of 10, 20, and 30 s between rate changes, Changing Rates$_{HLP}$ still outperforms Min. For 40 s, however, rate changes are no longer applied because the period between allowed changes is too long(compared with holding time), and the performance is closer to Max(recall that Changing Rates$_{H}$ policy H:Holding rates) We compare the performance of provisioning DDRs in Opaque and Hybrid networks (Opaque results are subscripted with O, MaxO e.g.,). The fraction of unprovisioned bytes for BD1.In the case of Max$_{H}$ and Max$_{O}$, the type of OXC is not as important, because Max does not groom requests (hence, the difference in's performance is mainly due to the different RWA algorithms being used for Hybrid and Opaque). We observe that Min$_{O}$ performs significantly better than Max$_{O}$. In addition, Mi$_{no}$ is able to provision more bandwidth than Min$_{H}$. This is because Opaque OXCs do full OEO conversion, so they can do better grooming than Hybrid OXCs. To evaluate the efficiency of our DDR provisioning approaches, we studied the utilization of network utilization) for both the Hybrid and the Opaque cases. Because of space considerations, in we only show the average transceiver utilization, computed as the average (Tx+Rx)/2(with Tx and Rx being the number of utilized transmitters and receivers)

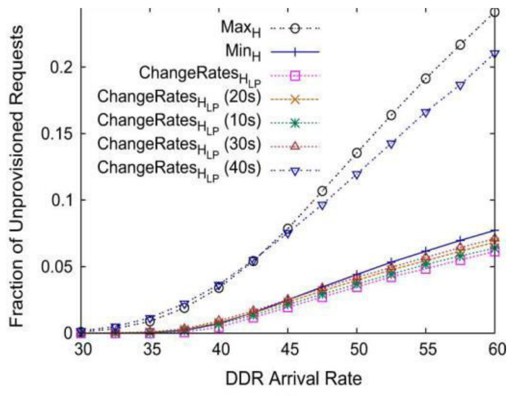


(a)

Fig. 2 : Performance of the bandwidth-allocation BD3and policies forBD2andBD3. rate changes

(a) Fraction of unprovisioned bytes for BD2

(b)



(b)

Fig 2(b) Fraction of unprovisioned requests for Effect of allowing a limited number of interval of time between two consecutive events.

We use bidirectional transceiver slots. Since today's networks are often over provisioned, for this experiment, we assume that the capacity of the links (i.e., number of wavelengths) is large enough to satisfy all requests. This way, we can compare the transceiver utilizations fairly, with a constant value (e.g., zero) of unprovisioned bandwidth for all our approaches. in this system architecture the mesh networks acts as a source node and wave length assumption is taken between source and destination node source and destination is transferring data by checking the k-shortest path its send data which path is shortest and bandwidth allocation algorithm used for giving bandwidth between the two nodes and uses network resources utilization by using this the data is reaching DDR and shows the network performance.

VIII. CONCLUSION

In this paper we studied the problem of provisioning DDR over WDM mesh networks by allowing flexible transfer rates. we are proposing the bandwidth allocation policies i.e. fixed, adaptive, changing rates . Allocate a fixed amount of bandwidth to request depending on its MinRate. Adaptive improve the performance of band width allocation algorithm. Changing rates allow the transmission rate of existing requests to change over time to accommodate new request that cannot be provisioned.

ACKNOWLEDGMENT

We are very thankful to the authors who have helps us to improve the quality and performance of our paper. We are grateful to our guide who guided us to our research work.

REFERENCES

[1] S. Balasubramanian and A. Somani, “On traffic grooming choices for IP over WDMnetworks,” in Proc. Broadnets, San Jose, CA, Oct. 2006,

[2] H. Miyagi, M. Hayashitani, D. Ishii, Y. Arakawa, and N. Yamanaka,“Advanced wavelength reservation method based on deadline-aware scheduling for lambda grid networks,” J. Lightw. Technol., vol. 25, no.10, pp. 2904–2910, Oct. 2007.

[3] B. Mukherjee, “Architecture, control, and management of optical switching networks,” in Proc. IEEE/LEOS Photon. Switch. Conf.,Aug. 2007, pp. 43–44.

[4] “Grid Network Services Use Cases from the e-Science Community,” T.Ferrari, Ed., 2007, Open grid forum informational document.

[5] J. Zheng and H. Mouftah, “Routing and wavelength assignment for advance reservation in wavelength-routed WDM optical networks,”in Proc. IEEE Int. Conf. Commun., Jun. 2002, vol. 5, pp. 2722–2726.

[6] J. Yen, “Finding the □ shortest loop less paths in a network,” Manage.Sci., vol. 17, no. 11, pp. 712–716, Jul. 1971.

◈◈◈

Artificial Bee Colony algorithm and its Application for Image Fusion

Bhavya V S,Navyashree K M, Prabhat Kumar Sharma, Sunil K S, Pavithra P

Dept. of ISE, Reva Institute of Technology and Management, Yelahanka, Bangalore– 560 064, India
E-mail : bhavyav.sridhar, navyashreekm26, prabhat19sharma, sunhill.ise, pavithraperumal)@gmail.com

Abstract - Artificial Bee Colony (ABC) is one of the latest swarm algorithm based on the intelligent foraging behavior of honey bees introduced in the year 2005 by Karaboga since then it has been used for optimization of various solutions. And it is recently introduced for processing and analysis of images such as segmentation, object recognition and image retrieval. Fusing images from a vast collection of different images has become one of the interesting challenges and has drawn the attention of researchers towards the development of fusion techniques. In this paper, we have proposed the usage of ABC for optimal fusion of multi-temporal images and studied the effect of variation in the source area.

Keywords - *Artifiical Bee Colony algorithm(ABC), Image Fusion, Quality measure.*

I. INTRODUCTION

Image fusion is the process of combining information from two or more images of a scene into a single composite image that is more informative and is more suitable for visual perception or computer processing. The goal of image fusion is to integrate complementary multi-sensor, multi-temporal and/or multi-view information into one new image[1].Image fusion is used in the field of image classification, aerial and satellite imaging[2], medical imaging[3], robot vision, concealed weapon detection[4], multi-focus image fusion [5], face recognition.

Fusing images from a vast collection of different images has become one of the interesting challenges and has drawn the attention of researchers towards the development of fusion techniques. A large number of image fusion techniques proposed were based on wavelet transformation. Image fusion using the DT-CWT was proposed in[6]. Image fusion based on a type of shift-invariant DWT was suggested in [7]. Information at higher levels of abstraction such as image edges and image segment boundaries are used to guide image fusion by pyramids [8].A new image fusion method that combines HIS transform and curvlet transform was proposed in [9].

In recent time many work has been done to introduce this ABC algorithm is field of image processing. In image retrieval system Artificial Bee Colony optimization algorithm is used to fuse similarity score based on color and texture features of an image thereby achieving very high classification accuracy and minimum retrieval time[10]. Image registration is a hot topic in the field of image processing, and it is widely used in various applications. A novel image registration technique was proposed. For the model the ABC algorithm was used[11].Image segmentation is still a crucial problem in image processing. It hasn't yet been solved very well. ABC algorithm is used to achieve multi-level thresholding image segmentation based on PSNR[12].An improved Artificial Bee Colony Algorithm (ABC) was introduced for the object recognition problem in complex digital images which resulted in very improved and efficient object recognition[13].Likewise many research work is going on to introduce this algorithm in the field of image processing. In the literature survey conducted by us we found that no work has been done in image fusion using ABC algorithm.

Many existing image fusion methods are based on wavelets and pyramid, which are very complex to implement. Some of the primitive image fusion techniques based on arithmetic operations on the pixels are simple. But these simple approaches often have serious side effects such as reducing or increasing the contrast[4].

Our proposed image fusion technique based on ABC is simple to implement and the quality of image obtained is better than that of the pixel-pixel arithmetic operation.

II. PROPOSED WORK

A. ABC concept

ABC algorithm is based on the intelligent way of the bees interacting with each other. Honey bees being social insects divide their work among themselves: Employed bees, Onlooker bees and Scout Bees[14]. Their activities are categorized into four main phases:

Initialization phase, Employed bee phase, Onlooker bee phase and Scout bee phase. In initialization phase, each employed bee is assigned with different food resources. In employed bee phase, each employed bee calculates the nectar amount of the food resource associated with it and the distance of it from the hive. After collecting the important information of the source the employed bee share the gathered information with the bees waiting in the hive. In onlooker bee phase**,** onlooker bees (the bees waiting in the hive) read information regarding different food resources and choose the best food resource. In scout bees phase, the employed bees whose food resource becomes abandoned turns into scout bee. The main job of scout bees is to search for new food resources.

In field of computer science and operation research, ABC is mainly used for solution of optimization problem. When related to optimization problem, the food resources are the set of different feasible solutions available .The nectar value of each food resource calculated by the employed bees is the fitness value of the a particular feasible solution[15]. The food resource chosen by the onlooker bees are the best optimal solution among the available set of feasible solution.

The main steps of the algorithm are given below[16]:

- Initialization
- Repeat

(a) Assign the employed bees onto the food sources in the memory;

(b) Place the onlooker bees on the food sources in the memory;

(c) Send the scouts to the search area for discovering new food sources.

- Until (requirements are met).

B. Proposed ABC in image fusion

The objective of the study is to apply Artificial Bee Colony algorithm for fusion of multitemporal images.

In our experiment, the initialization phase consists of assigning a source to the employed bees. Since the work is on fusion of two images, each of these images is divided into small areas which become the source for employed bees. As we are interested only in the information contained in the individual images taken at different time intervals, we have chosen entropy as the nectar and the entropy value is the measure of the nectar amount. These values are compared by the onlooker bees and they choose the pixel of the image from the source which has the highest nectar amount and put them into the hive (fused image).

1) Methodology

The methodology followed for conduction of the experiment is described in figure 1.

The arithmetic method of image fusion is performed by taking the average of corresponding pixels of two images and placing it in the corresponding position in the output image.

$$C[i,j] = (a[i,j] + b[i,j])/2$$

Where $0<i<m$ is height and $0<j<m$ is weight of the window.

2) Algorithm

Step 1: Start.

Step 2: Initialization phase:

Read two images(Source).

Step 3: Employed Bee phase:

Select a source area of size (m x m) in both the images.

Calculate the properties(nectar amount) for both the source area.

Step 4: Onlooker Bee phase:

Select the centre pixel(nectar) of the source area(source) having highest property value (nectar value).

Store the selected pixel in 2D buffer(hive).

Step 5: Scout Bee phase:

Select next source area (new source) and repeat the steps 3 to 5, (p-m/2)*(q- m/2) times.

Here, p= Height and q = width of selected image and w=window size.

Step 6: Stop

C. Advantages and Limitations

The merit of the proposed algorithm is that it's easy to implement and the quality of image is better. The demerit of our work is the determination of window size. For the images chosen for the experiment larger window size gave better result but the same may not hold good for all the images.

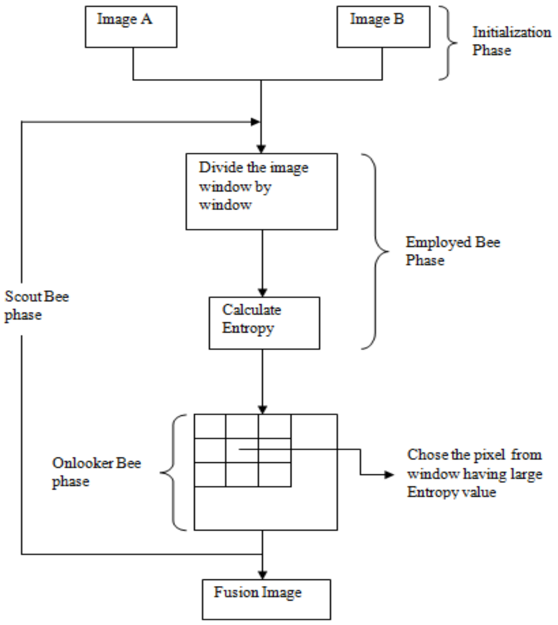


Fig. 1: Methodology

III. IMPLEMENTATION

A. *Experimental setup*

The experiment was conducted using VC++ 6.0 and the images were grayscale of the type RAW. The different types of images available for fusion were:

- Multi-view fusion of images from the same modality and taken at the same time but from different viewpoints.
- Multi-modal fusion of images coming from different sensors (visible and infrared, CT and MR, or panchromatic and multispectral satellite images).
- Multi-temporal fusion of images taken at different times in order to detect changes between them or to synthesize realistic images of objects which were not photographed in a desired time.
- Multi-focus fusion of images of a 3D scene taken repeatedly with various focal lengths.[17]

We have chosen three different sets of multi temporal images for our experimental study.

B. *Performance parameter*

The quality matrices for object analysis used for output image used in the experiment are[1]:

1. Entropy
2. Spatial frequency

1) Entropy

Entropy is a measure of the amount of information that can be derived from the image[6]. Entropy is calculated by the following equation:

$$E = -\sum_{n=0}^{n-1} P(i) log_2 P(i) \qquad (1)$$

Where,

n=maximum gray level considered

P(i)=normalized histogram of the graylevel i.

By taking entropy into consideration we can merge multi-temporal images by extracting the best content available in different images.

2) Spatial frequency

Spatial frequency is the measure of the overall activity level in an image[6]. For an MxN image F, with the gray value of pixel position (m, n) denoted by F(m, n), spatial frequency is computed as:

$$SF^2 = RF^2 + CF^2 \qquad (2)$$

Row Frequency(RF) and Column Frequency(CF) are given by the equations:

$$RF = \sqrt{\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=2}^{N}(F(m,n) - F(m,n-1))^2} \qquad (3)$$

$$CF = \sqrt{\frac{1}{MN}\sum_{n=1}^{N}\sum_{m=2}^{M}(F(m,n) - F(m-1,n))^2} \qquad (4)$$

IV. RESULTS AND DISCUSSION

A. *Fused Images*

The results of images fused by using ABC and the arithmetic based fusion are given below in figure 2,3 and 4.

(a)

(b)

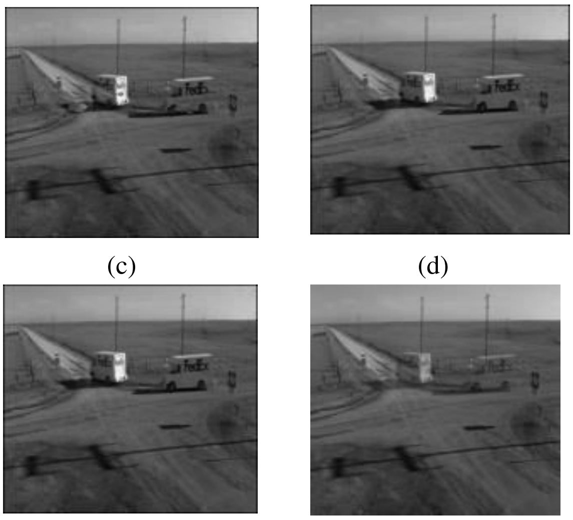


Fig. 2: Cab,Input images used for fusion-(a),(b); Images fused using ABC with varying source areas-(c),(d),(e); (c)-Source area=11x11,(d)-Source area=37x37,(e)-Source area=47x47,(f)-Image fused using average method.

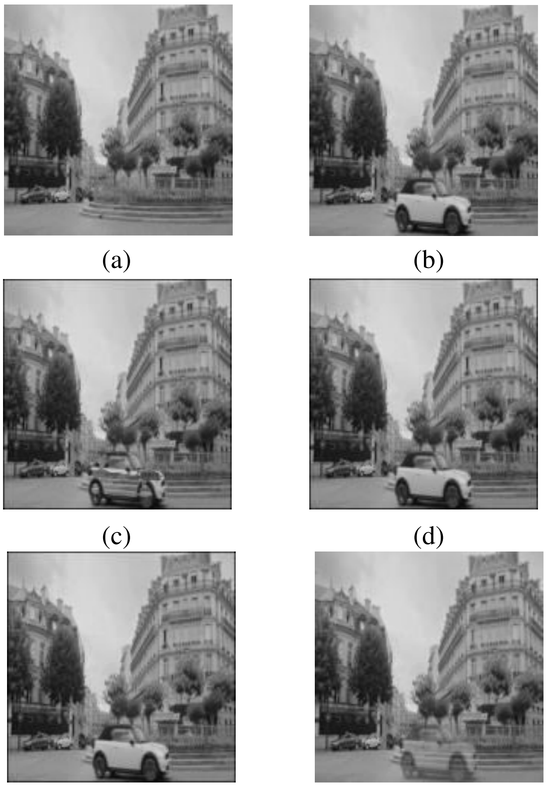


Fig. 3:Car input images used for fusion-(a),(b); Images fused using ABC with varying source area-(c),(d),(e); (c)-Source area=11x11,(d)-Source arear=37x37,(e)-Source area=47x47,(f)-Image fused using average method.

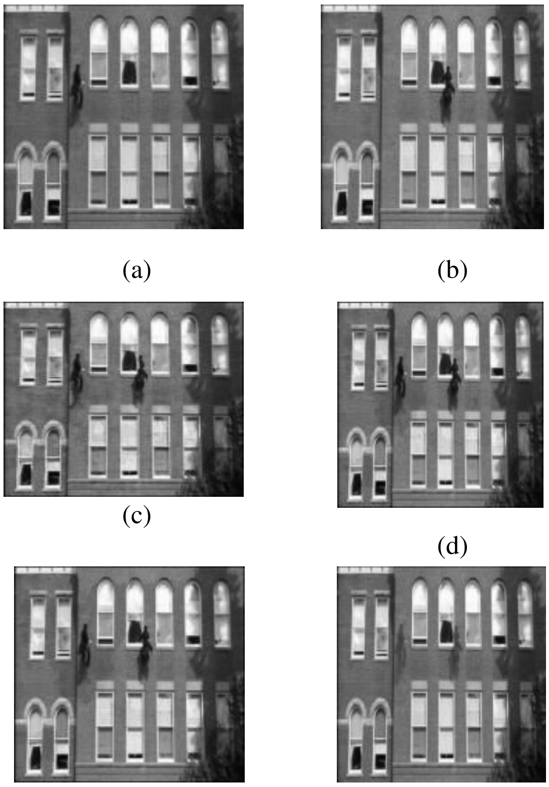
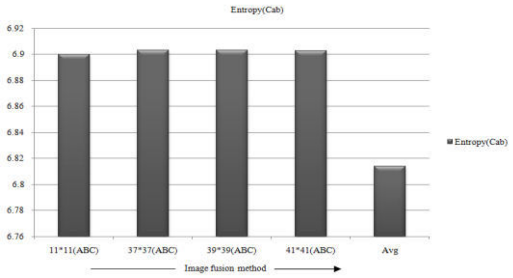


Fig. 4: Man input images used for fusion-(a),(b); Images fused using ABC with varying source area -(c),(d),(e); (c)-Source area=11x11,(d)- Source area=37x37,(e)-Source area=47x47,(f)-Image fused using average method

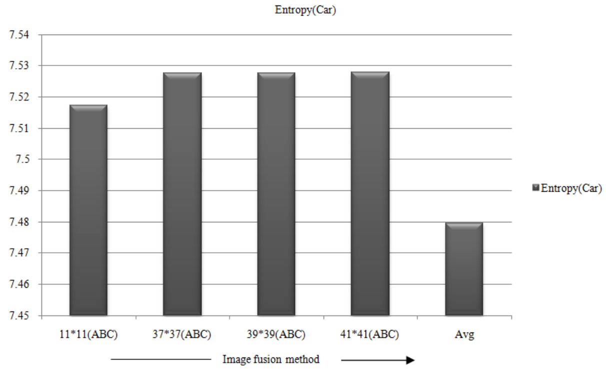
B. Analysis of entropy

All the output images obtained were subjected to objective and subjective quality analysis described in [1].

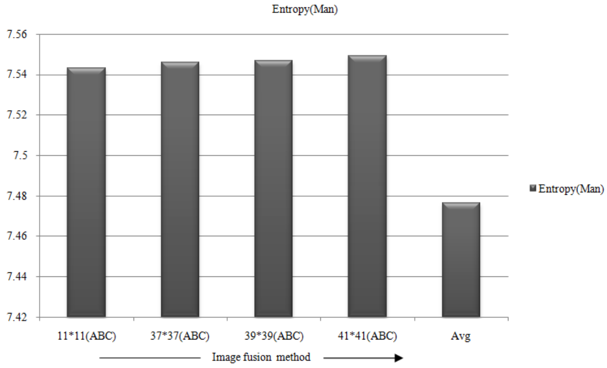
The following graph shows the analysis of the entropy of the different output images obtained



Graph I: Cab; Entropy values of the output image using ABC and average method of different source area



Graph II: Car;Entropy values of the output image using ABC and average method of different source area.

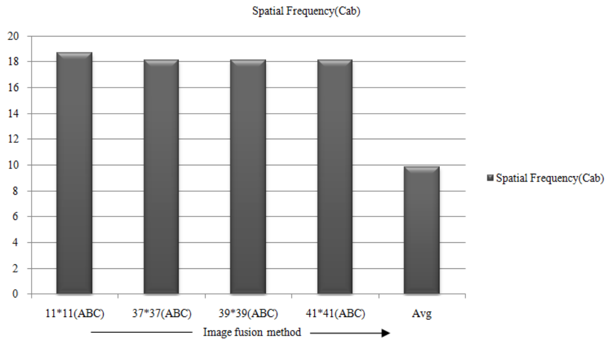


Graph III: Man; Entropy values of the output image using ABC and average method of different source area.

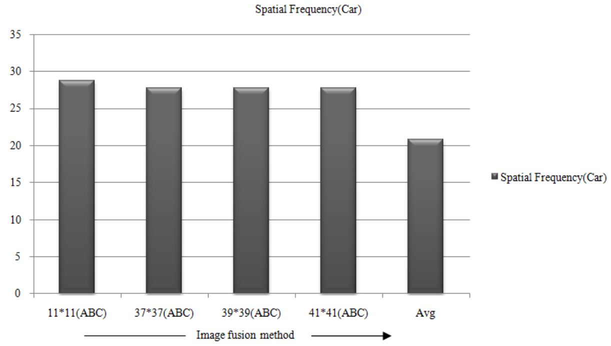
The maximum value of entropy of an image is 8 and the image with this value is said to have the maximum amount of information. Hence higher the value of entropy , better is the quality of the image [1]. The graphs I, II and III indicate that the images fused by ABC has higher entropy value hence is better than that of average method.

C. Analysis of spatial frequency

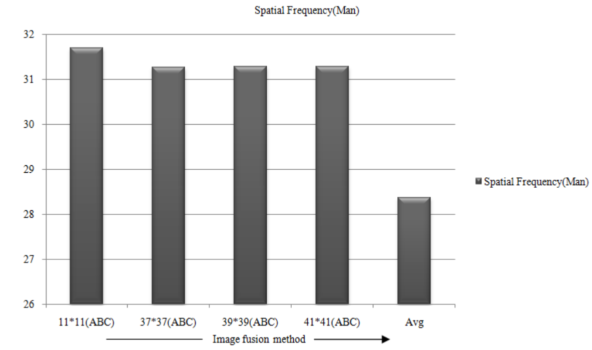
The following graph shows the analysis of the entropy of the different output images obtained.



Graph IV: Cab; Saptial frequency values of the output image using ABC and average method of different source area



Graph V: Car; Spatial frequecny values of the output image using ABC aand average method of different source area.



Graph VI: Man; Spatial frequecny values of the output image using ABC and average method of different source area

Image with higher SF value indicates that the image have greater activity level, hence has better quality [1]. The graphs I, II and III indicate that the images fused by ABC has higher SF value hence is better than that of average method.

V. CONCLUSION

The application of ABC for image fusion was successful and the output images fused based on ABC has more information content (higher Entropy value) and also looks better (higher Spatial Frequency values) than the images fused by arithmetic operation, figure 2,3 and 4. Although not much of variation was observed in the objective image quality metric values, a large amount of artifacts were observed by subjective analysis in the images which were fused using smaller source area than those of larger source areas.

Future work would include finding an optimized source area (window size) for fusion of images using ABC.

REFERENCES

[1] Ms. Shruthi T V , Ms. Ramyashree N , Ms. Pavithra P , Ms. Kamalam Balasubramani;"Analysis of Image Quality Using Quantitative Methods";ICICES,23rd-24th Feb 2011.

[2] Uttam Kumar, Chiranjit Mukhopadhyay and T. V. Ramachandra; "Pixel based fusion using IKONOS imagery", International Journal of Recent Trends in Engineering Vol 1, No. 1,2009.

[3] Wang Anna, Wu Jie, Li Dan, Chen Yu;" Research on Medical Image Fusion Based on Orthogonal Wavelet Packets Transformation Combined with 2v-SVM ", Complex Medical Engineering, 2007. CME 2007, ISBN: 978-1-4244-1078-1 Page no. 670-675. IEEE/ICME International Conference on 23-27 May 2007.

[4] Z. Xue, R S. Blum, and Y. Li;" Fusion of Visual and IR Images for Concealed Weapon Detection", Information Fusion, 2002. Proceedings of the Fifth IEEE International Conference on 2002,Vol 2, Page no: 1198 – 1205.

[5] Muhammad Hassan Arif and Syed Sqlain Shah;" Block Level Multi-Focus Image Fusion using Wavelet Transform", Signal Acquisition and Processing, 2009. ICSAP 2009. International Conference on 3-5 April 2009,Page no: 213 – 216.

[6] Paul Hill, Nishan Canagarajah and Dave Bull;" Image Fusion using ComplexWavelets ",BMVC Sept-2002,Page no: 487-496.

[7] Zhong Zhang,"A Categorization of Multiscale-Decomposition-Based Image Fusion Schemes with a Performance Study for a Digital Camera Application", Proceedings of the IEEE Aug 1999**,** 8 Volume:7, Issue no: 8 **Page no:** 1315 – 1326.

[8] Vladimir Petrović;" Multi-level Image Fusion", www.**i**magefusion.org.

[9] Man wang, Dai-yang cao, Jie-lin zang,"Fusion of multispectral and panchromatic satellite images based on HIS and curvlet transformation",ICWAPR-2007,Vol:1 Page no: 321-325.

[10] D.Chandrakala and S.Sumathi;" Application of Artificial Bee Colony Optimization Algorithm for Image Classification Using Color and Texture Feature Similarity Fusion"; SRN Artificial Intelligence; Volume 2012 (2012), Article ID 426957.

[11] Shuihua Wang," Artificial Bee Colony used for Rigid Image Registration", International Journal of Research and Reviews in Soft and Intelligent Computing (IJRRSIC) Vol. 1, No. 2, June 2011.

[12] Cao Yun-Fei, Xiao Yong-Hao, Yu Wei-Yu and Chen Yong-Chang," Multi-level Threshold Image Segmentation Based on PSNR using Artificial Bee Colony Algorithm", China Research Journal of Applied Sciences, Engineering and Technology Published: January 15, 2011.

[13] Chidambaram Chidambaram, "An Improved Artificial Bee Colony Algorithm for the Object Recognition Problem in Complex Digital Images using Template Matching",IJNCR,Vol 1,Issue 2,2010.

[14] Adil Baykaso lu, Lale Özbakır and Pınar Tapkan, "Artificial Bee Colony Algorithm and Its Application to Generalized Assignment Problem", ISBN 978-3-902613-09-7, pp. 532, December 2007.

[15] A Survey on Artificial Bee Colony Algorithm, Karnan Marcus , KamalamBalasubramani, Department of Computer Science and Engineering, Tamil Nadu College of Engineering, Coimbatore, India Department of Information Science and Engineering, Atria Institute of Technology, Anandanagar, Bangalore, India. IEEE ICCIC-2011, Dec 15th-18th2007,ISBN 978-1-61284-694-1.

[16] Manish Gupta, Govind Sharma, "An Efficient Modified Artificial Bee Colony Algorithm for Job Scheduling Problem", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-6, January 2012 .

[17] Jan Flusser, Filip Sroubek, and Barbara Zitov, "Image Fusion: Principles, Methods, and Applications",Tutorial EUSIPCO 2007.

◈◈◈

Generating Spring Graph to Infer RDF Data

Meera Mudholkar[1], Saloni Peswani[2], Seema Kela[3], Jagannath Aghav[4], Harish Jadhao[5] & Prashant Gaikwad[6]

[1,2,3,4&5]College of Engineering, Pune, India
[6]John Deere, Pune, India
E-mail : mudholkar.meera@gmail.com[1], peswani.saloni18@gmail.com[2], kela.seema15@gmail.com[3],
jva.comp@coep.ac.in[4], jadhaoharish@gmail.com[5], GaikwadPrashant@JohnDeere.com[6]

Abstract - RDF technology allows recognition of semantic context of a massive amount of data and a spring graph provides an effective visualization mechanism. In this paper, we propose solution for a market analyst using the spring graph technology which aims at depicting the resources available on the web-pages and the relationship between them. The implementation comprises of spring graph generation using Adobe Flash builder and manipulation and execution of queries accomplished by Jena (Java API) and SPARQL query engine. Current implementation takes domain specific file input as an illustration. The visual representation of a graph based on the proposed implementation demonstrates the details of relevant information desired by a market analyst.

Keywords - *Spring graph, RDF, Jena, Sparql, Ontology.*

I. INTRODUCTION

Existing systems for aggregating competitive research data in the field of market analysis simply focus on the acquirement and manipulation of obtained data, with little efforts made towards presenting the data to a user. In this paper, we propose a mechanism for semantically structuring data based on the concept of semantic web and making use of available semantic web technologies. The manipulation of this structured data results in a comprehensive amount of data and discerning this data in a text format is an arduous exercise for the user. To simplify this task of deciphering the information, we propose an approach to display the information in a graph visualization technique called spring graph. The system illustrates information related to company, products manufactured by the company, region, city and organization, which can be viewed by the market analyst.

A spring graph is a flex component used for graph visualization, which exhibits a graph of objects that are linked to each other using a force directed layout algorithm. The use of spring graph focuses on providing user interactivity, readability, user friendliness and output in a pleasing form. Spring graph generator will read a description file which contains the node's description and the relation between the other nodes connected to it. Each node is drawn as an ellipse, whereas the connection between two nodes is drawn as an arrow. The connected nodes move towards each other and this process gets repeated until the graph achieves constancy.

Adobe Flash Builder can be used for generating such spring graphs, where data can be provided in XML [1] format. It gives the graph three dimensional view. This graph is more reliable and readable than a general graph and focuses on good quality results. These graphs are flexible, simple, interactive and intuitive. Spring graph removes node overlapping problem which is generally found in conventional graphs.

User can perform certain operations on the results of the SPARQL queries such as searching for particular organization, product, person or any company. This result is provided to the spring graph generator. Spring graph makes use of Force based algorithms. Equilibrium Maintainer is a function of spring graph and it maintains equal distance of each node from central node. User can change repulsion factor according to view. User can also specify how many nodes should be visible at a time. Spacing between items can also be changed according to usage. It is dynamic way of generating graphs which are more interactive.

II. RELATED WORK

A few graph visualizations have been developed for representing ontology extraction in a graphical format. But these visualizations are implemented in the conventional graph representation and no current work is being carried out in the context of spring embedded graph generation. A web based search engine based on the spring graph model provides results which is an aggregation of information obtained from varied sources like news and magazine articles, audio and video content, blog entries and many other sources. It outputs the result in a graphical representation along with charts provides a 360-degree search and allows for varying the relevancy of a topic as per user requirement.

Another implementation produces visualization of graphs using a spring embedded algorithm that eliminates edge crossings [6]. This implementation generates three dimensional graph visualizations that provides graph with better quality and attractiveness.

Another paper discusses the use of graph theory for social network analysis [8]. This paper proposes that by determining the conceptual distance between people and groups, information about the type of communication in an organization can be inferred. It focuses mainly on analysis of military organizations and visualization of the relationship between people in the organization through diagrams and graphs. It makes use of graph theoretic techniques and traditional statistical approach for graph visualization Spring embedding algorithm is used here to visualize the social network, where the spring distance correlates to the actual distance between two nodes (i.e. the link distance).

Yet another paper proposes an extension of the spring embedding algorithm to the three dimensional realm. This extension called the GEM-3D (Graph Embedder 3D)[9] provides graph representation at interactive speed and better display quality. It also provides graphs for real life examples accommodating hundreds of nodes and makes use of visual clues for user readability. This new algorithm can be applied to both real world graphs as well as artificial graphs to represent the topology of undirected graphs. The algorithm was also successful in presenting a visualization of the Petersen graph which was impossible to draw using the earlier spring embedding paradigm. The same implementation is applied to both directed and undirected graphs at present.

Another implementation uses spring graph to visualize the information extracted from a digital library after semantic analysis [10]. The graph representation signifies a meaningful relationship between the documents and provides an efficient way of extracting the documents. The spring embedding algorithm represents semantic relationship between the entities, with lesser spring distance representing more similarity between the nodes, and vice versa.

This paper presents a preprocessor to enhance the performance of a conventional spring embedder, which can be used in parallel to numerous optimization and approximation techniques [11]. This preprocessor is invoked before the spring embedder and it operates in two phases. In the first phase, it maintains a user specified distance between the nodes in a graph. In the second phase, it distributes the nodes equally on a grid. The spring embedder is invoked after these two phases. This spring embedder is based on Fruchterman and Reingold algorithm which improves speed of the spring embedding process.

Another implementation uses spring graph to visualize the information extracted from a digital library after semantic analysis [10]. The graph representation signifies a meaningful relationship between the documents and provides an efficient way of extracting the documents. The spring embedding algorithm represents semantic relationship between the entities, with lesser spring distance representing more similarity between the nodes, and vice versa.

III. PROPOSED METHOD

The main components of system architecture are: RDF data and ontologies, Jena , Sparql engine and Spring graph generator.

A. RDF data and ontologies

RDF, which stands for Resource Description Framework, is a method used to elicit information from web pages on the worldwide web and structure it in a form understandable by a computer system. RDF represents each small piece of information in the form of a resource. A specific, domain related semantic is affixed to each resource and RDF serves as the basis for manipulating such kind of data. It displays data in triple format where every assertion is represented in following three parts: subject-predicate-object, more generally written as P(S; O).

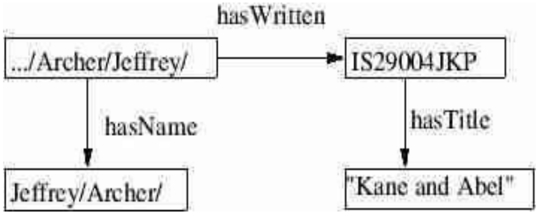


Fig. 1 : An RDF Example

Figure 1 illustrates an RDF subject-predicate-object triple. The URI ../Archer/Jeffrey/.. is used to represent the subject Jeffrey Archer, who is a novelist. The predicate has Written signifies the relationship between Jeffrey Archer and the book written by him. The URI IS29004JKP represents the book with name Kane and Abel. Thus this subject-predicate-object forms the RDF triple.

B. Jena and Sparql Engine

Jena [7] is a Java Framework that supports semantic web applications and provides a number of useful tools and libraries for working on RDF data. The RDF data can be written in xml format using Jena API. The input to the system is data in the form of RDF or OWL. A file consisting of RDF [1] data or OWL [2] ontology is stored in the database in the form of subject predicate-object triple. This data is nothing but information extracted from a number of web page sources which is compiled together and organized into a structured ontology. Search is carried out on this structured data to obtain desired result. The user wants to retrieve precise information about a particular topic. The structured data stored in the RDF file is searched to get the relevant information desired by the end user. A search query is passed to the Jena and SPARQL search engine, where it is processed to yield the required result. Essentially, Jena is used as a reasoner for RDF and OWL data since it has large capacity to store RDF triplets.

The main components of a SPARQL [3] engine are the query evaluation engine and storage manager. Low level instructions are passed to the query evaluation engine, where the instructions are processed to produce the required result. File manager component of the SPARQL engine deals with data structures and memory management.

Data dictionary stores additional information of the data structure being used. Query evaluation engine has latest SPARQL specification. It provides functions like aggregate, select, update and many more, whereas storage manager is responsible for storing all the data and results of the queries.

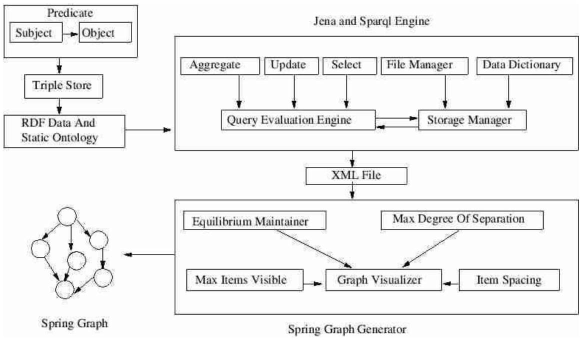


Fig. 2 : System Architecture.

C. Spring Graph Generator

The spring graph generator contains following components: Equilibrium maintainer, Maximum degree of separation, Maximum elements visible, Item spacing, and Graph visualizer.

The equilibrium maintainer maintains a constant distance between nodes in a graph. The function of this module is to check whether the distance between two nodes is the same as defined. If not, it will convert the current distance to the desired constant distance. The maximum degree of separation module determines the number of objects that will be connected to the central object. The maximum elements visible module determines the number of nodes that will be directly connected to the subnodes i.e. the objects of central node. Item spacing module ensures that the entire graph is visible in one panel as the maximum degree of separation and maximum elements values are increased. If these two values are less, the item spacing between nodes can be more. If those two values are more, the item spacing between the nodes will reduce.

IV. ALGORITHM

The spring graph that is generated contains a large amount of nodes, with each node connected to a number of other nodes. On clicking on the link connecting any two nodes, additional information representing the relationship between these two nodes will be displayed. The textbox displaying this information will also have a link to the main news article from where this information is retrieved. Since the RDF file contains voluminous data about a large number of nodes, the major challenge is to determine which nodes will be displayed onto the spring graph. We propose an algorithm for this purpose.

Algorithm: From RDF to Spring Graph

Require: RDF OR OWL DATA

Ensure: SPRING GRAPH GENERATED FROM GIVEN DATA.

```
1:  Accept string from user.
2:  if (string==null) then
3:      Goto step 7.
4:  else
5:      Goto step 8.
6:  end if
7:  Set the string to a predefined default value.
8:  Store the string as first element on queue.
9:  while (queue != empty) do
10: Pop element from front of queue and store it in a
    variable named subject.
11: Open rdf file for comparing subject with the triple
    stored in file.
12: while(!EOF)
13: Read subject from the subject-predicate-object
    triple stored in file into variable named
    filesubject.
```

```
14:  if (subject == filesubject) then
15:  Store the predicate onto predicate list, object onto
     object list. Also push the object at the end of
     queue.Goto step 12.
16: else
17:       GOTO step 12.
18:   end if
19: end while
20: Open the the static ontology file for reading
     description about subject.
21: while (!EOF)
22: Read subject from the subject-predicate-object
     triple stored in file into variable named
     descrsubject.
23: if (subject == descrsubject) then
24: Read the description of subject, which is stored in
     node description attribute into the variable
     description.
          Goto step 28.
25: else
26:       Goto step 21.
27: end if
28: Store the description in a description list.
29: Goto step 9.
30: Open a file in write mode.Write the subject, object
     from object list, predicate from predicate list and
     description from description list into the file.
31: end while.
32: STOP.
```

The proposed algorithm is simple and easy to understand. It requires limited storage in the form of a single queue and storage area for the subject-predicate-object triple. RDF data manipulation and extraction is simplified by the use of Jena and SPARQL query engine. This algorithm is efficient and trivial to implement. Search string entered by the user will be the initial subject that will be stored on the queue. The RDF file will be queried using Jena for objects related to this subject. Randomly subject is selected in case the user does not provide any input.

The search string entered by the user will be the initial subject that will be stored on the queue. The RDF file will be queried using Jena for objects related to this subject. A predefined subject is selected in case the user does not provide any input.

Consider the example where the user wants to search for information related to Steve Jobs. This search string will be accepted as the first input and pushed onto the initially empty queue. The subject at the front of the queue is then popped, in this case Steve Jobs, and SPARQL query language is used to query the RDF file in search of objects associated with this subject. These objects are then pushed at the end of the queue one by one, which will be later parsed. The object, predicate and description related to each subject is also stored in a list, and later used to write the triple to an XML file. This process continues recursively until there are no more objects related to the subject and the queue is exhausted. All of this information is then collectively used to generate an XML file, which in turn is passed to the spring graph generator to generate the spring graph.

V. IMPLEMENTATION

A spring graph is designed for providing information about the relationship between different entities. Adobe Flash builder provides a spring graph component class that allows building a spring graph and representing the relationship between the nodes. After clicking each link, it must display additional information about the relationship between the two entities. For the generation of spring graph, the RDF file and OWL file (containing description) are fetched. The data is extracted in triples form using sparql query language. The result that is obtained from the query gets stored in the form of list. The subject is pushed in the queue. The objects related to each of the subject are then extracted and added to the end of queue. This traversal is carried out until the queue becomes empty. The XML file is generated and given as input to adobe flash builder. The adobe flash builder reads the XML file to generate spring graph.

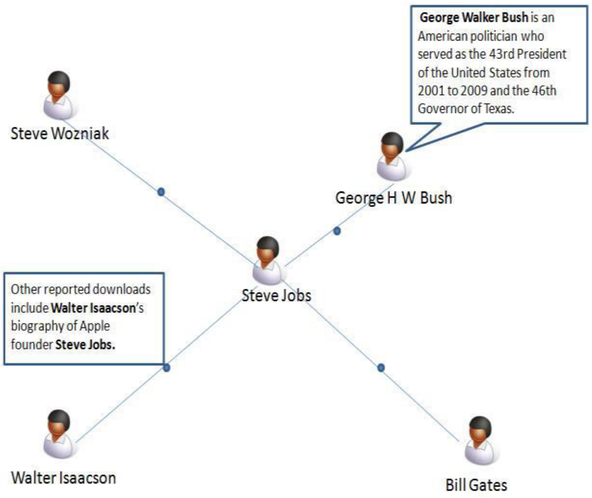

George Walker Bush is an American politician who served as the 43rd President of the United States from 2001 to 2009 and the 46th Governor of Texas.
Steve Wozniak
George H W Bush
Other reported downloads include Walter Isaacson's biography of Apple founder Steve Jobs.
Steve Jobs
Walter Isaacson
Bill Gates

Fig. 3 : A Spring Graph Example

VI. CODE SNIPPET

The following code demonstrates generation of spring graph for entities related to Steve Jobs using the spring graph component of Adobe Flash builder. Initially an empty graph is created using the Graph component of the spring graph package. Items are created and added to the graph and they are connected together by using the link method. The program code is presented below.

```
start:mx:Script
![CDATA[
%Importing packages
import mx.events.ScrollEvent;
import com.adobe.flex.extras.
controls.springgraph.Item;
import com.adobe.flex.extras.
controls.springgraph.Graph;
%Creating nodes
private var g: Graph = new Graph();
private function newItem(): void
var item1: Item = new Item(“Steve Jobs”);
var item2: Item = new Item(“Steve Wozniak”);
var item3: Item = new Item(“George H W Bush”);
var item4: Item = new Item(“Bill Gates”);
var item5: Item = new Item(“Walter Isaacson”);
%Add nodes to the graph
g.add(item1);
g.add(item2);
g.add(item3);
g.add(item4);
g.add(item5);
g.link(item1, item2);
g.link(item1, item3);
g.link(item1, item4);
g.link(item1, item5);
s.dataProvider = g;
]]
end:mx:Script
```

VII. RESULTS

The system provides an excellent tool for presenting a voluminous amount of data through the spring graph visualizer. Use of the system results in saving of time and effort on the part of the market analyst, wherein he doesn't have to scan through an enormous amount of information in order to deduce the desired result. The structured ontology leads to efficient and organized retrieval as well as manipulation of semantic data.

VIII.CONCLUSION

We have illustrated the generation of spring embedded graph using Adobe Flash builder. This system takes input from an RDF file and demonstrates a graphical visualization of the key values acquired from search result. System uses SPARQL querying facility in order to simplify the querying process and for efficient retrieval of RDF data. Spring graph displays relevant information required by the user, along with detail description of each node and additional information about the relationship between connected nodes. This system provides a solution for inferring meaningful data from a semantically structured static ontology and RDF data, which can be used as a business solution by a market analyst, sales representative, corporate organizations and many others. In future, we plan to determine the similarity between resources and accordingly vary the length of links, with more similarity resulting in shorter distance between nodes and vice versa. Also similar resources will be placed together in a cluster, with each cluster having a unique color for efficient identification.

REFERENCES

[1] Stefan Decker and Sergey Melnik ,Stanford University,Frank Van Harmelen, Dieter Fensel, And Michel Klein, Vrije University of Amsterdam Jeen Broekstra, Administrator Nederland B.V., Michael Erdmann, University of Karlsruhe, IAN Horrocks, University of Manchester, The Semantic Web:The Roles Of XML and RDF,2000.

[2] Milea V,Frasincar , A temporal Web Ontology Language.

[3] Jiewen Huang, Yale University, Daniel J. Abadi, Yale University and Kun Ren, Northwestern Polytechnical University, China Scalable, SPARQL Querying of Large RDF Graphs.

[4] Faruk Bajramovic,Arne Tauber,Ralph Wozelka and Wlorister Ferdinand, A Taxonomy of Force-Directed Placement Techniques.

[5] Philipp Cimiano, Johanna Volker,Institute AIFB, University of Karlsruhe. Text2Onto, A Framework for Ontology Learning and Data-driven Change Discovery.

[6] Weimin Wu, Yongfeng Cao, Quing Su, Yonghe Zhang, Guangdong University of Technology,Guangzhou,Visualization of Graph based on the Three-dimensional Spring Model.

[7] Jeremy J. Carroll, Ian Dickinson, Chris Dollin,Dave Reynolds, Andy Seaborne, Kevin Wilkinson,Digital Media Systems Laboratory, Jena:Implementing the Semantic Web Recommendations,2003.

[8] Anthony Dekker,C3 Research Centre,Defence Science Technology Organization, Visualization Of Social Networks Using CAVALIER,2006.

[9] Ingo Brub and Arne Frick,Universitat Karlsruhe, Fakultat fur Informatik,D-76128 Karlsruhe, Germany, Fast Interactive 3-D Graph Visualization.

[10] Junliang Zhang,University of North Carolina, Chapel Hill,Javed Mostafa,Himansu Tripathy,Laboratory for Applied Informatics Research Indiana University,Information Retrieval by Semantic Analysis and Visualization of the Concept Space of D-Lib Magazine,2010.

[11] Paul Mutton, Peter Rodgers,University of Kent at Canterbury, Demonstration of a Preprocessor for the Spring Embedder. http://http://kar.kent.ac.uk/13754/1/ DemonstratePaul.pdf.

[12] Jennifer Golbeck and Paul Mutton,Spring-Embedded Graphs for SemanticVisualization. http://trust.mindswap.org/papers/ GolbeckMuttonChapter.pdf.

◈◈◈

Self-Managing Performance in Application Servers – Modelling and Data Architecture

Ravi Kumar G[1], C.Muthusamy[2] & A.Vinaya Babu[3]

[1]HP Bangalore, Research Scholar JNTUH, Hyderabad, India, [2]Yahoo, Bangalore, India
[3]College of Engineering, JNTUH, Hyderabad, India
E-mail : ravikgullapalli@gmail.com[1], chelgeetha@yahoo.com[2], dravinayababu@yahaoo.com[3]

Abstract **-** High performance is always a desired objective in computing systems. Managing performance through manual intervention is a well-known and obvious mechanism. The attempts to self-manage performance with minimal human intervention are predominant in the recent advances of research. Control Systems theory is playing a significant role in building such intelligent and autonomic computing systems. We are investigating in building Intelligent Application Servers by enabling control system as a first class feature in component software, at design time and runtime. In this direction, it is important to build efficient data access mechanisms that capture the control system models, performance data, analyze the data efficiently, identify patterns and build a knowledge base. In this paper we propose a data organization and architecture as a building block of developing Intelligent Application Servers.

Keywords **-** *self-managing performance; Intelligent Application Servers; Control Systems; modeling and data architecture;*

I. INTRODUCTION

High performance is most desired non-functional requirement of any computing system. The rapidly increasing usage of IT resources triggered self-managing and self-correcting IT architectures from hardware to the application layer. Such self-managing and self-correcting solutions are referred as Autonomic Computing. There are various solutions in providing such autonomic and self-managing mechanisms in the recent research investigations [1]. Control Systems is obvious choice due to the inherent feedback and adaptive control capabilities to build such autonomic IT systems [2], supported by mathematical foundations. Control Systems proved successful applications in networking, database systems, and data centers. There is a great emphasis of control systems application in application layer involving Application and Web Servers [3]. Most of the research is around using classic controllers, building hybrid controllers for automated performance management. In this paper we intend to exploit generic modeling and data architecture providing simple ways to model, capture, organize and analyze the real time data that help in building more robust self-managing abilities in Application Servers.

II. PROBLEM AND RELATED WORK

Application Layers play a significant role in today's IT environments. They host applications and provide services ranging from individual consumers to the Enterprise users. It is obvious that such Application Servers exhibit high performance all the times. There are solutions applying control systems to self-manage the performance using adaptive controllers. Many of them are specific solutions to manage the performance of web servers [4][5], web services [6], web caching [7][8], EJB Servers [9], JMS Servers [10], JDBC drivers [11][12]. Most of such solutions are specific to the components in the Enterprise Application Server stack. Although the controllers, control algorithms are important in building self-managing computing systems, it is even more necessary to provide simple, faster, efficient data organization mechanisms that support generic data modeling, data analysis, and pattern identification mechanisms and construct knowledge base. In this paper we address the above said issues by a simple Data Architecture for efficient functioning of these controllers. The proposed Data Architecture is a building block of a bigger research problem where we are investigating to build an Expert Control System for Application Servers [13][14].

III. INTELLIGENT CONTROL ARCHITECTURE – OVERVIEW

The Fig 1 below shows Intelligent Control Architecture consisting of Data Architecture block in the outer loop and a Feedback block in the inner loop. The

Feedback block consists of a set of classic controllers such as P, PI, PID [15] and advanced controllers such as Time-Series, Fuzzy and other intelligent controllers. The Data Architecture block contains the functional blocks to capture the performance data of the Application Server component control system models, analyze data, identify patterns and create knowledge base. The subsequent sections explain this block in detail and the data that is exchanged between these blocks.

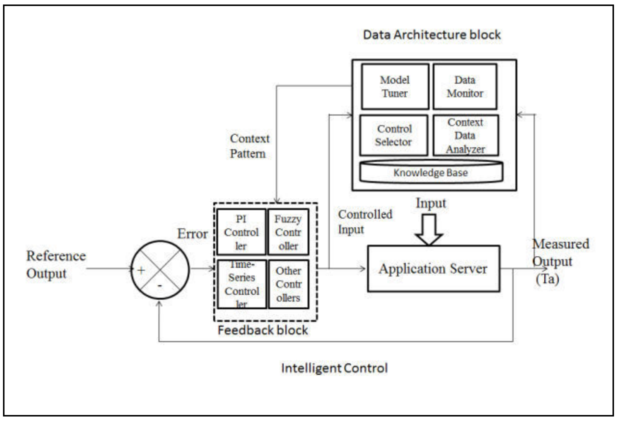


Fig. 1: Intelligent Control Architecture

At every sample interval the control loops shown are run, where the data architecture block analyzes the real time data, sends a “Context Pattern” to the Feedback block. The context pattern object contains the controller to be used and other information required for tuning the controlled input to the Application Server to achieve the desired performance.

IV. MODEL TEMPLATES – UML DESIGN

Control System solutions require the Application Server components models to tune their performance. The typical models adopted are ARMA models [16]. Equation (1) below shows a simple SISO system model.

$$\mathrm{y}(t + 1) = a\mathrm{y}(t) + b\mathrm{u}(t) \tag{1}$$

A. Model Templates

The feedback control system uses ARMA models and we have considered the same to represent the Application Server components in our solution.

The models of different JEE server [17] components such as JMS Providers [18], EJB Servers [19] , Web Servers are captured as SISO or MISO model [20] templates consisting of all possible output parameters for the server components. These are pre-defined models used to capture the associated performance data in the faster access database.

The pre-built model templates are separately packaged as XML resource files by the server component developers along with the compiled source. These templates are used at the Application Server initialization to create the database and estimate the different model parameters at runtime. The Fig 2 below shows the definition of a model template

```
<model>
 <component-name>
  <outputParam>  </outputParam>
  <inputParam1>  </inputParam1>
  <inputParam2>  </inputParam2>
  <inputParam3>  </inputParam3>
  <inputParam4>  </inputParam4>
</component-name>
</model>
```

Fig. 2 : Sample MISO Model Template in xml

B. Example Models – JMS Providers

The following Fig 3 shows the xml definition of representing the ARMA model of JMS Provider.

```
<model>
 <jms>
  <outputParam>
    msgTPut
  </outputParam>
  <inputParam1>noSub</inputParam1>
</jms>
</model>
```

Fig. 3 : JMS Provider SISO Model Template.

The above shown model template can be represented by the following (2) as SISO model

$$\mathrm{msgTput}\ (t + 1) = a\ \mathrm{msgTput}(t) + b\ \mathrm{noSub}(t) \tag{2}$$

‘*a*’ and ‘*b*’ are the model parameters. This model will be transformed into a column oriented database during the initialization of the Application Servers.

V. DATA ARCHITECTURE

This section explains the Data Architecture block of the Intelligent Control Architecture in the Fig 1. Effective data organization is important for analyzing the performance data; infer meaningful patterns to enable tuning the managed system. We propose to use a Column oriented database to capture performance data and will be periodically moved to a persistent storage when its patterns are defined as rules in the knowledge base. The Fig 4 below shows the proposed Data Architecture with two parts in it

- Data Organization: It deals with the parsing the model templates and capturing the data. The Data Modeler deals with transforming model templates of the Application Server components into the column oriented database. The Data Monitor updates the performance data.
- Data Analysis: It has a generic data access layer to fetch the performance data, which will be consumed by the Data Analyzer to analyze the data, identify patterns and generate Context Patterns.

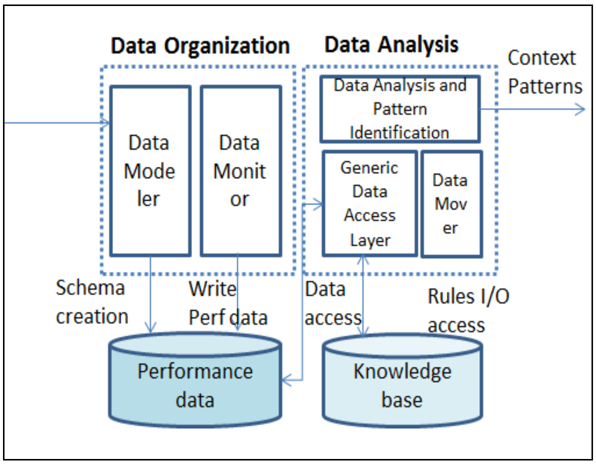


Fig. 4 : Data Architecture

A. *Context Pattern*

Context Pattern is an object that contains the predicted values for the future output parameter values, that reflects the behavior of the components of the Application Server. Additional parameters are also predicted such as percentage of CPU usage, Memory usage. Additionally for such a predicted behavior the suitable controller required to be applied is also set. The Fig 5 below shows a sample Context Pattern.

```
<ContextPattern>
  <Pattern>SuddenVariations</Pattern>
  <OutParam>MessageThroughput</OutParam>
  <OutParamVal>50</OutParamVal>
  <InParam>Smax</InParam>
  <InParamVal>75</InParamVal>
  <ControllerType>Fuzzy</ControllerType>
  <CPU>75</CPU>
  <MEM>68</MEM>
</ContextPattern>
```

Fig. 5 : Context Pattern

The context pattern either determines to choose an appropriate controller or it suggests an action to be taken to cater to the future needs. The Table I below shows the action to be taken based on the context pattern generated. The mapping between the Context Pattern values and the Controller to be chosen or the action to be taken is stored as control selection rules in the Data Architecture block in Fig 1.

TABLE I : CONTEXT PATTERNS CONTROLLER MAPPING

Sl. No	**Context Patterns**		
	Pattern Identified	***Controller Type***	***Action***
1.	Sudden Variations	Fuzzy	
2.	Quick Adaptation	PID	
3.	Gradual Increase	Time-Series	
4.	Constant	Time-Series	
5.	Resource shortage		New Server Instance
6.	Drastic Performance Degrdation		System Audit

B. *Data Organisation – Performance Data*

The model templates defined for different server components will have a lot of data associated during runtime. It is desired that there is a faster data query mechanism, and the growing data volume should not be an overhead on the self-managing architecture.

The first challenge of faster data access is addressed by choosing a column oriented database [21]. It is used to capture the data that provides data sequence for a given parameter which is a useful input for easier data analysis and identifying the patterns. The second challenge is handled by moving the data to a persistent storage periodically. The data analyzer generates context patterns periodically, generates rules from those context patterns, capture in the knowledge base. Such rules represent the patterns associated with the data and hence the data will be moved to a persistent storage. Thus the amount of data associated to study the past data; the mechanism to analyze and predict the future patterns is much simpler and light weight mechanism.

The following are the different data elements involved based on the model templates:

- Input parameters
- Output parameters
- Model Parameter Constants
- Controller Gains

The Column oriented database supports SQL [22] and fetches the data column wise rather than row wise thereby providing faster data access. Such a data structure enables the Pattern Analysis algorithms to easily extract the behavior. The following Table II below shows a sample Table structure of a column oriented database for a JMS Provider.

TABLE II : JMS PROVIDER COLUMN ORIENTED DATA

Message Throughput	*Subscribers*	*Publishers*	*brokers*	*CPU*

It is easy to add additional input parameters at runtime when it is observed that there are other factors affecting the system performance.

C. Data Analysis - Knowledge base

Data Analysis and Pattern Identification is another important functional building block of the Data Architecture shown in Fig 1. Once the performance data is available in the database, through generic data access layer, the Data Analyzer fetches the data. Initially the knowledge base is empty and every time a Context Pattern is generated it is written to the knowledge base. We used a simple custom rule mechanism which is XML based that stores condition and action to be taken. The following Fig 6 shows a XML rule template Context Pattern used to store as a rule.

```
<rule>
 <name> </name>
 <component> </component>
  <if>
   <property-1>
    <name> </name>
    <value> </value>
   </property-1>
   <property-2>
    <name> </name>
    <value> </value>
   </property-2>
   <property-3>
    <name> </name>
    <value> </value>
   </property-3>
   <property-n>
    <name> </name>
    <value> </value>
   </property-n>
  </if>
 <then>
  <ControllerType></ControllerType>
 </then>
</rule>
```

Fig. 6 : Sample Rule Template

There is a rich collection of Pattern Recognition [23] and Data Mining [24] techniques and we use some of them in our solution such as Time-Series, Episode Discovery, Outlier Analysis, and Association Rules.

D. Model and Data Organisation - Algorithm

The following Fig 7 shows the algorithm of creating templates the knowledge base is created. The control loop is run periodically and in each loop 'n' columns of the output parameter is analyzed. After each loop context pattern is transformed into a rule, written to the knowledge base.

- Create model templates
- Set the periodicity of control loop (p)
- For every interval 'p'
  - Read 'n' columns of output param
  - Analyze the data and generate Context Pattern(prediction algorithms are run)
  - Convert the Context Pattern into rules of the knowledge base
  - After 'm' control loops, call Data Mover to move all the performance data into a persistent storage

Fig. 7 : Data Organisation Algorithm

VI. IMPLEMENTATION

A primitive implementation of the proposed solution is implemented using Java and XML.

A. Data Model

The Data Modeler is a simple Java class implemented to read model templates and create tables in the column oriented database. We have used MonetDB [25] to store the model templates. But we are exploring to identify a light weight in memory column database. Currently the XML model templates have to be created manually. A tool implementation is in progress that allows creating the templates easily.

B. *Knowledge base*

A simple custom rule engine is developed that stores the knowledge base in XML format. A set of java classes are implemented to create the rules using the context pattern object generated. The following Fig 8 shows a sample knowledge base that is created for a context pattern of JMS Providers.

```
<rule>
 <name>JMS-MsgTput-Pattern1</name>
 <component>jms</component>
  <if>
   <property-1>
    <name>Smax</name>
    <value>125</value>
   </property-1>
   <property-2>
    <name>MsgTput</name>
    <value>160</value>
   </property-2>
   <property-3>
    <name>Pattern</name>
    <value>SuddenVariation</value>
   </property-3>
   <property-4>
    <name>CPU</name>
    <value>64</value>
   </property-4>
   <property-4>
    <name>ThresholdViolation</name>
    <value>Yes</value>
   </property-4>
  </if>
 <then>
  <ControllerType>Fuzzy</ControllerType>
    <NwSrvrInstance>False</NwSrvrInstance>
  <modelparam-1>1</modelparam-1>
  <modelparam-2>0.28</modelparam-2>
 </then>
</rule>
```

Fig. 8 : A sample Rule for JMS Provider tuning

VII. DISCUSSION AND FUTURE WORK

In this paper we proposed control system model and data architecture that provides templates to represent the various server components of the Application Server. Also, the solution exploited the mechanisms to capture, organize and use the data associated with these models for effective prediction of the future patterns of the components. The current solution is a subset of the bigger research problem that we are trying to address dealing with development of a generic end-to-end framework that enables in creating robust Application Servers that are more adaptive and self-managing in performance management. We have implemented various controllers for the JEE server components with encouraging results [10][11]112] but most of the implementation is simulation based. We intend to extend our work in validating our theory by running in actual Application Server environments. The end goal our research is to enable the design, modeling and runtime of Application Servers with inherent self-managing capabilities. We have implemented a prototype solution that supports control system concepts as first class elements in UML modeling [26]. We inted to integrate the model templates creation discussed in this paper during the UML design of the server components.

Additionally we want to explore Java based rule engines such as Jess [27] and evaluate against our custom rule engine. The proposed architecture implementation is primitive and it requires enahancement to complete the implemenation of all the components shown and evalute the results in different run time environments.

REFERENCES

[1] Mohammad Reza Nami, Koen Bertels, "A Survey of Autonomic Computing Systems", ICAS '07

[2] What Does Control Theory Bring to Systems Research? Xiaoyun Zhu, Mustafa Uysal, Zhikui, Wang , Sharad Singhal, Arif Merchant, Pradeep Padala, Kang Shin, ACM SIGOPS Operating Systems Review, Volume 43 Issue 1, January 2009

[3] Ravi Kumar G, C.Muthusamy, A.Vinaya Babu, "Control Systems application in Java based Enterprise and Cloud Environments – A Survey", IJACSA Vol 2 No 8, 2011

[4] N. Gandhi and D. M. Tilbury, Y. Diao, J. Hellerstein, and S. Parekh "MIMO Control of an Apache Web Server, Modeling and Controller Design", IEEE American Control Conference, 2002

[5] C. Lu, T.F. Abdelzaher, J.A. Stankovic and S.H. Son, "A feedback control approach for guaranteeing relative delays in web servers" Proc. of the 7th IEEE Real-Time Technology and Applications Symposium, pp 51-62, 2001

[6] Tarek Abdelzaher. Yina Lu, Ronahua Zhana, Dan Henriksson, "Practical Application of Control Theory to Web Services", American Control Conference, 2004

[7] Ying Lu, Avneesh Saxena and Tarek E Abdelzaher Differentiated Caching Services; A Control-Theoretical Approach, IEEE International Conference on Distributed Sysytems, 2001

[8] Ying Lu, Tarek Abdelzaher and Gang Tao, "Direct Adaptive Control of A Web Cache System", Proceedings of the American Control Conference, Denver, Colorado, 2003

[9] Yan Zhang, Wei Qu, Anna Liu, "Adaptive Self-Configuration Architecture for J2EE-based Middleware", Vol 9, HICSS"06

[10] Ravi Kumar G, Dr.Chelliah Muthusamy and Dr.A.Vinaya Babu , "Self-regulating Message Throughput in Enterprise Messaging Servers – A Feedback Control Solution" , IJACSA, Volume 3, No 1, Jan2012

[11] Ravi Kumar Gullapalli, Dr.Chelliah Muthusamy, Dr.A.Vinaya Babu and Raj N. Marndi, "A FEEDBACK CONTROL SOLUTION IN IMPROVING DATABASE DRIVER CACHING", IJEST, Vol 3, No 7, July 2011

[12] Ravi Kumar G, Dr.Chelliah Muthusamy, Dr.A.Vinaya Babu and Raj N. Marndi, "Autonomic Database Driver – An Adaptive Control Solution", Proc. ICITEC 2012, Mar 2012,pp 40-44,

[13] Ravi Kumar G, Dr.Chelliah Muthusamy, Dr.A.Vinaya Babu "Intelligent Application Servers – A Vision of Self-managing of performance" – unpublished, accepted in ICAdC 2012

[14] Ravi Kumar G, Dr.Chelliah Muthusamy, Dr.A.Vinaya Babu "SELF-MANAGING THE PERFORMANCE OF DISTRIBUTED COMPUTING SYSTEMS – AN EXPERT CONTROL SYSTEM SOLUTION" - unpublished

[15] Z.Vukic, Ognejen Kuljaca:"Lecture on PID Controllers", http://arri.uta.edu/acs/jyotirmay/EE4343/Labs_Projects/pidcontrollers.pdf , Apr 2002

[16] ARMA: http://en.wikipedia.org/wiki/ Autoregressive_moving_average_model

[17] JEE Specification : http://www.oracle.com/ technetwork/ java/javaee/tech/index.html"

[18] JMS Providers :http://en.wikipedia.org/wiki/ Java_Message_Service

[19] EJB:" http://en.wikipedia.org/wiki/ Enterprise_JavaBeans"

[20] System Analysis and Modeling: "http://en.wikipedia.org/wiki/System_analysis"

[21] DJ.Abadi, PA.Boncz, S. Harizopoulos, "Column-oriented Database Systems", Proc ACM, VLDB'09

[22] SQL:, http://en.wikipedia.org/wiki/SQL

[23] Robert J. Scholkoff "Pattern Recognition: Statistical, Structured, Neural Approaches", John Wiley 1992

[24] Jiawei Han, Micheline Kamber, , "Data Mining – Concepts and Techniques", Morgan Kaufmann Publishers, 2006

[25] MonetDB: "http://www.monetdb.org/"

[26] Ravi Kumar Gullapalli, Dr.Chelliah Muthusamy, Dr.A.Vinaya Babu, "Design and Modeling Autonomic aware Software in UML – A Control System Solution", ICCIT,July 2012 - inpress

[27] Jess Rule Engine , "http://www.jessrules.com/"

◈◈◈

Decision Tree-Rough Set Based System with HHMM For Predicting the Stock Market Trends

Shweta Tiwari & Rekha Pandit

Dept. of Computer Science & Engineering, RGPV University Bhopal, India
E-mail : shwetatiwari2601@gmail.com, rekhapandit@rediffmail.com

***Abstract* -** Stock Market is an organized market, where shares are issued and traded. These shares are either traded through Stock exchanges or Over-the-Counter in physical or electronic form. Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on a financial exchange. The successful prediction of a stock's future price could yield significant profit. Prediction of stock market by data mining techniques has been receiving a lot of attention recently, to those who wish to profit by trading stocks in the stock market and for researchers attempting to uncover the information hidden in the stock market data. This paper presents a decision tree- rough set based system for predicting the trends in the Bombay Stock Exchange (BSESENSEX) with the combination of Hierarchical Hidden Markov Model. For extracting features Technical indicators are used in the present study from the historical SENSEX data. CART is used to select the relevant features and a rough set based system is then used to induce rules from the extracted features, this supervised data are then evaluated by of Hierarchical Hidden Markov Model.

***Keywords* -** *Stock market, Data mining, CART, Rough set, HHMM, technical indicator.*

I. INTRODUCTION

Around the world, trading in the stock market has gained enormous popularity as a means through which one can reap huge profits. Attempting to successfully and accurately predict the financial market has long attracted the interests and attention of economists, bankers, mathematicians and scientists alike. Thus, stock price movement prediction has long been a cherished desire of investors, speculators and industries. Stock Market is an organized market where shares are issued and traded. These shares are either traded through Stock exchanges or Over-the-Counter in physical or electronic form. Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on a financial exchange. The successful prediction of a stock's future price could yield significant profit. However, according to the efficient market hypothesis, all such attempts at prediction are futile as all the information that could affect the behavior of stock price or the market index must have been already incorporated into the current market quotation. There have been several studies, which question the efficient market hypothesis by showing that it is, in fact, possible to predict, with some degree of accuracy, the future behavior of the stock markets. Technical analysis has been used since a very long time for predicting the future behavior of the stock price. Different technical indicators are used in the present study to extract information from the financial time series and hence, they act as the features that are used for stock market trend prediction.

The financial markets form the bedrock of any economy. There are a large number of factors and parameters that influence the direction, volume, price and flow of traded stocks. This coupled with the markets' vulnerability to external and nonfinancial related factors and the resulting intrinsic volatility makes the development of a robust and accurate financial market prediction model an interesting research and engineering problem. In this paper a CART decision tree is used to select the relevant features (technical indicators) from the extracted feature set. The selected features are then applied to a rough set based system for predicting one-day-ahead trends in the stock market. These trends are evaluated by HHMM and final predictions are generated.

II. LITERATURE SURVEY

Data mining techniques and Artificial neural networks have been widely used prediction of financial time series. In [1] Technical indicators and rough-set based system were used to predict on-day-ahead trend of SENSEX. In [4], the effectiveness of time delay, recurrent and probabilistic neural networks for prediction of stock trends based on historical data of the daily closing price was attempted. In [5] technical indicators were used as inputs to a feedforward neural network for prediction of Taiwan Stock Exchange and NASDAQ. In [6], technical indicators and a backpropagation neural network was used to create a decision support system for exchange traded funds trading. Technical indicators and neural networks were used in [7] to predict the US Dollar Vs British Pound exchange rates. In [8] a framework for intelligent interaction of automatic trading algorithms with the user was presented. In [9] a back propagation neural network was employed to predict the buy/sell points for a stock and then applied a case based dynamic window to further improve the forecast accuracy. In [3] a survey of more than hundred articles which used neural networks and neuro-fuzzy models for predicting stock markets was presented. It was observed that soft computing techniques outperform conventional models in most cases. Defining the structure of the model is however, a major issue and is a matter of trial and error. In [10], review of data mining applications in stock markets was presented. [11], used a two-layer bias decision tree with technical indicators feature to create a decision rule that can make recommendations when to buy a stock and when not to buy it. [12] combined the filter rule and the decision tree technique for stock trading. In[13] a hybrid fuzzy time series model with cumulative probability distribution approach and rough set rule induction was proposed, to forecast stock markets. Cumulative probability distribution approach was used to discretize the observations in training datasets based on the characteristics of data distribution and rules were generated using rough set algorithm. Forecasting was then done based on rule support values from rough set algorithm.

III. PROPOSED SYSTEM

The trend prediction system proposed in this paper works in the following way: First the features are extracted from the daily stock market data. Then the relevant features are selected using decision tree. A rough set based classifier is then used to predict the next day's trend using the selected features. Then these trends are evaluated using HHMM and final predictions will generate. In the present study, the prediction accuracy of the proposed system is validated using the Bombay Stock Exchange Sensitive Index (BSE-SENSEX or commonly, SENSEX) data. The performance of trend prediction systems are evaluated using the cross validation method.

IV. DATA MINING

A. *Definition of data mining*

Data mining is one of the most cutting-edge researches in the field of the current international databases and information decision-making. From a technical point of view, it refers to extracting the previously unknown, the potentially useful patterns or knowledge from large databases including association rules, time series, artificial intelligence, statistics, databases, etc. The term *data mining* has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. The phrase *knowledge discovery in databases* (KDD) refers to the overall process of discovering useful knowledge from data, and *data mining* refers to a particular step in this process (Fayyad et.al., 1996). In this research, data mining techniques will be applied to the data on stock market in order to predict the next day's stock market scenario.

B. *The necessity of data mining in stock market*

Things that are likely to affect the price of a stock include: 1) What people expect its future dividends will be. 2) When the dividends are expected to be paid. 3) The amount of risk involved. Data Mining is a process of abstracting unaware, potential and useful information and knowledge from plentiful, incomplete, noisy, fuzzy and stochastic data. These information and knowledge can't be achieved relying on a simple data search. The key of data mining include three parts: data, information and business decisions. Data is the most valuable only when mobilized or converted into useful information. Accessing to data is not the ultimate goal of data mining. In fact, the final aim of data mining is using that information to improve business decision-making efficiency and to develop more appropriate decisions. The stock market data is stream data, at the same time, the stock market data shares sequential nature, which can be used to analyze stream data time-series pattern mining methods. In addition, there are many affecting factors of stock prices, making the price data show non-linear features, which bring new challenges to the traditional data mining algorithms.

V. STOCK MARKET

Stock Market is an organized market where shares are issued and traded. These shares are either traded through stocks exchange or Over-the-Counter in physical or electronic form.

Stock market can be divided into 2 parts:

1) Primary Market
2) Secondary Market

Primary Market deals with securities that are channelized through the Initial Public Offer (IPO) route. After the assurances of the stocks to the general public, these stocks are then bought and sold by the investors between themselves in the secondary market. Here the stock issuing corporation has no direct influence on these trades.

Stocks in the stock market are either traded through Stock Exchanges or Over-the-Counter. Stock Exchange is organized market place where stocks bonds are other equivalents are traded between buyers and sellers. The contracts are standardized ones. But in case of OTC, the trade takes place through a network of dealer and the contracts are bilateral customized ones.

The platforms through which the stocks are traded are:

a) Offline stock trading: In this customer has to place order to the dealer of the stock broking firm either in person or over phone.
b) Online stock trading: Whereas here the cline could place his order on his own from any place he wants, provided he has computer with an internet connection.

Whatever the time horizon is, trades have in common uncertainty related to future market movements. Although this characteristic is highly undesirable for the investor, it is also unavoidable whenever the stock market is chosen as an investment tool. In addition, automated trading requires instruments able to manage and reduce uncertainty on quantitative basis. Therefore, prediction (forecasting) of future market movements is one step in the process of designing a reliable automated trading strategy.

VI. TREND PREDICTION

In literature a number of different methods have been applied in order to predict stock market movements and trends. In general, methods can be classified in four major categories: (i) Technical Analysis, (ii) Fundamental Analysis, (iii) Traditional Time Series Forecasting and (iv) Machine Learning Methods. Technical Analysis is probably the most common approach to trend forecasting. A large literature is available. Technical analysis makes use of composite functions, such as *indicators* and *oscillators*, derived by time series, and heuristic rules able to reveal signals of change in the market trends. Popular examples of methods are Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), Stochastic oscillator. This approach relies on the belief that markets are mostly driven by psychology, more than economics.

VII.TECHNICAL INDICATORS

In finance, technical analysis is security analysis discipline for forecasting the direction of prices through the study of past market data, primarily price and volume. Behavioral economics and quantitative analysis build on and incorporate many of the same tools of technical analysis, which being an aspect of active management, stands in contradiction to much of modern portfolio theory. The efficacy of both technical and fundamental analysis is disputed by efficient-market hypothesis which states that stock market prices are essentially unpredictable.

Technicians employ many techniques, one of which is the use of charts. Using charts, technical analysts seek to identify price patterns and market trends in financial markets and attempt to exploit those patterns. Technicians use various methods and tools, the study of price charts is but one. Technicians using charts search for archetypal price chart patterns, such as the well-known head and shoulders or double top/bottom reversal patterns, study technical indicators, moving averages, and look for forms such as lines of support, resistance, channels, and more obscure formations such as flags, pennants, balance days and cup and handle patterns.

Technical analysts also widely use market indicators of many sorts, some of which are mathematical transformations of price, often including up and down volume, advance/decline data and other inputs. These indicators are used to help assess whether an asset is trending, and if it is, the probability of its direction and of continuation. Technicians also look for relationships between price/volume indices and market indicators. Other avenues of study include correlations between changes in options (implied volatility) and put/call ratios with price. Also important are sentiment indicators such as Put/Call ratios, bull/bear ratios, short interest, Implied Volatility, etc. Technical analysis employs models and trading rules based on price and volume transformations, such as the relative strength index, moving averages, regressions, inter-market and intra-market price correlations, business cycles, stock market cycles or, classically, through recognition of chart patterns.

Technical analysis stands in contrast to the fundamental analysis approach to security and stock analysis. Technical analysis analyzes price, volume and other market information, whereas fundamental analysis looks at the facts of the company, market, currency or commodity. Most large brokerage, trading group, or financial institutions will typically have both a technical analysis and fundamental analysis team.

VIII. CART

The CART decision tree is a binary recursive partitioning procedure capable of processing continuous and nominal attributes both as targets and predictors. Data are handled in their raw form; no binning is required or recommended. Trees are grown to a maximal size without the use of a stopping rule and then pruned back (essentially split by split) to the root via cost-complexity pruning. The next split to be pruned is the one contributing least to the overall performance of the tree on training data (and more than one split may be removed at a time). The procedure produces trees that are invariant under any order preserving transformation of the predictor attributes. The CART mechanism is intended to produce not one, but a sequence of nested pruned trees, all of which are candidate optimal trees. The "right sized" or "honest" tree is identified by evaluating the predictive performance of every tree in the pruning sequence. CART offers no internal performance measures for tree selection based on the training data as such measures are deemed suspect. Instead, tree performance is always measured on independent test data (or via cross validation) and tree selection proceeds only after test-data-based evaluation. If no test data exist and cross validation has not been performed, CART will remain agnostic regarding which tree in the sequence is best. This is in sharp contrast to methods such as C4.5 that generate preferred models on the basis of training data measures. The CART mechanism includes automatic (optional) class balancing, automatic missing value handling, and allows for cost-sensitive learning, dynamic feature construction, and probability tree estimation. The final reports include a novel attribute importance ranking.

A. Splitting rules

CART splitting rules are always couched in the form. *An instance goes left if CONDITION, and goes right otherwise,* where the CONDITION is expressed as "attribute *Xi* <= *C*" for continuous attributes. For nominal attributes the CONDITION is expressed as membership in an explicit list of values. The CART authors argue that binary splits are to be preferred because (1) they fragment the data more slowly than multi-way splits, and (2) repeated splits on the same attribute are allowed and, if selected, will eventually generate as many partitions for an attribute as required. Any loss of ease in reading the tree is expected to be offset by improved performance. A third implicit reason is that the large sample theory developed by the authors was restricted to binary partitioning. The CART monograph focuses most of its discussion on the Gini rule, which is similar to the better known entropy or information-gain criterion. For a binary (0/1) target the "Gini measure of impurity" of a node *t* is

$$G(t) = 1 - p(t)^2 - \left(1 - p(t)\right)^2$$

where *p(t)* is the (possibly weighted) relative frequency of class 1 in the node, and the improvement (gain) generated by a split of the parent node P into left and right children L and R is

$$I(P) = G(P) - qG(L) - (1 - q)G(R)$$

Here, q is the (possibly weighted) fraction of instances going left. The CART authors favor the Gini criterion over information gain because the Gini can be readily extended to include symmetrized costs (see below) and is computed more rapidly than information gain. (Later versions of CART have added information gain as an optional splitting rule.) They introduce the modified twoing rule, which is based on a direct comparison of the target attribute distribution in two child nodes:

$$I(split) = \left[.25\left(q(1-q)\right)^u \sum_k |pL(k) - pR(k)|\right]^2$$

where *k* indexes the target classes, *pL( )* and *pR( )* are the probability distributions of the target in the left and right child nodes respectively, and the power term *u* embeds a user trollable penalty on splits generating unequal-sized child nodes. They also introduce a variant of the twoing split criterion that treats the classes of the target as ordered; ordered twoing attempts to ensure target classes represented on the left of a split are ranked below those represented on the right. In our experience the twoing criterion is often a superior performer on multi-class targets as well as on inherently difficult-to-predict (e.g. noisy) binary targets. For regression (continuous targets), CART offers a choice of Least Squares (LS) and Least Absolute Deviation (LAD) criteria as the basis for measuring the improvement of a split. Three other splitting rules for cost-sensitive learning and probability trees are discussed separately below.

IX. FEATURE EXTRACTION

The inputs of the proposed system are the daily open, high, low, close trading volumes of the SENSEX and technical indexes. Nine commonly used technical indicators are considered initially. Three types of indicators, namely, volume based, price based and overlay indicators were chosen. Since the proposed system is capable of choosing the relevant features automatically, the exact number of indicators is not of much importance, only those indicators which contain any relevant information will be selected by the system. Literature on the technical indices considered above, is widely available. Classification of the trend into up, down and no trend was done in the following way: The market is formally classified as being in an uptrend (downtrend) when all the following conditions are satisfied:

1. The closing value must lead (lag) its 25 day moving average
2. The 25 day moving average must lead (lag) 65 day moving average.
3. The 25 day moving average must have been rising (falling) for at least 5 days.
4. The 65 day moving average must have been rising (falling) for at least 1 day.

If the movement of the market cannot be classified as either an uptrend or a downtrend, it is assumed that there is no trend in the market movement. The above information gives the present day's trend and is also used to predict the next day's trend. The aim of all the trend prediction techniques considered in the present study, is to identify the relationship between the information variables and the decision variable. All the technical indexes considered, may or may not be necessary to arrive at the decision. It is also possible that some features may have a negative impact on the accuracy of the system. Hence, feature selection is used to identify those features (technical indexes) which are necessary for trend prediction and help in maintaining or improving the prediction accuracy. Features which are not selected are deemed unnecessary and are discarded.

X. ROUGH SET BASED TREND PREDICTION

Rough sets are extremely useful in dealing with incomplete or imperfect knowledge. This section presents a brief overview of the rough set concepts used in the present study and also presents their application for stock market trend prediction. A rough set uses the concepts of lower and upper approximations to classify the domain of interest into disjoint categories. The lower approximation is a description of the domain objects that are known with certainty to belong to the subset of interest, whereas the upper approximation is a description of the objects that possibly belong to the subset.

A. *Information and decision system*

In rough set theory knowledge is represented in information tables. Each row in the table represents an object. Each column in the table represents an attribute (also referred to as feature, in this study), in the present case, each attribute is a technical index obtained from the historical SENSEX data.

An information system Λ is defined as:

$$\Lambda = (U, A)$$

U= Non empty finite set of objects called the universe

A= Non empty finite set of features such that

$$\alpha : U \rightarrow V_x \forall \alpha \in A$$

And V_x is the value set for α

Inclusion of decision features (attributes) along with the features (attributes) of the information system results in a decision system. If Q is the set of decision attributes, a decision system S can be represented as

$$S = (U, A \cup Q)$$

The elements of A are called conditional attributes or conditions and U is the universal set. The decision attribute for the present study is the trend of SENSEX (up trend, down trend and no trend).

B. *Indiscernibility*

The concept of indiscernibility is used to define equivalence classes for the objects. Given a subset of attributes P A, each such subset defines an equivalence relation IND(P) called an indiscernibility relation.This indiscernibility relation is defined as

$$IND(P) = \{(x, y) \in U^2 \mid \alpha \in P, \alpha(x) = \alpha(y)\}$$

Above equation means that the subset of features, P, will define the partition of the universe into sets such that each object in a set cannot be distinguished from other objects in the set using only the features in P. The sets into which the objects are divided, are called the equivalence classes. The equivalence classes of the indiscernibility relation with respect to P are denoted $[x]P$,where x U. The partition of U, determined by IND(P), is denoted U/IND(P) or U/P , which is simply the set of equivalence classes generated by IND(P).

C. *Set approximation*

Given an information system, $\wedge$ = (U, A), the set X U,

can be approximated using only the information in the set P using the P-upper and P-lower approximations of X. P-Lower approximation of X:

$$\underline{P}X = x \mid \overline{x}_P \subseteq X$$

The lower approximation is the set containing all objects for which the equivalence class corresponding to the object is a subset of the set. This set contains all objects which certainly belong to the set *X*. P-Upper approximation of X:

$$\overline{P}X = x \mid \overline{x}_P \cap X \neq 0$$

XI. HIERARCHICAL HIDDEN MARKOV MODELS

The hierarchical hidden Markov model (HHMM) is an extension of the hidden Markov model, in which states are organized in a hierarchy. In other terms, HHMMs are structured multi-level stochastic processes. They generalize ordinary HMM by making states probabilistic models on their own. Therefore, HHMMs are recursively defined, so that each state at level l relies on an HHMM of layer l+1. When a state in an HHMM is activated, it becomes active also its own probabilistic model and one of the states of the underlying HHMM is activated recursively. This process is repeated until a *production state*, which is a state that emits a single observation symbol, is activated. The states that do not directly emit observations symbols are called *internal states*. Production states do not hold a sub-model, and they are not able to transit to other states. So, when the production state is reached, a symbol of sequence is produced, the control goes back to the calling state, that in turn gives the control to another state at the same level. When a *terminal state* is reached, the control is moved to the upper level. An example of HHMM structure is given in Fig.1.

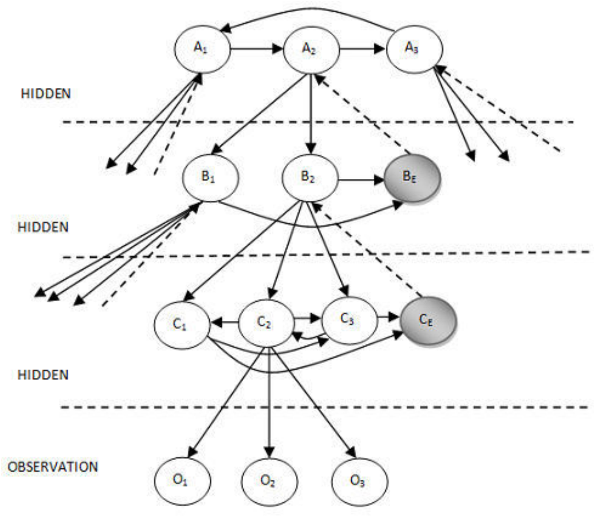


Fig. 1

States are represented by circles, and transitions between states by arrows. In particular, black solid lines show intra-level transitions. For our convenience, we only provided transitions with non-zero probability, but it is possible to transit from any state to another within a level. Transitions from upper to lower level are denoted by fine-dashed lines. Transitions back to upper levels are denoted by dashed lines. Bold circles are the production states, which are the solely allowed to produce a symbol $\sigma_k \in \Sigma$, Σ is the alphabet of emitted symbols, observed in sequences. Gray circles are terminal. The other circles are internal.

Formally an HHMM is defined as a pair <G, λ>, where G is D-layer directed graph, and $\lambda \equiv \{\lambda_d\}_{d=1..}D$ a collection of parameter, specific of each layer d. The graph G is made of inner, production and terminal states. An inner state at level d of an HHMM is denoted by q_d, while $q^d{}_E$ is the terminal state and We collect the inner states and the production state at level d in $Q_d \equiv \{q_i^d, q_E^d\}$. It is not required an internal state to have the same number of substates, although any HHMM can be transformed into a model with an equal number of sub-states for each internal state. For each internal state $q^d{}_i$ there is a probability transition matrix $A^{q^d} = \{a_{ij}^{q^d}\}$, where $a_{ij}^{q^d} = \Pr(q_j^{d+1} \mid q_i^{d+1})$ is the probability of moving $o^d{}_k$ is a production state emitting the symbol $\sigma_k \in \Sigma$. from i-th to j-th sub-state of q^d. Similarly, $\pi_i^{q^d} = \Pr(q_i^{d+1} \mid q_d)$ is the initial probability assigned to sub-states by q^d. It can be regarded as the probability of performing a vertical transition, that is the probability by which the state q^d activates the sub state q_i^{d+1} . We denote the activation probability distribution by $\pi^{q^d} = \{\pi_i^{q^d}\}$ .Finally, $B^{q^d} = \{b^{q^d}(o_k)\}$ is the probability that internal state q will activate the production state o_k^{d+1} , which in turn will emit the symbol σ_k. Therefore $b^{q^d}(o_k)$ is also the probability that the symbol $\sigma_k \in \Sigma$ is produced when the state q_d is activated. In summary, we get $\lambda_d = \{A^{q^d}, \pi^{q^d}, B^{q_E^d}\}_{q^d, q_E^d \in Q_d}$. States and observations can be either discrete or continuous. Similarly to ordinary hidden Markov models, HHMMs are useful in applications dealing with sequences of symbols, such as in signal identification and classification, behavior recognition, handwritten character recognition, text analysis, and other pattern recognition problems. Indeed, it can be proven that each HHMMs can be transformed into an equivalent HMM. But, the structure in layers of HHMM can be exploited in order to adopt more efficient inference algorithms and robust learning algorithms. More specifically, HHMMs can be employed to (i) calculate the likelihood of a sequence, that is the probability $\Pr(\overline{O} \mid \lambda)$ of a sequence $\overline{O}$ to be generated by the model λ; (ii) to find the most probable state sequence $S^* = \arg\max_s \Pr(\overline{O} \mid \lambda, S)$, given a sequence $\overline{O}$ and the model λ; (iii) to find the most probable states and observations given a partial subsequence of symbols.

XII. CONCLUSION

Predicting the future is one aspect in designing profitable day trading strategies. Technical analysis stands in contrast to the fundamental analysis approach to security and stock analysis. Technical analysis analyzes price, volume and other market information, whereas fundamental analysis looks at the facts of the company, market, currency or commodity. Most large brokerage, trading group, or financial institutions will typically have both a technical analysis and fundamental analysis team. Here in this paper technical analysis is considered for the study. The design of decision tree-rough set based stock market trend prediction system with the combination of HHMM for predicting the future trend of the SENSEX is presented in this paper. Features are extracted from the historical SENSEX data by using CART. Extracted features are used to generate the prediction rules using rough set theory because rough sets are extremely useful in dealing with incomplete or imperfect knowledge. HHMMs are useful in applications dealing with sequences of symbols, such as in signal identification and classification, behavior recognition, handwritten character recognition, text analysis, and other pattern recognition problems. Trends derived from rough set are then evaluate by HHMM and final predictions are generated.

REFERENCES

[1] Binoy.B.Nair, V.P Mohandas and N. R. Sakthivel,' A Decision Tree- Rough Set Hybrid System for Stock Market Trend Prediction', International Journal of Computer Applications (0975 – 8887)

[2] Luigi Troiano and Pravesh Kriplani,' Predicting Trend in the Next-Day Market by Hierarchical Hidden Markov Model', 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM),pg:201-204

[3] Atsalakis G. S., and Valavanis K. P., (2009), „Surveying stock market forecasting techniques – part II: soft computing methods‟, Expert Systems with Applications, vol.36, pp. 5932–5941.

[4] Saad E. W., Prokhorov D.V., and Wunsch, D.C., (1998), „Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks‟, IEEE Transactions on Neural Networks, Vol. 9, No. 6, pp. 1456-1470.

[5] Lee, C-T., and Chen,Y-P. 2007. The efficacy of neural networks and simple technical indicators in predicting stock markets. In Proceedings of the International Conference on Convergence Information Technology, pp.2292-2297.

[6] Kuo, M-H., and Chen, C-L.2006. An ETF trading decision support system by using neural network and technical indicators. In Proceedings of the International Joint Conference on Neural Networks, pp. 2394-2401.

[7] Nagarajan,V., Wu,Y., Liu,M. and Wang Q-G. 2005. Forecast Studies for Financial Markets using Technical Analysis. In Proceedings of the International Conference on Control and Automation (ICCA2005), pp. 259-264.

[8] Bansal, Archit, Mishra ,Kaushik, Pachouri, Anshul, (2010), „Algorithmic Trading (AT) - Framework for Futuristic Intelligent Human Interaction with Small Investors‟„ International Journal of Computer Applications, vol. 1, no. 21, pp.01-05.

[9] Chang, P.-C. , Liu, C.-H. , Lin, J.-L. , Fan, C.-Y. , & Celeste, S.P. Ng., (2009) , „A neural network with a case based dynamic window for stock trading prediction‟, Expert Systems with Applications, vol.36, pp.6889–6898.

[10] Setty, Venugopal D., Rangaswamy, T.M. and Subramanya, K.N.,(2010), „A Review on Data Mining Applications to the Performance of Stock Marketing‟, International Journal of Computer Applications, vol. 1, no. 3,pp.33-43.

[11] Wang, J-L. , and Chan, S-H.,(2006), „Stock market trading rule discovery using two-layer bias decision tree‟, Expert Systems with Applications, 30, pp.605–611.

[12] Wu, M-C., Lin, S-Y., & Lin, C-H.,(2006) , „An effective application of decision tree to stock trading‟, Expert Systems with Applications, 31, pp.270–274.

[13] Teoh, H. J., Cheng,C-H., Chu, H-H., and Chen, J-S., (2008), „Fuzzy time series model based on probabilistic approach and rough set rule induction for empirical research in stock markets‟, Data & Knowledge Engineering, 67, pp.103–117.

[14] S. Fine, "The hierarchical hidden markov model: Analysis and applications," Machine Learning, vol. 32, pp. 41–62, 1998.

◈◈◈

Computer Aided Visualization Model for Speech Perception and Intelligibility

A Rathinavelu [AR] & M. Karthikeyan

CSE Dept., Dr. Mahalingam College of Engineering and Technology, Pollachi-642003, TamilNadu, India
E-mail : starvee@drmcet.ac.in, mkarthime@gmail.com

Abstract - The aim of this research is to increase the phonetic skills of the early language learners and second language learners by visualizing the inner articulators such as tongue and lips during the pronunciation of Tamil phonemes. Speech perception interface helps to increase the perception capability of the subjects for new words and sounds. Computer aided intelligibility model is developed to evaluate the performance of the subjects after the completion of training. MRI is used to extract the parameters from the speech sound. Using polygons the tongue and lip models are designed. Tongue is modeled using 86 polygons, 49 control points which are arranged in 7×7 grid. Lip model is designed using 28 polygons. Model has 42 control points which are arranged in 6×7 grid. Control points are identified for the articulation. Speech perception interface is developed to provide the perception exercise to the subjects. Using minimal pair with images, the perception interface is presented to the subjects. Perception interface will give better user interaction and giving the feedback to the subject. Computer aided intelligibility model has the interface with the images. Assessor evaluates the performance of the subjects and provides mark to the individual subject. To provide better look and feel to the interface, Human Interface Guidelines (HIG) are used for designing the interface. Computer aided visualization model is found to be more interactive with the subject and it will increase the phonetic skill of the subject. With the help of our system early language learners and second language learners can easily perceive the Tamil phonemes and words.

Keywords - *Articulation, Speech perception, Control points, Minimal pair, Human Interface Guidelines, Intelligibility.*

I. INTRODUCTION

Computer aided visualization of internal speech articulators will provide the better training of speech perception to the subjects. Visualization of speech is one of the complex tasks. The idea behind the computer aided speech visualization is to visualize the articulation of the inner articulators. Visualization of inner speech articulators such as lips and tongue is useful for the subjects to learn the movements of the inner articulators for a pronunciation of a word or a sound. Front and side view of an animated face will be presented to the subjects. Speech production of the subjects will be good, only if they effectively perceive the speech sounds from the computer aided visualization model. Perception exercise can help us to achieve good speech production. Perception exercise provides effective method to perceive new words and sounds. Intelligibility of a subject would be better, only when the system provide good teaching process. Intelligibility experiments [1] are needed to evaluate the intelligibility of the subject after training process has been finished.

This research focused on the visualization of inner articulators; provides the perception exercise to the subjects. After training is accomplished the performance of the subjects can be evaluated by conducting intelligibility experiments on the subjects. Human Interface Guidelines [2] (HIG) are included in the interfaces.

Perception can be done commonly by two ways [3]. Minimal pair of words called minimal pair and the other one is lexical stress pattern. Intelligibility experiments can be done in many ways. Manual intelligibility experiments are done by asking the questionnaires to the subjects after completing the perception. Next type is that the intelligibility is calculated by using various calculations on different measures [4]. This is a complete automated process carried out by the system. Another method contains intervention of both system and manual.

Remaining part of the paper is organized as follows. Section II gives the summary of the related work regarding this research. Section III contains the overall system design and implementation information. Section IV gives you the future work for the system evolution.

II. RELATED WORK

This part of the paper provides the related work in the context of modeling the inner articulators, speech perception interface and intelligibility experiments.

Early language learners and second language learners will learn the pronunciation of new words and sounds with the help of Computer aided visualization interface [5] that contains visualization of inner articulators. Even though they have the visualization interface, subjects are still lacking in the speech production. To overcome this difficulty they need some speech perception interface for their practice. Also experimental results used to evaluate the subjects performance based on their speech production.

To model the 3D animated face, 3D wireframe structure is commonly used [6] [7] [8]. Magnetic Resonance Imaging (MRI), Electromagnetic Articulatory (EMA) coils with Qualysis-Movetrack (QSMT), and EPG are strategies to extract parameters of the inner articulators. In EMA [8] electromagnetic field is used to track movements of smaller receiver coils glued on the speech articulators, QSMT is the Qualisys Motion Capture System in which Infrared cameras are used which will record the 3D positions of reflective markers placed on the subject. [9] Kinematic data from Electropalatography (EPG) are the common methods to extract the parameters of tongue, lips to visualize the inner articulators. EPG provides the custom-made artificial palate which is moulded to fit against the speaker's hard palate. Artificial palate contains electrodes. When contact occurs between tongue surface and electrodes electronic signals are sent to an external processing unit. MRI is amenable to computerized three dimensional modeling and provides better structural differentiation. Based on the control points of the tongue and lips, the articulator models are visualized to the subjects. MRI [5] [9] strategy is the simple video capturing technique during the speaker pronouncing the word or phoneme.

Drawback of EMA with QSMT technique is, since electronic magnetic field and receiver coils are implemented on the speaker's inner articulators, it may produce health risks to the speaker. EPG also have the disadvantage that, it may cause some health risks to the subject since it uses the electrodes on the artificial palate. Since MRI is the simple video capturing technique it will not cause any side effects to the speaker.

Animated agent ville provides the speech perception and production exercises. With the help of minimal pair the perception exercise is modeled. Ville will say the name among the minimal pair, and then user has to select the name which is said by the ville. Ville gives the verbal feedback about the user's selection. Perception technique will give better training to the subject. Another kind of perception is that reading short stories [10] in corresponding language. It will grow the reading skill of the user.

Reading short stories is difficult for early language learners and second language learners at the initial stage of learning new language. Ville with feedback method uses simple words rather than stories and it is interactive with the user.

Diagnostic rhyme test is performed using monosyllabic words in the form of consonant- vowel-consonant sound sequence [11]. Truncation and Coarticulation method provides the words in the form of consonant-vowel (CV) sound sequence with the truncation either at the end or initial part of the word. [6] Were as in keyword test reproduction of the short sentence by the subject is evaluated. Speech Reception Test (SRT) [12] is conducted based on the signal-to-noise ratio. Another kind of testing is by using Virtual Reality (VR) objects. VR object will be shown to the subject, and then subject have to say the name of the image which is shown. It is the simple technique for the implementation.

Difficulty in truncation and coarticulation technique is that it is not working well for small words. It is difficult to use the signal-to-noise ratio with respect to this system. Using VR objects intelligibility can be calculated, since there is no need to automate the system completely to evaluate intelligibility.

III. SYSTEM DESIGN

System provides the model of animated talking head that visualizes the inner articulators such as tongue and lips. Model involves front view of lips and side view of lips and tongue. By using seven parameters inner articulators are articulated. Parameters are lip opening and closing, lip protrusion, lip rounding, tongue body raise, tongue front and back, tongue contact with palate, tongue tip raise. Speech perception interface is the next component in the system. It comprises of minimal pair, images of the word and feedback about the subject's selection. Intelligibility testing module provides the experiment on the subject's performance after training has been completed. Help menu is included in both first and second interfaces, to guide the subjects for using the system properly. Help menu gives the detailed description about each interface. Also gives the steps to perform the functions in each interface of the system.

A. Speech User Interface

It comprises of front and side view of the face, Inner articulators in the face are transparent to the subjects. Hence they can watch the movements of the articulators. Tongue consists of 86 polygons with 98 control points arranged in 7×7 grid. Each top and bottom layer of the tongue consists of 49 control points. Among that 7 major control points are selected to animate the tongue model. Lip model consist of 28 polygons with 42 control points arranged in 6×7 grid. 6 major control points are selected to articulate the lip model. For each letter and word, the control point information is extracted from the MRI video. MRI video is recorded while the word or letter is pronounced by [AR]. Interface has list of words (30) and letters (25). After subject choosing a word or letter by the subject the talking head performs the visualization of articulation of inner articulators (tongue and lips). Speech User Interface (Fig.1) helps the subjects to see how the inner articulators are articulated for each Tamil sound or a word. Scroll bars are used to set the values of the seven parameters. Scroll bars are enabled only for the letter articulation not for the word articulation. Scroll bar values can be set automatically by the system during the articulation of Tamil letters. In Fig.1 scroll bar values for lip open & close and tongue body raise are set during the articulation of letter "RA". Subject can also set the values of any parameter by using scroll bar and they can see the movements of tongue and lips which are in the interface.

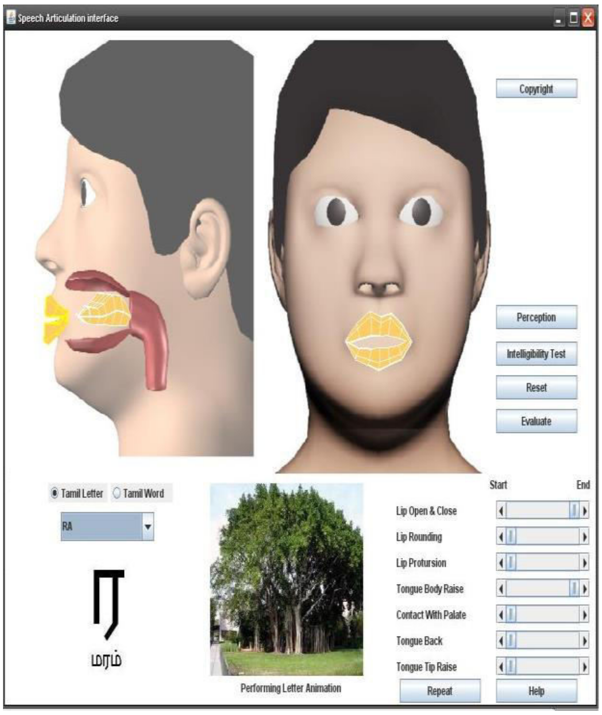


Fig. 1: Speech User Interface

Repeat option can be used by the subject to make the system to repeat the articulation which is recently performed. So the subject can listen the articulation again. Evaluate option is useful for the subject to compare the articulation for the particular phoneme in the computer aided visualization model and the MRI video which is already recorded.

B. Speech Perception Interface

Speech perception interface provides the minimal pair, which consists of set of words. Each set contains 3 words which looks similar in pronunciation but do not have exact pronunciation. Like speech user interface this interface also provides visualization of inner articulators for the word among the 3 words which are displayed in the interface. Mean while, the image of the word is also displayed. Among the 3 words the talking head will say one of the words; there are many set of minimal pairs will be presented to the subject. The subject has to practice the minimal pair word sets one by one. Subject's task is to select the word which is articulated by the talking head. For each selection the system gives the feedback that whether the selected word by the subject is right or wrong. Subject can also make the system to repeat the same set of words in order to listen the articulation of inner articulators again. Speech Perception interface is shown in Fig. 2. Perception exercise will improve the subject's skill to perceive new words or sounds.

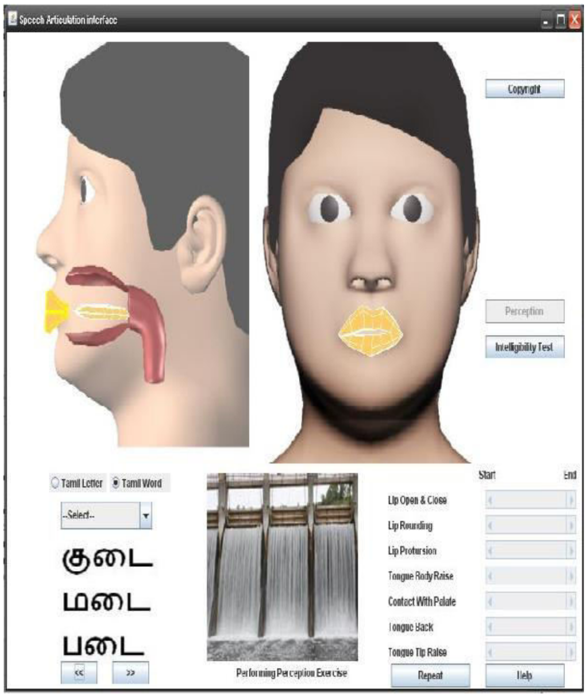
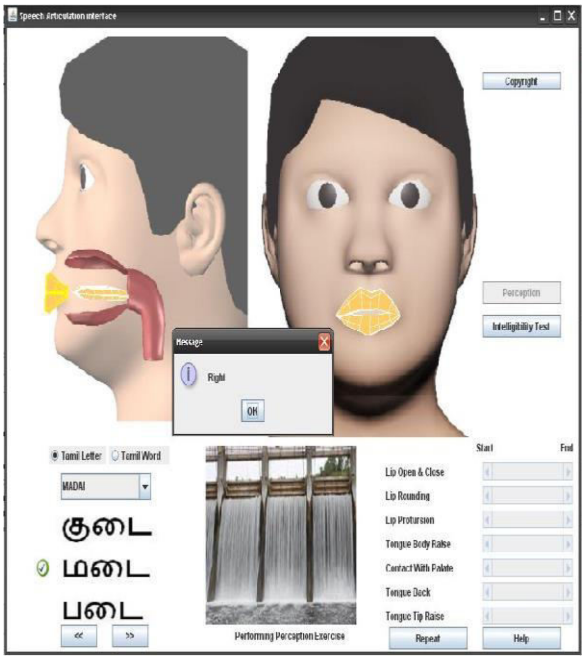


Fig. 2 : Perception Interface

This interface provides the user interaction with the system by giving the feedback (Fig. 3) about the word which is chosen by the subjects, which will help the subjects to evaluate themselves. By providing the repeat option they can also move to the next or previous word set. Assessor record the marks for individual subject based on their ability of perceiving new words and phonemes.



Speech Articulation interface

Message

Right

OK

● Tamil Letter ○ Tamil Word

MADAI

குடை

✓ மடை

படை

Performing Perception Exercise

Fig. 3 : Perception with Feedback

C. *Intelligibility Testing*

Intelligibility of speech refers to the accuracy with which a listener can understand spoken word or phrase. It is a kind of performance evaluation method. Intelligibility will be good only when the subject remembers the words and their pronunciation which are presented in the speech user interface and speech perception interface. The performance of the subjects is evaluated based on their pronunciation. Intelligibility testing interface provides only the images which are presented in speech user and speech perception interface. Subject's task is to say the word of the image which is presented in the interface; subject's pronunciation is recorded by the assessor. Assessor will evaluate the subject's intelligibility based on their pronunciation. Based on 4 categories the intelligibility is evaluated which are,

- Omissions
- Additions
- Substitutions
- Distortions

Above four are the common errors performed by the subject during articulation. With the help of system, these errors can be rectified. System provides better visualization of inner articulators during the articulation of each phoneme and word. Also it makes the subject to perform exercises in the perception interface to perceive the new sounds and words efficiently.

Schedule is planned to give training to the subjects. After the training is completed intelligibility testing is conducted on the trained subjects. The audio is recorded by the assessor while the subject pronouncing the name of the images during the intelligibility testing. Assessor will evaluate the performance of the subject by analyzing their pronunciation.

IV. DISCUSSIONS

This system provides good perception over the new sounds and words to the early language learners and second language learners. Articulation of Tamil phonemes and Tamil words is taught by the computer aided visualization model to the subject. There is a good interaction between the system and the subjects in speech user interface and perception interface. It increases the perception skill of the subjects. By using this interface the subjects can practice to perceive the new words and sounds. System also evaluates the subject's performance by the intelligibility testing method. Research work can be extended by increasing the realism and smoothness of the talking head.

V. ACKNOWLEDGEMENT

This research work is financially supported by Dr.Mahalingam College of Engineering and Technology, Pollachi, South India. We wish to thank the members of college management. Also we wish to thank the faculty member G.Anupriya Assistant Professor (SG) PG (CSE) and G.Yuvaraj, Alumni of PG (CSE) for their encouragement.

REFERENCES

[1] A Rathinavelu, H Thiagarajan and S R Savithri, "Evaluation of computer aided 3D lip sync instruction model using virtual reality objects", Proceeding of 6th International conference on Disability, Virtual Reality & Association Technology, 2006, pp 67-73.

[2] "Mac OS Human Interface Guidelines user Experience", Apple Inc. 2011, pp 27-28.

[3] Preben Wik, Rebecca Hincks, Julia Hirschberg, "Responses to Ville: A virtual language teacher for Swedish", Proceedings of SLaTE Workshop on speech and Language Technology in Education, 2009.

[4] Jont B.Allen "Articulation and Intelligibility", Morgan and Claypool Publishers, ISBN: 1598290088, 2005, Page No 98-106.

[5] A Rathinavelu and G Yuvaraj, "Data Visualization Model for Speech Articulators", Proceedings of AICERA 2011, pp.155-159.

[6] Giampiero Salvi, Jonas Beskow, Samer Al Moubayed, and Bjorn Granstrom, “SynFace-Speech-Driven Facial Animation for Virtual Speech-Reading Support”, EURASIP Journal on Audio, Speech, Music Processing, Volume 2009, 2009, DOI:10.1155/2009/191940.

[7] J. Beskow, “Rule-based visual speech synthesis,” in Proceedings of the European Conference on Speech Communication and Technology (Eurospeech ’95), 1995, pp. 299–302.

[8] Olov Engwall, preben Wik, “Can you tell if the tongue movements are real or synthesized”, International Conference on Auditory-Visual Speech Processing, Sep 2009.

[9] O. Engwall, “Combining MRI, EMA & EPG in a three dimensional tongue model”, Speech Communication, Vol. 41/2-3, 2003, pp. 303–329.

[10] Tina Absullah, Noor Azma Abu Bakar, “A study of second language learners’ perception of using short story in learning English”, Journal Article, unpublished.

[11] http://www.meyersound.com/support/papers/speech/ section3.htm, Accessed on 10/28/2011.

[12] B. Hagerman and C. Kinnefors, “Efficient adaptive methods for measuring speech reception threshold in quiet and in noise”, Scandinavian Audiology, Vol. 24, No. 1, 1994, pp. 71–77.

◈◈◈

Frequent Pattern Analysis For Effective Classification Using Parallel Apriori

N.S. Jagadeesh & E.Madhusudhana Reddy

Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science,
Madanapalle, Andhra Pradesh, India.
E-mail : jagadeeshns4u@gmail.com, e_mreddy@yahoo.com

Abstract - With the extensive competition in the domestic and international business, the Customer Relationship Management has become one of matters of concern to the enterprise. CRM takes the customers as the center; it gives a new life to the enterprise organization system and optimizes the business process. In an effort to help enterprises understand their customers' shopping behavior and the ways to retain valued customers, we propose data mining techniques. As a rising subject, data mining is playing an increasingly important role in the decision support activity of every walk of life. CRM can be defined as the process of predicting customer behavior and selecting actions to influence that behavior to benefit the company. It gives a new life to the enterprise organization system and optimizes the business process. Customer satisfaction can also be improved through more effective marketing. Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data Mining Approach is used which is a sophisticated statistical processing or artificial intelligence algorithms to discover useful trends and patterns from the extracted data so that it can yield important insights including prediction models and associations that can help companies understand their customer better.

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. The Apriori algorithm is the most established algorithm for frequent item sets mining (FIM). Several implementations of the Apriori algorithm have been reported and evaluated. we revised Apriori implementation into a parallel one where input transactions are read by a parallel computer. The effect a parallel computer on this modified implementation is presented.

Keywords - *Apriori, Association Rules, Data Mining, Frequent Itemsets Mining (FIM), Parallel Computing, CRM (Customer Relationship Management),Data Warehouse, Clustering.*

I. INTRODUCTION

Enterprise data mining applications, such as mining public service data and telecom fraudulent activities, inevitably involve complex data sources, particularly multiple large scale, distributed, and heterogeneous data sources embedding information about business transactions, user preferences, and business impact. In these situations, business people certainly expect the discovered knowledge to present a full picture of business settings rather than one view based on a single source. Knowledge reflecting full business settings is more business friendly, comprehensive, and informative for business decision makers to accept the results and to take operable actions accordingly.

It is challenging to mine for comprehensive and informative knowledge in such complex data suited to real-life decision needs by using the existing methods. The challenges come from many aspects, for instance, the traditional methods usually discover homogeneous features from a single source of data while it is not effective to mine for patterns combining components from multiple data sources. It is often very costly and sometimes impossible to join multiple data sources into a single data set for pattern mining.

In our project, we proposed the concepts of *combined association rules*, *combined rule pairs*, and *combined rule clusters* to mine for informative patterns in complex data by catering for the comprehensive aspects in multiple datasets. A combined association rule is composed of multiple heterogeneous itemsets from different data sets while combined rule pairs and combined rule clusters are built from combined association rules. Analysis shows that such combined rules cannot be directly produced by traditional algorithms such as the FPGrowth. This paper builds on the existing works and proposes the approach of *combined mining* as a general method for directly identifying patterns enclosing constituents from multiple sources or with heterogeneous features covering demographics, behavior, and business impacts. Its deliverables are *combined patterns* such as the aforementioned combined association rules.

The general ideas of combined mining are as follows.

1) By involving multiple heterogeneous features, combined patterns are generated which reflect multiple aspects of concerns and characteristics in businesses.

2) By mining multiple data sources, combined patterns are generated which reflect multiple aspects of nature across the business lines.

3) By applying multiple methods in pattern mining, combined patterns are generated which disclose a deep and comprehensive essence of data by taking advantage of different methods.

4) By applying multiple interestingness metrics in pattern mining, patterns are generated which reflect concerns and significance from multiple perspectives.

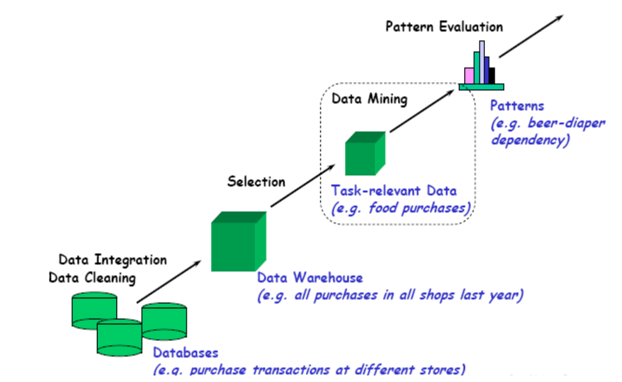
The main contributions of our paper are as follows:

1) Building on existing works, generalizing the concept of combined mining that can be expanded and instantiated into many specific approaches and models for mining complex data toward more informative knowledge.

2) Discussing general frameworks and their paradigms and basic processes of *multi feature* and *multi method combined mining* for supporting combined mining, which contribute to *multisource combined mining*—they are flexible to be instantiated into specific needs.

3) Proposing various strategies for conducting pattern interaction when instantiating the aforementioned proposed frameworks—as a result, novel combined pattern types, such as incremental cluster patterns, can result from combined mining, which have not been investigated before.

4) Illustrating the corresponding interestingness metrics for evaluating certain types of combined patterns.



Architecture

II. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods. In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

The wide-spread use of distributed information systems leads to the construction of large data collections in business, science and on the Web. These data collections contain a wealth of information, that however needs to be discovered. Businesses can learn from their transaction data more about the behavior of their customers and therefore can improve their business by exploiting this knowledge. Science can obtain from observational data (e.g. satellite data) new insights on research questions. Data mining provides methods that allow to extract from large data collections unknown relationships among the data items that are useful for decision making. Thus data mining generates novel, unsuspected interpretations of data.

General Definitions

- **Itemset:** Set of items that occur together
- **Association Rule:** Probability that particular items are purchased together.
  - ○ $X \rightarrow Y$ where $X \cap Y = 0$

- **Support**, supp(X) of an itemset X is the ratio of transactions in which an itemset appears to the total number of transactions.
- **Share** of an itemset is the ratio of the count of items purchased together to the total count of items purchased in all transactions.
- **Confidence** of rule $X \to Y$, denoted conf(X∟$\to Y$) = supp($X \cup Y$) / supp(X).
    - Confidence can also be defined in terms of the conditional probability.

      conf($X \to Y$) = P($Y \mid X$) = P($X \cap Y$) / P(X).
- **Transaction Database** stores transaction data. Transaction data may also be stored in some other form than a m x n database.

Applications of relational data mining

The use of RDM has enabled applications in areas rich with structured data and domain knowledge, which would be difficult to address with single table approaches. RDM has been used in different areas, ranging from analysis of business data, through environmental and traffic engineering to web mining, but has been especially successful in bioinformatics (including drug design and functional genomics).Bioinformatics applications of RDM are discussed in the article by Page and Craven in this issue.

daughter (X,Y) ← female(X),mother(Y,X).

daughter (X,Y) ← female(X),father(Y,X).

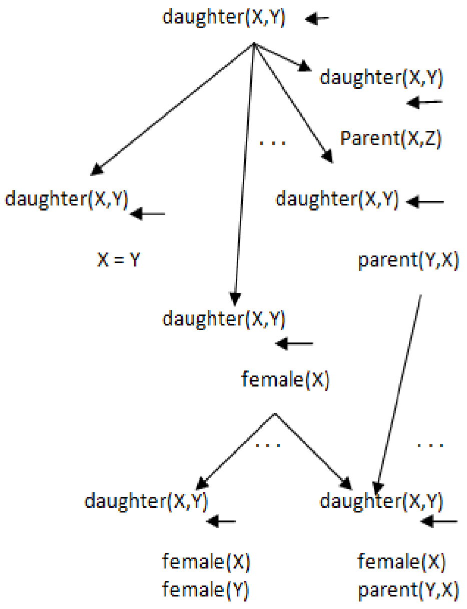

daughter(X,Y) ←
daughter(X,Y) ← Parent(X,Z)
. . .
daughter(X,Y) ← X = Y
daughter(X,Y) ← parent(Y,X)
daughter(X,Y) ← female(X)
. . .
. . .
daughter(X,Y) ← female(X) female(Y)
daughter(X,Y) ← female(X) parent(Y,X)

Figure 1: Part of the refinement graph for the family relations problem.

Relational Association Rules

The discovery of frequent patterns and association rules is one of the most commonly studied tasks in data mining. Here we first describe frequent relational patterns and relational association rules.

Association Rules

The general aim of data mining is to find patterns in data that can become actionable information. In the case of association discovery, the involvement of domain experts is critical to the process of identification of relevant rules. One of the reasons behind maintaining any database is to enable the user to find interesting patterns and trends in the data. Once this information is available, we can perhaps get rid of the original database. The output of the data-mining process should be a "summary" of the database. One such type constitutes the association rule.

Table joining is widely used in order to mine patterns from multiple relational tables by putting relevant features from individual tables into a consolidated one.

TABLE I

CUSTOMER DEMOGRAPHIC DATA (F-FEMALE, M-MALE)

Customer ID	Gender	. . .
1	F	
4	M	

TABLE II

Customer ID	Policies	Activities	Debt
1	(c1,c2)	(a1,a2)	Y
4	(c2,c4)	(a1-a2-a3)	N
4	(c1,c2,c4)	(a1-a3)	N
4	(c1,c2,c3)	(a1-a3-a4)	Y
4	(c2,c3)	(a2-a4)	N

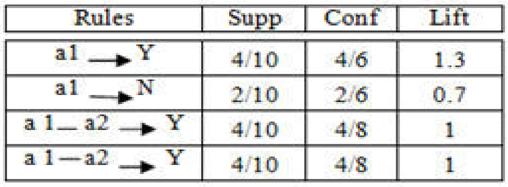
TABLE III

TRADIATIONAL ASSOCIATION RULES

Rules	Supp	Conf	Lift
c1 → Y	4/10	4/6	1.3
c1 → N	2/10	2/6	0.7
c2 → Y	4/10	4/8	1
c2 → N	4/10	4/8	1

TABLE IV

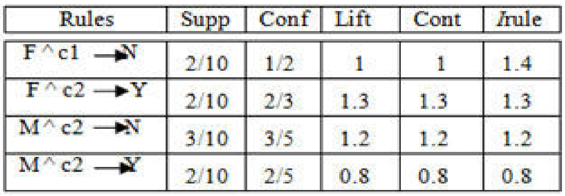
TRADIATIONAL ASSOCIATION RULES



Rules	Supp	Conf	Lift
a1 → Y	4/10	4/6	1.3
a1 → N	2/10	2/6	0.7
a 1_ a2 → Y	4/10	4/8	1
a 1—a2 → Y	4/10	4/8	1

TABLE V

COMBINED ASSOCIATION RULES



Rules	Supp	Conf	Lift	Cont	Irule
F ^ c1 → N	2/10	1/2	1	1	1.4
F ^ c2 → Y	2/10	2/3	1.3	1.3	1.3
M ^ c2 → N	3/10	3/5	1.2	1.2	1.2
M ^ c2 → Y	2/10	2/5	0.8	0.8	0.8

As a result, a pattern may consist of features from multiple tables. This method is suitable for mining multiple relational databases, particularly for small data sets.

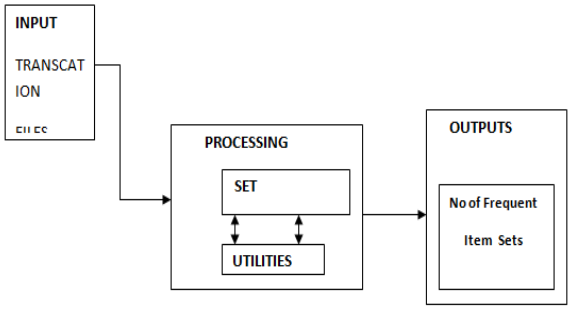
The Apriori Algorithm

The Apriori algorithm finds frequent itemsets according to a user-defined minimum support. In the first pass of the algorithm, it constructs the candidate 1-itemsets. The algorithm then generates the frequent 1-itemsets by pruning some candidate 1-itemsets if their support values are lower than the minimum support.

Implementation for parallel computing

The Apriori algorithm has been revised in several ways. One revision of the Apriori algorithm is to partition a transaction database into disjoint partitions TDB1, TDB2, TDB3, …, TDBn. Node 0 then computes the sum of all candidate k-itemsets and prunes the candidate kitemsets to the frequent k-itemsets.

Flow Diagram



The Apriori Algorithm : Pseudo code

```
$C_k$: Candidate itemset of size k
$L_k$ : frequent itemset of size k
$L_1$ = {frequent items};
for (k = 1; $L_k$ !=$\varnothing$; k++) do begin
    $C_{k+1}$ = candidates generated from $L_k$;
 for each transaction t in database do
      increment the count of all candidates in $C_{k+1}$   that are contained in t
$L_{k+1}$=candidates in $C_{k+1}$ with min_support  end
return $\cup_k L_k$;
```

III. CONCLUSION AND FUTURE ENHANCEMENT

Typical enterprise applications, such as telecom fraud detection and cross-market surveillance in stock markets, often involve multiple distributed and heterogeneous features as well as data sources with large quantities and expect to cater for user demographics, preferences, behavior, business appearance, service usage, and business impact. This challenges existing data mining methods such as post analysis and table joining based analysis. Building on existing works, this paper has presented a comprehensive and general approach named combined mining for discovering informative knowledge in complex data. The frameworks are extracted from our relevant business projects conducted and currently under investigation from the domains of government service, banking, insurance, and capital markets. Several real-life cases studies have been briefed which instantiate some of the proposed frameworks in identifying combined patterns in multiple sources of governmental service data. They have shown that the proposed frameworks are flexible and customizable for handling a large amount of complex data involving multiple features, sources, and methods as needed, for which data sampling and table joining may not be acceptable.

REFERENCES

[1] Combined Mining: Discovering Informative Knowledge in Complex Data Longbing Cao, Senior Member, IEEE, Huaifeng Zhang, Member, IEEE, Yanchang Zhao, Member, IEEE,Dan Luo, and Chengqi Zhang, Senior Member, IEEE Transactions on systems,man,and cybernetics-Part B: Cybernetics, vol. 41, no.3, june.2011

[2] L. Cao, Y. Zhao, and C. Zhang, “Mining impact-targeted activity patterns in imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 20, no. 8, pp. 1053–1066, Aug. 2008.

[3] L. Cao, Y. Zhao, H. Zhang, D. Luo, and C.Zhang, “Flexible frameworks for actionable knowledge discovery,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 9, pp. 1299–1312, Sep. 2010.

[4] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, “Discriminative frequent pattern analysis for effective classification,” in Proc. ICDE, 2007,pp. 716–725.

[5] S. Dzeroski, “Multirelational data mining: An introduction,” ACM SIGKDD Explor. Newslett., vol. 5, no. 1, pp. 1–16, Jul. 2003.

[6] G. Dong and J. Li, “Efficient mining of emerging patterns: Discovering trends and differences,” in Proc. KDD, 1999, pp. 43–52.

[7] K. K. R. Hewawasam, K. Premaratne, and M.-L. Shyu, “Rule mining and classification in a situation assessment application: A belief-theoretic approach for handling data imperfections,” IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 37, no. 6, pp. 1446–1459, Dec. 2007.

[8] A. Jorge, “Hierarchical clustering for thematic browsing and summarization of large sets of association rules,” in Proc. SDM, 2004, pp. 178–187.

[9] B. Lent, A. N. Swami, and J. Widom, “Clustering association rules,” in Proc. ICDE, 1997, pp. 220–231.

[10] B. Liu, W. Hsu, and Y. Ma, “Pruning and summarizing the discovered associations,” in Proc. KDD, 1999, pp. 125–134.

[11] Y. Zhao, C. Zhang, and L. Cao, Eds., Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction. Hershey, PA: Inf. Sci.Ref., 2009.

[12] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc. of SIGMOD, pages 207–216, 1993.

[13] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. of VLDB, pages 487–499, 1994.

[14] X. Yin and J. Han. Cpar: Classification based on predictive association rules. In Proc. of SDM, 2003

[15] [Online].Available:http://datamining.it.uts.edu.au/group/projects.php

◈◈◈

Secured Data Tranamission using Snort Rules and Mining Techniques

R.Venkatramana & M.Sreedevi

Department of CSE, Madanapalli Institute of Technology and Science,
Madanapalli, Andhra Pradesh, India
E-mail : rvramana.r@gmail.com

***Abstract* -** Network traffic analysis becomes more and more crucial in the IP network infrastructure as the amount of IP packets transmitted on the Internet at any given moment of time increases enormously. A thorough understanding of the IP traffic will help us better design our network topology and utilize bandwidth more effectively. From the perspective of security, it can also protect our system from attacks, such as intrusions, our model employs feature selection so that the binary classifier for each type of attack can be more accurate, which improves the detection of attacks that occur less frequently in the training data. Based on the accurate binary classifiers, our model applies a new ensemble approach which aggregates each binary classifier's decisions for the same input and decides which class is most suitable for a given input. During this process, the potential bias of certain binary classifier could be alleviated by other binary classifiers' decision. Our model also makes use of multi boosting for reducing both variance and bias. The clients have some rules to communicate between them using snort rules. Any Communications (such as FTP, SMTP, etc) between the clients are monitored by the snort. If it continues again, then that particular client will be disconnected from this network (means cannot be able to communicate with other clients in that network.) But, that client will be physically connected with the network. The proposed work describes a network traffic analysis software tool, which provides searching, visualization, and preprocessing functions with a user-friendly GUI implemented in Java language. Within the huge network traffic data collected, a user can identify any particular packets using various searching functions provided. Visualization presents the analyzed result in a different setting to further enhance the analysis. The GUI in Java allows the tool to be used in different platforms. This tool is tested and demonstrated through several real network datasets.

***Keywords* -** *Algorithms, filtering algorithms, finite-state automata (FSA), mathematics, packet filters, packet processing, predicate optimization, protocol description languages (PDLs), run-time safety, snort rules and mining techniques.*

I. INTRODUCTION

Packet filters are a class of packet manipulation programs used to classify network traffic in accordance to a set of user-provided rules; they are a basic component of many networking applications such as shapers, sniffers, demultiplexers, firewalls, and more.The modern networking scenario imposes many requirements on packet filters, mainly in terms of processing speed (to keep up with network line rates) and resource consumption (to run in constrained environments). Filtering techniques should also support modern protocol formats that often include cyclic or repeated structures (e.g., MPLS label stacks, IPv6 extension headers). Finally, it is also crucial that filters preserve the integrity of their execution environment, both in terms of memory access safety and termination enforcement, especially when running as an operating system module or on the bare hardware. Although at first sight this aspect might not seem crucial, it is a fact that many of the limitations built into existing packet filters derive directly from safety issues. As an example, the impossibility of automatically proving termination for a generic computer program led the BPF [1] designers to generate acyclic filters only, thus preventing the parsing of packets with multiple levels of encapsulation or repeated field sequences.

Existing packet filters focus invariably on subsets of these issues but, to the best of our knowledge, do not solve all of them at the same time. As an example, two widely known generators, BPF [2] and PathFinder [3], do not support recursive encapsulation; NetVM-based filters [4], on the other hand, have no provision for enforcing termination, either in filtering code or in the underlying virtual machine.

This paper presents Stateless PAcket Filter (SPAF), a finite-state automata (FSA)-based technique to generate fast and safe packet filters that are also flexible enough to fully support most layer-2 to layer-4 protocols, including optional and variable headers and recursive encapsulation. The proposed technique specifically targets the lower layers of the protocol stack and does not directly apply for deep packet inspection nor for stateful filtering in general. Moreover, for the purpose of this paper, we consider only static situations where on-the-fly rule set updates are not required. While these limitations exclude some interesting use cases, SPAF filters are nevertheless useful for a large class of applications, such as monitoring and traffic trace filtering, and can serve as the initial stage for more complex tools such as intrusion detection systems and firewalls.

A stateless packet filter can be expressed as a set of predicates on packet fields, joined by boolean operators; often these predicates are not completely independent from one another, and the evaluation of the whole set can be short-circuited. One of the most important questions in designing generators for high-performance filters is therefore how to efficiently organize the predicate set to reduce the amount of processing required to come to a match/mismatch decision. By considering packet filtering as a regular language recognition problem and exploiting the related mathematical framework to express and organize predicates as finite-state automata, SPAF achieves by construction a reduction of the amount of redundancy along any execution path in the resulting program: Any packet field is examined at most once. This property emerges from the model, and it always holds even in cases that are hard to treat with conventional techniques, such as large-scale boolean composition. Moreover, thanks to their simple and regular structure, finite automata also double as an internal representation directly translatable into an optimized executable form without requiring a full-blown compiler.

Finally, safety (both in terms of termination and memory access integrity) can be enforced with very low run-time overhead.

The rest of this paper is structured as follows. Section II presents an overview of the main related filtering approaches developed to this date. Section III provides a brief introduction to the FSAs used for filter representation and describes the filter construction procedure. Section IV focuses on executable code generation and on enforcing the formal properties of interest, Finally, Section V reports conclusions and also highlights possible future developments.

II. RELATED WORK

Given their wide adoption and relatively long history, there is a large corpus of literature on packet filters. A first class of filters is based on the CFG paradigm; the best-known and most widely employed one is probably BPF [1], the Berkeley Packet Filter. BPF filters are created from protocol escriptions hardcoded in the generator and are translated into a bytecode listing for a simple, *ad hoc* virtual machine. The bytecode was originally interpreted, leading to a considerable run-time overhead impact that can be reduced by employing JIT techniques [5]. BPF disallows backward jumps in filters in order to ensure termination, thus forgoing support for, e.g., IPv6 extension headers; memory protection is enforced by checking each access at run-time. Multiple filter statements can be composed together by boolean operators, but in the original BPF implementation, only a small number of optimizations are performed over predicates, leading to run-time inefficiencies when dependent or repeated predicates are evaluated. Two relevant BPF extensions are BPF and xPF. BPF [2] adds local and global data-flow optimization algorithms that try to remove redundant operations by altering the CFG structure. xPF [6] relaxes control flow restrictions by allowing backward jumps in the filter CFG; termination is enforced by limiting the maximum number of executed instructions through a run-time watchdog built into the interpreter, but its overhead was not measured, and extending this approach to just-in-time code emission has not been proposed and might prove difficult.

A further CFG-based approach, unrelated to BPF, is described in [4]. Its main contribution is decoupling the protocol database from the filter generator by employing an XML-based protocol description language, NetPDL [7]. Filtering code is executed on the NetVM [8], a special-purpose virtual machine targeting network applications that also provides an optimizing JIT compiler that works both on filter structure and low-level code. The introduction of a high-level description language reportedly does not cause any performance penalties; this approach, however, delegates all safety considerations to the VM and does not provide an effective way to compose multiple filters.

In general, CFG-based generators benefit from their flexible structure that does not impose any significant restriction on predicate evaluation order; for the same reason, however, they are prone to the introduction of hard-to-detect redundancies, leading to multiple unnecessary evaluations if no further precautions are taken. Even when optimizers are employed and are experimentally shown to be useful, they work on an opportunistic basis and seldom provide any hard guarantees on the resulting code.

A second group of filter generators chooses tree-like structures to organize predicates. PathFinder [3] transforms predicates into template masks (atoms), ordered into decision trees. Atoms are then matched through a linear packet scan until a result is reached. Decision trees enable an optimization based on merging prefixes that are shared across multiple filters. PathFinder is shown to work well both in software and hardware implementations, but it does not take protocol database decoupling into consideration, and no solution to memory safety issues is proposed for the software implementation. FSA-based filters share a degree of similarity with PathFinder as packets are also scanned linearly from the beginning to the end, but predicate organization, filter composition, and safety considerations are handled differently. DPF [9] improves over PathFinder by generating machine code just-in-time and adding low-level optimizations such as a flexible switch emission strategy. Moreover, DPF is capable of aggregating bounds checks at the atom level by checking the availability of the highest memory offset to be read instead of considering each memory access in isolation; our technique, described in Section IV-E, acts similarly but considers the filter as a whole, thus further reducing run-time overhead.

While organizing predicates into regular structures makes it easier to spot redundancies and other sources of overhead, it also introduces different limitations. As an example, generators restricted to the aforementioned acyclic structures do not fully support tunneling or repeated protocol portions. Moreover, it has been noted that performing prefix coalescing is not sufficient to catch certain common patterns, resulting in redundant predicate evaluation [2].

A third approach is to consider packet filtering as a language recognition problem. Jayaram *et al.* [10] use a pushdown automaton to perform packet demultiplexing; filters are expressed as LALR(1) grammars and can therefore be effectively composed using the appropriate rules. This solution improves filter scalability, but there are downsides related to the push-down automaton: A number of specific optimizations are required to achieve good performance. It is also quite unwieldy to express protocols and filter rules as formal grammars that must be kept strictly unambiguous: The authors marginally note that the simpler FSA model would be sufficient for the same task.

Apart from the specialized solutions for fast packet filtering mentioned, one of the most widely used packet filtering programs is the NetFilter framework.1 NetFilter is a component of the Linux kernel that performs packet filtering, firewalling, mangling operations (e.g., network address translation), and more, acting through a set of hooks and callbacks that intercept packets as they traverse the networking stack. In contrast with all the aforementioned approaches, NetFilter uses the relatively simple method of applying all the specified rules in sequence when performing packet filtering, leading to poor performance and scalability; moreover, it appears not possible to specify an arbitrary predicate, filters being limited to preset protocols and statements that are specialized by specifying actual network addresses and ports.

Besides the generation technique, there have also been improvements along other dimensions such as architectural considerations, as demonstrated by xPF, FFPF [19], and nCap [20], or dynamic rule sets support, as shown by the SWIFT tool [21]. We consider these aspects out of scope for the purpose of this paper, being either orthogonal to the technique we present or the object of future works.

Definition of Near-Duplicate

The central idea of near-duplicate spam detection is to exploit reported known spams to block subsequent ones which have similar content. For different forms of e-mail representation, the definitions of similarity between two e-mails are diverse. Unlike most prior works representing e-mails based mainly on content text, we investigate representing each e-mail using an HTML tag sequence, which depicts the layout structure of e-mail, and look forward to more effectively capturing the near-duplicate phenomenon of spams.

Let I ¼ ft1; t2; . . . ; ti; . . . ; tn;

<mytext=>;

<anchor>g be the set of all valid HTML tags

with two types of newly created tags,

<mytext=> and

<anchor>, included. An e-mail abstraction derived from

procedure SAG is denoted as <e1; e2; . . . ; ei; . . . ; em>, which

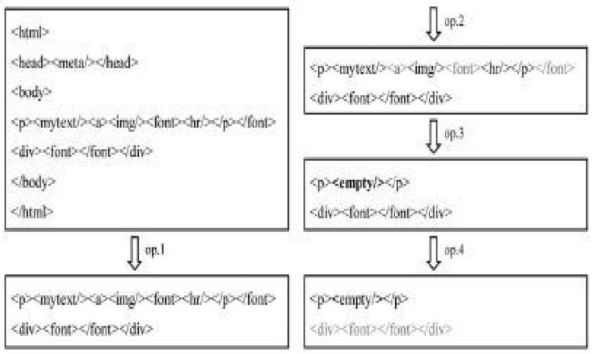
is an ordered list of tags, where ei 2 I. The definition of near duplicate

is: “Two e-mail abstractions _ ¼ <a1; a2; . . . ;

ai; . . . ; an> and _ ¼ <b1; b2; . . . ; bi; . . . ; bm> are viewed as

near-duplicate if 8ai ¼ bi and n ¼ m.”

The tag length of an e-mail abstraction is defined as the number of tags in an e-mail abstraction.

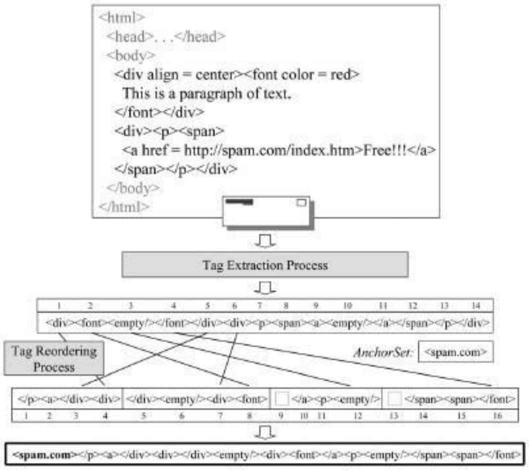


The following sequence of operations is performed in the preprocessing step.

1. Front and rear tags are excluded.
2. Nonempty tags2 that have no corresponding start tags or end tags are deleted. Besides, mismatched nonempty tags are also deleted.
3. All empty tags2 are regarded as the same and are replaced by the newly created <empty=> tag.

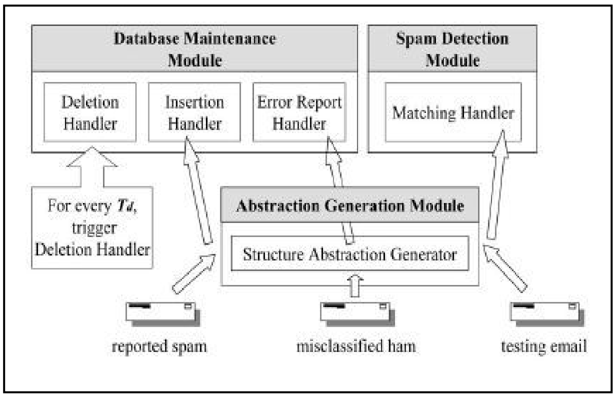
   Moreover, successive <empty=> tags are pruned and only one <empty=> tag is retained.
4. The pairs of nonempty tags enclosing nothing are removed.

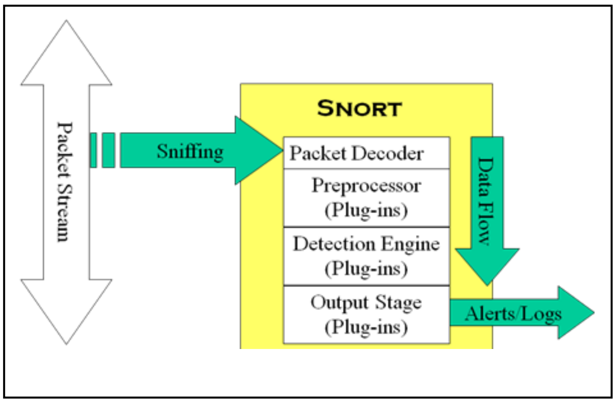
Example for Mail


<html>
<head>. . .</head>
<body>
<div align = center><font color = red>
This is a paragraph of text.
</font></div>
<div><p><span>
<a href = http://spam.com/index.htm>Free!!!</a>
</span></p></div>
</body>
</html>
Tag Extraction Process
<div><font><empty/></font></div><div><p><span><a><empty/></a></span></p></div>
Tag Reordering Process
AnchorSet: <spam.com>
</p><a></div><div> </div><empty/><div><font> </a><p><empty/> </span><span></font>
<spam.com></p><a></div><div></div><empty/><div><font></a><p><empty/></span><span></font>

SYSTEM ARCHITECTURE

In the packet filtering process with the increasing popularity of electronic mail (or e-mail), several people and companies found it an easy way to distribute a massive amount of unsolicited messages to a tremendous number of users at a very low cost. These unwanted bulk messages or junk emails are called spam messages. The majority of spam messages that has been reported recently are unsolicited commercials promoting services and products including sexual enhancers, cheap drugs and herbal supplements, health insurance, travel tickets, hotel reservations, and software products. They can also include offensive content such pornographic images and can be used as well for spreading rumors and other fraudulent advertisements such as make money fast.

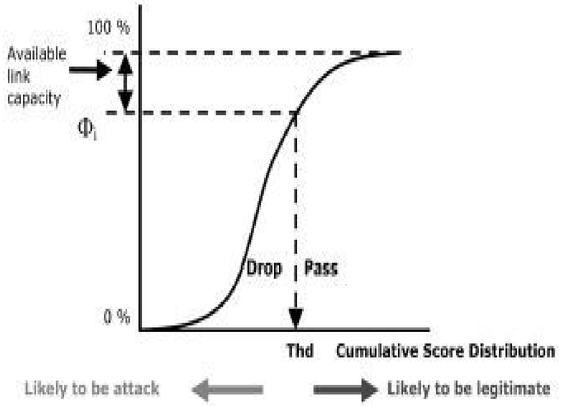




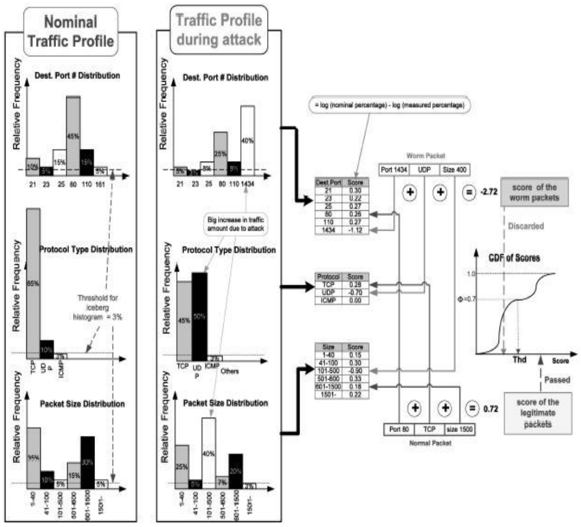
As a result, spam has become an area of growing concern attracting the attention of many security researchers and practitioners. In addition to regulations and legislations, various anti-spam technical solutions have been proposed and deployed to combat this problem. Front-end filtering was the most common and easier way to reject or quarantine spam messages as early as possible at the receiving server. However most of the early anti-spam tools were static; for example using a blacklist of known spammers, a white list of good sources, or a fixed set of keywords to identify spam messages. Although these list-based methods can substantially reduce the risk provided that lists are updated periodically, they fail to scale and to adapt to spammers' tactics.

Selective Packet Discarding

Once the score is computed for a packet, selective packet discarding, and overload control can be performed using the score as the differentiating metric. Since an exact prioritization would require offline, multiple-pass operations, e.g., sorting and packet buffering, the following alternative approach is taken into account. First, the cumulative distribution function (CDF) of the scores of all incoming packets in time period (Ti) is maintained. Second, the cut-off threshold score is calculated. Third, the arriving packets in time period T (i+1) if its score value is below the cut-off threshold are discarded. At the same time, the packets arriving at T (i+1) create a new CDF.



Selective packet discarding


Nominal Traffic Profile
Traffic Profile during attack
Relative Frequency
Dest. Port # Distribution
Protocol Type Distribution
Packet Size Distribution
Threshold for iceberg histogram = 3%
Big increase in traffic amount due to attack
= log (nominal percentage) - log (measured percentage)
Worm Packet
Port 1434
UDP
Size 400
= -2.72
score of the worm packets
Discarded
CDF of Scores
Φ=0.7
Thd
Score
Passed
= 0.72
score of the legitimate packets
Port 80
TCP
size 1500
Normal Packet
Dest Port
Protocol
Size
TCP
UDP
ICMP
Others

Discarding SQL Slammer Worm attack packets

III. FILTER GENERATION TECHNIQUE

In this section we describe about filtering techniques using snort rules and mining techniques.

One of the key concepts in PacketScore is the notion of "Conditional Legitimate Probability" (CLP) based on Bayesian theorem. CLP indicates the likelihood of a packet being legitimate by comparing its attribute values with the values in the baseline profile. Packets are selectively discarded by comparing the CLP of each packet with a dynamic threshold. The concept of using a baseline profile with Bayesian theorem has been used previously in anomaly-based IDS (Intrusion Detection System) applications, where the goals are generally attack detection rather than real-time packet filtering.

In this research, the basic concept to a practical real-time packet filtering scheme using elaborate processes is extended. In this method, the PacketScore operations for single-point protection is described, but the fundamental concept can be extended to a distributed implementation for core-routers.

To make it more suitable for real-time processing, conversion of floating-point division/multiplication operations into subtraction/addition operations is made. Scoring a packet is equivalent to looking up the scorebooks, e.g., the TTL scorebook, the packet size scorebook, the protocol type scorebook, etc. After looking up the multiple scorebooks, the matching CLP entries in a log-version scorebook are added. This is generally faster than multiplying the matching entries in a regular scorebook. The small speed improvement from converting a multiplication operation into an addition operation is particularly useful because every single packet must be scored in real-time. This speed improvement becomes more beneficial as the number of scorebooks increases. On the other hand, generating a log-version scorebook may take longer than a regular scorebook generation. However, the scorebook is generated only once at the end of each period and it is not necessary to observe every packet for scorebook generation; thus, some processing delay can be allowed. Furthermore, scorebook generation can be easily parallelized using two processing lines, which allows complete sampling without missing a packet.

The purpose of a stateless packet filter generator is to create a program that, given a finite-length byte sequence (a packet) as its input, returns a binary match/mismatch decision. The input of the generator itself consists of a set of filter rules provided by the user that specify the desired properties of matching packets; each rule, in turn, consists of multiple predicates expressed in a simple high-level language (where header fields and protocols appear symbolically), combined together with boolean operators. In older generators, the set of supported protocols was fixed; in modern ones protocol header formats are kept into an external database that can be updated without modifying the generator.

In order to develop a successful FSA-based filtering technique, it is first of all necessary to show that any filter of interest can be expressed as a finite automaton, then provide a method to transform a high-level filter statement and a protocol database into FSA form.

Finally, the resulting automaton must be translated into an efficiently executable form.

A. Protocol Database Compilation

The first phase in the SPAF generation process consists of parsing the protocol database and building template automata that recognize all the correctly formatted headers for a given protocol. These automata will be reused and specialized in later phases to create the final filter.

In order to decouple filter generation from the protocol database, we have employed an XML-based protocol description language (NetPDL [7]) designed to describe the on-the-wire structures of network protocols and their encapsulation relationships. NetPDL descriptions are stored in external files that can be freely edited without modifying the generator itself.

A precise description of NetPDL is beyond the scope of this paper. Nevertheless, we shall provide a quick overview of the features supported by the FSA generator. The language provides a large number of primitives that enable the description of header formats of layer-2–7 protocols, but for the scope of this work we have restricted our support to those designed for layer-2–4 decoding. The basic building block of a protocol format is the header field, a sequence of bytes or bits that can be either fixed or variable in size. Adjacent fields are by default laid out in sequence, but more complex structures such as optional or repeated sections can be created using conditional choices and loops; these statements are controlled by expressions that can contain references to the values of previously encountered fields.

A second NetPDL portion contains a sequence of control flow operations (if, switch) that predicate encapsulation relationships. In general, the control flow is followed until a nextproto tag is encountered, specifying which is the next protocol to be found in the packet. A NetPDL database thus describes an oriented encapsulation graph where the vertices are protocols and the edges are encapsulation relationships. Currently, the graph begins with a single user-specified root that usually represents the link-layer protocol, but an extension to multiple ones would be trivial. Starting from this root, the FSA generator follows the encapsulation graph and builds a FSA for every reachable protocol using the method explained later in this section.

```
<protocol  name =”ipv6”>
   <format>
      <field>
<field type=”bit” name = ”ver” mask=”0xF0000000” size =”4”/>
<field type=”bit” name = ”tos” mask=”0x0F000000”size =”4”/>
<field type=”bit” name = ”flabel” mask=”0x00FFFFFF”size =”4”/>
<field type=”fixed” name=”plex”  size =”2”/>
<field type=”nexthdr” name=”plex”  size =”1”/>
<field type=”hop” name=”plex”  size =”16”/>
<field type=”src” name=”plex”  size =”16”/>
<field type=”dst” name=”plex”  size =”16”/>
<loop type =”while” expr=”1”>
<switch expr=”nexthdr”>
<case value=”0”><includeblk name=”HBH”/></case>
<case value=”0”><includeblk name=”AH”/></case>
….
<default>
<loopctrl type=”break”/>
</default>
   </switch>
      </loop>
         </fields>
            </format>
<encapsulation>
<switch expr=”nexthdr”>
<case value=”4”> <nextproto proto=”#ip”/></case>
<case value=”4”> <nextproto proto=”#tcp”/></case>
<case value=”4”> <nextproto proto=”#udp”/></case>
….
</switch>
   </encapsulation>
      </protocol>
```

IPV6 NetPDL excerpt

As an example, a simplified NetPDL description of the IPv6 header format is presented in Fig. 1. IPv6 starts with a sequence of fixed-size fields; bitfields (such as ver) are specified by the mask attribute. The initial portion is followed by a set of extension headers, each one containing a “next header” information (nexthdr). This sequence is of unspecified (but implicitly finite, as any packet is finite) length, and it is described using a switch nested within a loop: At each iteration, the newly read nexthdr field is evaluated, and if no more extension headers are present, the loop terminates. Encapsulation relationships are also specified in a similar fashion by jumping to the correct protocol depending on the value of the last nexthdr encountered.

SPAF currently supports the full versions of the most common layer-2–4 protocols in use nowadays, such as Ethernet, MPLS, VLAN, PPPoE, ARP, IPv4, IPv6, TCP, UDP, and ICMP; this set can be easily extended as long as no stateful capabilities are required.

An important point regarding FSA creation from NetPDL descriptions is that, as long as it is correctly performed, it is not be a critical task for filter performance: Any resulting automaton ultimately will be determinized and minimized, yielding a canonical representation of the filter that does not depend on the generation procedure. For this reason, and given the complexity involved, the NetPDL-to-FSA conversion procedure is not fully described in this paper, and it can be regarded as an implementation detail. Nevertheless, in order to exemplify how the conversion can be done, we report the key steps for translating the NetPDL snippets of Fig. 2 into the corresponding automata.

The purpose of this initial conversion step is not to generate automata immediately suitable for filtering. On the contrary, the results are templates for the following generation steps, representing the “vanilla” version of protocol headers, with no other conditions imposed, to be specialized according to the filter rules. Since they are strictly related to header format, any inputconsuming transition in these templates can be related to a specific portion of one3 header field; this information must be preserved to accommodate the imposition of filtering rules. For this reason, template automata are augmented by marking all the relevant transitions with the related field’s name.4

The simplest example is generating an automaton that parses a fixed-length header field [Fig. 2(a)]: It is sufficient to build a FSA that skips an appropriate amount of bytes, resulting in Fig. 2(b). During the construction process, header fields are given well-defined start and end5 states that are used as stitching points to join with any predecessors or successors by -transitions, as required. A more complex example involving a conditional choice is shown in Fig. 2(c). The generation procedure starts by creating automata representations for all the initial fields in the NetPDL description; upon encountering the switch construct, however, the generator backtracks the transition graph until it encounters the type field. Once found, all the states/transitions that follow type (the block in the figure) are replicated. The original copy is left as is, while in the replica the transitions for type are specialized to recognize the bytes of interest for the switch, so the right path will be taken depending on the actual input values. Finally, the correct trailing block (or) is joined in the right place. The last example [Fig. 2(e) and (f)] shows the automata generated for a header structure similar to the IPv6 extension headers case. In this case, a loop is interlocked with a switch construct, and a greater amount of block replication is required to ensure that independent paths exist into the automaton for every possible combination of the current nexth value (upon which the outcome of the switch depends) and the next nexth value, which might cause the loop to end.

Encapsulation relationships are handled in a similar fashion by spawning new paths in the automaton graph that end with a special state marked with the protocol that should follow. The exact usage of these marked states is explained in Section III-D. The generation procedure acts to counter the absence of explicit storage locations in the FSA model; when it becomes necessary to use the values of previously encountered fields for subsequent computations, the only solution is to spawn a number of parallel branches within the automaton, each one associated with a specific value of the field under consideration.

B. Multicast Packet Delivery

Here we discuss about packet forwarding to the nodes

Packet sending from the source

After the multicast tree is constructed, all the sources of the group could send packets to the tree and the packets will be forwarded along the tree. In most tree-based multicast protocols, a data source needs to send the packets initially to the root of the tree.

The source node want send the data to the members at that time we perform the security action, i.e. whenever the source node want to send the data , the source node can encrypt the data by using AES (Advanced Encryption Standers) the encrypted data can be transferred to the group members , in the transmission of packets the intermediate nodes want to read the data , if suppose the nodes can access the data that time we don’t have any problem because the data is in the encryption form i.e. cipher text , due to this text the intermediate nodes can’t get the data it can simply transfer the data to the destination, in the destination side the receiver can decrypt the data using AES algorithm.

For providing the security we use the Advanced Encrypted Standards Algorithm

The algorithm described by AES is a symmetric-key algorithm, meaning the same key is used for both encrypting and decrypting the data. The strength of a 128-bit AES key is roughly equivalent to 2600-bits RSA key. AES data encryption is a more mathematically efficient and elegant cryptographic algorithm the time required to crack an encryption algorithm is directly related to the length of the key used to secure the communication (It takes less time). AES allows you to choose a 128-bit, 192-bit or 256-bit key, making it exponentially stronger than the 56-bit key of DES (RSA). The algorithm was required to be royalty-free for use worldwide .AES has defined three versions, with 10, 12, and 14 rounds. Each version uses a different cipher key size (128, 192, or 256), but the round keys are always 128 bits.

IV. CONCLUSION

We have designed, prototyped, and evaluated SPAF, a packet filter generator based on the creation of finite-state automata from a high-level protocol format database and filter redicates. SPAF aims at emitting fast and efficient filters while preserving all the relevant safety properties, both in terms of memory access correctness and termination. The PacketScore scheme is used to defend against DDoS attacks. The key concept in PacketScore is the Conditional Legitimate Probability (CLP) produced by comparison of legitimate traffic and attack traffic characteristics, which indicates the likelihood of legitimacy of a packet. As a result, packets following a legitimate traffic profile have higher scores, while attack packets have lower scores. This scheme can tackle never-before-seen DDoS attack types by providing a statistics-based adaptive differentiation between attack and legitimate packets to drive selective packet discarding and overload control at high-speed.

Thus, PacketScore is capable of blocking all kinds of attacks as long as the attackers do not precisely mimic the sites' traffic characteristics. The performance and design tradeoffs of the proposed packet scoring scheme in the context of a stand-alone implementation is studied. By exploiting the measurement/scorebook generation process, an attacker may try to mislead PacketScore by changing the attack types and/or intensities. We can easily overcome such an attempt by using a smaller measurement period to track the attack traffic pattern more closely. We are currently investigating the generalized implementation of PacketScore for core networks.

In order to prove this technique on the field, we have developed a filter generator that creates filters from an external protocol database and user-specified rules. Filter DFAs can be used as they are by existing hardware or software engines or translated into C code by the back end.We also developed an *ad hoc* DFA execution engine that adapts its operations to the word size of the underlying machine instead of processing a byte at a time and enforces memory safety and termination through run-time fully aggregated bound checks.

The run-time performance and memory occupation of SPAF filters have been evaluated both in synthetic and real-world benchmarks. Test results show that FSA-based filters perform on a similar or improved level as other modern approaches such as BPF+, both on simple and complex filters; SPAF filters are also shown to scale better with increasing numbers of filtering rules. The measured overhead of run-time safety checks is small and does not cause any significant penalties both in times of run-times (few checks are executed per packet) and memory occupation (few checks are inserted per filter). Overall, the SPAF approach is an effective and simple way to generate packet filters that are easy to compose and efficient to run, even with increasing complexity.Among the potential problems, a widely known issue affecting specifically DFAs is an explosion occurring in the state space when treating certain critical patterns; this problem is the limiting factor for DFA adoption in other pattern-based detectors such as intrusion detection systems

The SPAF approach can be easily extended to perform packet demultiplexing in addition to packet filtering. This is partial supported by our current generator by labeling final states with identifiers of the matching filtering rules; full support would require dynamic automata creation and code generation, tasks that will be the object of future studies. Another future extension to SPAF could be enabling interactions (e.g., look-ups and updates) with stateful constructs such as session tables, useful for higher-layer filtering and traffic classification. In conclusion, SPAF has been shown as an approach that improves the state of the art by generating packet filters that combine most of the desired properties in terms of processing speed, memory consumption, flexibility and simplicity in specifying protocol formats and filtering rules, effective filter composition, and low run-time overhead for safety enforcement. The development of the filter generator and the test results support the viability of our claims.

REFERENCES

[1] S. McCanne and V. Jacobson, "The BSD packet filter: A new architecture for user-level packet capture," in Proc. USENIX, 1993, p. 2.

[2] A. Begel, S. McCanne, and S. L. Graham, "BPF□: Exploiting global data-flow optimization in a generalized packet filter architecture," SIGCOMM Comput. Commun. Rev., vol. 29, no. 4, pp. 123–134, 1999.

[3] M. L. Bailey, B. Gopal, M. A. Pagels, L. L. Peterson, and P. Sarkar, "PathFinder: A pattern-based packet classifier," in Proc. Oper. Syst. Design Implement., 1994, pp. 115–123.

[4] O. Morandi, F. Risso, M. Baldi, and A. Baldini, "Enabling flexible packet filtering through dynamic code generation," in Proc. IEEE ICC, May 2008, pp. 5849–5856.

[5] L. Degioanni, M. Baldi, F. Risso, and G. Varenni, "Profiling and optimization of software-based network-analysis applications," in Proc. 15th Symp. Comput. Arch. High Perform. Comput., Washington, DC, 2003, p. 226.

[6] S. Ioannidis and K. G. Anagnostakis, "xPF: Packet filtering for lowcost network monitoring," in Proc. HPSR, 2002, pp. 121–126.

[7] F. Risso and M. Baldi, "NetPDL: An extensible XML-based language for packet header description," Comput. Netw., vol. 50, no. 5, pp. 688–706, 2006.

[8] L. Degioanni, M. Baldi, D. Buffa, F. Risso, F. Stirano, and G. Varenni, "Network virtual machine (NetVM): A new architecture for efficient and portable packet processing applications," in Proc. 8th Int. Conf. Telecommun., Jun. 15–17, 2005, vol. 1, pp. 163–168.

[9] D. R. Engler and M. F. Kaashoek, "DPF: Fast, flexible message demultiplexing using dynamic code generation," in Proc. ACM SIGCOMM, New York, 1996, pp. 53–59.

[10] M. Jayaram, R. Cytron, D. Schmidt, and G.Varghese, "Efficient demultiplexing of network packets by automatic parsing," in Proc. Workshop Compiler Support Syst. Softw., 1996.

[11] S. Kumar, J. Turner, and J. Williams, "Advanced algorithms for fast and scalable deep packet inspection," in Proc. ACM ANCS, New York, 2006, pp. 81–92.

[12] S. Kumar, S. Dharmapurikar, F. Yu, P. Crowley, and J. Turner, "Algorithms to accelerate multiple regular expressions matching for deep packet inspection," in Proc ACM SIGCOMM, New ork, 2006, pp. 339–350.

[13] M. Becchi and P. Crowley, "An improved algorithm to accelerate regular expression evaluation," in Proc. ACM ANCS, New York, 2007, pp. 145–154.

[14] F. Yu, Z. Chen, Y. Diao, T. V. Lakshman, and R. H. Katz, "Fast and memory-efficient regular expression matching for deep packet inspection," in Proc. ACM ANCS, New York, 2006, pp. 93–102.

[15] M. Becchi and P. Crowley, "A hybrid finite automaton for practical deep packet inspection," in Proc. ACM CoNEXT, New York, 2007, pp. 1–12.

[16] R. Smith, C. Estan, and S. Jha, "XFA: Faster signature matching with extended automata," in Proc. IEEE Symp. Security Privacy, 2008, pp. 187–201.

[17] M. Becchi, M. Franklin, and P. Crowley, "A workload for evaluating deep packet inspection architectures," in Proc. IEEE Int. Symp. Workload Characterization, Sep. 2008, pp. 79–89.

[18] T. Hruby, K. van Reeuwijk, and H. Bos, "Ruler: High-speed packet matching and rewriting on NPUs," in Proc. ACM ANCS, New York, 2007, pp. 1–10.

[19] H. Bos, W. D. Bruijn, M. Cristea, T. Nguyen, and G. Portokalidis, "FFPF: Fairly fast packet filters," in Proc. OSDI, 2004, pp. 347–363.

[20] L. Deri, "nCap: Wire-speed packet capture and transmission," in Proc. IEEE E2EMON, Washington, DC, 2005, pp. 47–55.

[21] Z. Wu, M. Xie, and H. Wang, "Swift: A fast dynamic packet filter," in Proc. 5th USENIX Symp. Netw. Syst. Design Implement., Berkeley,CA, 2008, pp. 279–292.

◈◈◈

SAT: A Security Architecture Achieving Routing Anonymity in Wireless Mesh Networks

Tanuja Chinta & K. Madhavi

CSE Dept., JNTUA College of Engineering, Anantapur, Andhra Pradesh, India
E-mail : tanuja.chinta@gmail.com, kasamadhavi@yahoo.com

Abstract - Wireless Mesh Network(WMN) is a promising technology and is expected to be widespread due to its low investment feature and the wireless broadband services it supports, attractive to both service providers and users. Some security conflicts will occur in WMN's. In this paper, we propose a security architecture achieving routing anonymity, anonymity and traceability. The proposed architecture resolve security conflicts that will occur in WMN's. In addition, architecture guarantee fundamental security requirements including authentication, confidentiality, data integrity and non repudiation. Thorough analysis on security and efficiency is incorporated demonstrating the feasibility and effectiveness of the proposed architecture.

Keywords - Routing anonymity, anonymity, traceability, pseudonym, misbehavior, wireless mesh networks(WMN).

I. INTRODUCTION

Wireless Mesh Networks is a promising technology and it is attractive to both service providers and users. Security issues inherent in WMN need to be considered before the deployment and proliferation of these networks. In WMN's, fundamental operations need to be secured. Anonymity, Routing anonymity and traceability is achieved to secure the fundamental operations in WMN's. The requirement for anonymity is to unlink a user's identity to his or her specific activities. Anonymity is also required to hide the location information of a user to prevent movement tracing. Routing anonymity is required to resolve traffic analysis attacks. Routing anonymity hides the confidential communication relationship of two parties by building an anonymous path between them. Traceability is required to detect misbehaving users in WMN's. In this paper, security conflicts namely routing anonymity, traceability and anonymity can be resolved in the emerging wireless communication systems. Blind signature can be implemented in the architecture to achieve anonymity. Pseudonym technique is used to hide the user location information. RSA algorithm is used to achieve routing anonymity.

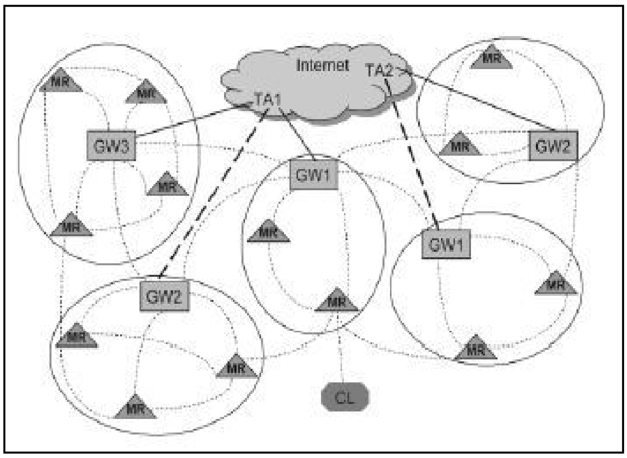
II. PRELIMINARIES

Blind signature-Blind signature is introduced by Chaum. Blind signature scheme allows a receiver to obtain a signature on a message such that both the message and the signature remain unknown to the signer. Brands developed restrictive blind signature scheme. Restrictive blind signature scheme means restrictiveness property is incorporated into blind signature. This property restricts the user in the blind signature scheme to embed some account related information. Restrictiveness property is used to guarantee traceability in the proposed system. Restrictive partially blind signature schemes serve as a building to the proposed architecture.

Restrictiveness - Let a message m be such that the receiver knows a representation of (a_1,a_2,..., a_k) of m with respect to a generator tuple (g_1,...,g_k) at the beginning of a blind signature protocol. Let (b_1,...,b_k) be the representation the receiver knows of the blinded message m_1 of m after the completion of the protocol. If there exist two functions I_1 and I_2 such that I_1(a_1,a_2,..., a_k)=I_2 (b_1,...,b_k), regardless of m and the blinding transformations applied by the receiver, then the protocol is called a restrictive blind signature protocol. The functions I_1 and I_2 are called blinding-invariant functions of the protocol with respect to (g_1,...,g_k).

Partial Blindness - A signature scheme is partially blind if, for all probabilistic polynomial-time algorithm A, A wins the game in the signature issuing protocol with probability at most 1/2+1/k for sufficiently large k and some constant €. The probability is taken over coin flips of KG,U_0,U_1, and A, where KG is the key generation function, U_0 and U_1 are two honest users.

III. NETWORK ARCHITECTURE



The above figure represents network topology of a WMN. The WMN consists of mesh routers(MR's) and gateways(GW's) which are interconnected by wireless links(shown as dotted curves). Mesh routers and gateways serve as access points in WMN. An architecture is divided in to wireless mesh domains(WMN's). Each domain contains one administrator called trusted authority(TA). Each domain is managed by trusted authority. The client(CL) access the network services from the internet through gateways and mesh routers.

IV. SAT SECURITY ARCHITECHTURE

Ticket based approach is implemented in this architecture to provide security. so, architecture is named as ticket-based security architecture. Ticket-based security architecture contains four protocols. They are ticket issuance, ticket deposit, fraud detection and ticket revocation.

Ticket Issuance-

Ticket issuance protocol can be implemented as follows:

1. $CL \rightarrow\rightarrow TA: IDCL.m.t1, HMACK(m||t1);$
2. $TA \rightarrow\rightarrow CL: IDTA,\ X = e(m,TTA),\ Y = e(P,Q),$

   $Z = e(m,Q),\ U = rH1\ (IDTA),$

   $V = rP, t2, HMACk(X||Y||Z||U||V||t2);$
3. $CL \rightarrow\rightarrow TA{:}IDCL,$

   $B = \frac{1}{\lambda} H_2(m'||U'||V'||R||W||X'||Y'||Z')$

   $+\mu,\ t_3, HMAC_K(B||\ t_3);$ and
4. $TA \rightarrow\rightarrow CL : ID_{TA},\ \sigma_1 = Q + B\Gamma_{TA}$

   $\sigma_2 = (r + B)\ \Gamma_{TA} + r\ H_1(c), t_4,$

   $HMAC_k(\sigma_1 \ ||\ \sigma_2\ ||\ t_4).$

Ticket Deposit-

Ticket deposit protocol can be implemented as follows:

1. $CL \rightarrow\rightarrow GW : PS_{CL},\ m',\ W,\ c,$

   $\sigma = (U',\ V',\ X',\ \rho, \sigma'_1,\ \sigma'_2),\ t_5,$

   $SIG_{TCL}\text{~}(m'\ ||\ w\ ||\ c\ ||\ \sigma\ ||\ t_5);$
2. $GW \rightarrow\rightarrow CL{:}ID_{GW}, d{=}H_3(R\ ||\ W\ ||\ ID_{GW}\ ||\ T), t_6,$

   $HMAC_{K'}(d\ ||\ t_6);$
3. $CL \rightarrow\rightarrow GW : PS_{CL}, r_1{=}d(u_1\alpha){+}v_1, r_2{=}d\alpha{+}v_2,$

   $t_7,\ HMAC_{K'}(r_1\ ||\ r_2\ ||\ t_7);$ and
4. $GW \rightarrow\rightarrow CL{:}ID_{GW}, misb, exp, t_8,$

   $SIG_{TGW}(PS_{CL}\ ||\ ID_{GW}\ ||\ misb\ ||\ exp\ ||\ t_8);$

Fraud Detection-

Fraud detection protocol can be implemented as follows:

1. $GW \rightarrow TA : ID_{GW}, m', W, c,$

   $\sigma{=}(U',V',X',\rho,\sigma_1',\sigma_2'), r_1, r_2, T, t_9,$

   $HMAC_{K''}(m'\ ||\ W\ ||\ c\ ||\ \sigma\ ||\ r_1\ ||\ r_2\ ||\ T\ ||\ t_9);$

Ticket Revocation-

Ticket revocation protocol can be implemented as follows:

- $CL \rightarrow\rightarrow MR : PSTCL, aP0, t12,$

  $HIDS\varphi\ CL, SCL(H1'(PSTCL\ ||\ ap0\ ||\ t12));$
- $MR \rightarrow CL : IDTMR, bP0, t13,$

  $HIDS\varphi MR, SMR(H1'(IDTMR\ ||\ bP0\ ||\ t13));$ and
- $CL \rightarrow\rightarrow MR : PSTCL, PSCL, SK\varepsilon k(IDCL\ ||\ m),$

  $t14, HMACK(PSCL\ ||\ SK\varepsilon\ ||\ t14).$

V. ROUTING ANONYMITY

Routing anonymity can be implemented by using RSA algorithm in the proposed security architecture. Public key algorithm invented in 1977 by Ron Rivest, Adi Shamir and Leonard Adleman (RSA). It supports Encryption and Digital Signatures. It is the most widely used public key algorithm. It gets its security from integer factorization problem. It is relatively easy to understand and implement. It is patent free(Since 2000). In order to implement RSA you will need arbitrary precision arithmetic (multiple precision arithmetic), Pseudo Random Number Generator (PRNG) and prime number generator. The difficulty of implementation greatly depends on the target platform, application usage and how much of the tools you need to implement from scratch.

RSA Algorithm

Key generation:

- Select random prime numbers p and q, and check that $p != q$
- Compute modulus $n = pq$
- Compute phi, $= (p - 1)(q - 1)$
- Select public exponent e, $1 < e <$ such that $gcd(e,) = 1$
- Compute private exponent $d = e - 1 \ mod$
- -Public key is $\{n, e\}$, private key is d

Encryption: $c = me \ mod \ n$, decryption: $m = cd \ mod \ n$

Digital signature: $s = H(m)d \ mod \ n,$ verification: $m' = semod \ n,$

if $m' = H(m)$ signature is correct. H is a publicly known hash function.

RSA Key Generation

If the RSA keys does not exist, they need to be created. The key generation process is usually relatively slow but fortunately it is performed seldom (the very first time and then only if keys need to be regenerated). The key generation starts by finding two distinct prime numbers p and q. First PRNG is used to generate random numbers, then they are tested for primality and will be regenerated until prime numbers are found. The p and q must same length in bits, must not be equal, and they should not be close to each other (that is $p - q$ should not be small number). If primes are chosen random, and even when they are same in length, it is extremely likely these conditions are met. Compute modulus $n = pq$ and = (p-1)(q-1). The n will be stored for later as it is part of the public key. To have 1024 bit public key, then *p and q* are about 512 bits each. Select public exponent e, which is used as public key with n. It is used to encrypt messages and to verify digital signatures. The e is stored for later with n. The e is usually small number but it can be $1 < e <$. The e must be relatively prime to , hence $gcd(e, = 1$ *(*gcd = greatest common divisor, use Euclidean algorithm*)*. Usually e is small to make encryption faster. However, using very small e (<16 bit number) is not recommended. A popular starting value for e is 65537. If e is not relatively prime to , then it is usually added by 2 until it becomes relatively prime. This makes the finding of e as fast as possible. Compute private exponent d, which is the actual RSA private key.

The d must not be disclosed at any time or the security of the RSA is compromised. The d is found by computing the multiplicative inverse $d = e - 1 \ mod.$ The extended Euclidean algorithm is commonly used to compute inverses. The d exponent is used to decrypt messages and to compute digital signatures. Implementations try to find as small d as possible to make decryption faster. This is fine as long as it is assured that d is about the same size as n. If it is only one quarter of size it is not considered safe to be used. It is possible to find a smaller d by using $lcm(p-1,q-1)$ instead of (lcm = least common multiple, $lcm(p-1,q-1) =/ gcd(p-1, q-1)).$ Key generation is the most important part of RSA, it is also the hardest part of RSA to implement correctly. Prime numbers must be primes, otherwise the RSA will not work or is insecure. There exists some rare composite numbers that make the RSA work, but the end result is insecure. Find fast implementation of the extended Euclidean algorithm. Do not select too small e. Do not compute too small d. Compute at least 1024 bit public key. Smaller keys are nowadays considered insecure. If you need long time security compute 2048 bit keys or longer. Also, compute always new n for each key pair. Do not share n with any other key pair (common modulus attack). Test the keys by performing RSA encryption and decryption operations.

VI. SECURITY ANALYSIS

In this section, we are going to analyze the security requirements our system can achieve. Our security architecture satisfies the security requirements for authentication, data integrity and confidentiality by implementing digital signature, message authentication code and encryption in our system.

Anonymity - Anonymity can be easily shown that a gateway cannot link a client's network access activities to his real identity. It can be implemented by using pseudonyms in authentication which reveals no information on the real id. Client's home TA cannot link client's network access activities to the real id by implementing restrictive partially blind signature scheme in our system.

Traceability- Traceability means conditional anonymity. It can be needed for misbehaving users. Unconditional anonymity is needed for honest clients. Traceability can be implemented by using restrictive partially blind signature scheme.

VII. EFFICIENCY ANALYSIS

Most of the pairing-based cryptosystems are needed to work in 1) a subgroup of the elliptic curve E(Fq) of sufficiently large prime order p, and 2) a sufficiently large finite field Fqk, where is the size of the field and k is the embedding degree. For minimum levels of security, p>2160 and qk>21024 is required to ensure the hardness of the DLP in G1 and G2.To improve the computation and communication efficiency when working with E(Fq), we need to put q value as small while maintaining the security with large values of k. SHA-1 is used to compute the keyed-hash message authentication code which yields a 160-bit output to improve efficiency.

Communication- Our ticket-based security architecture consists of four intradomain protocols. These protocols are distributed in nature. So, the communication cost incurred is more affordable.

Storage- TA may contain several servers to store client's necessary information. The storage capacity of these high end servers is not a concern. So, we need to focus on the storage overhead encountered at the low-end client side. There is a tradeoff between storage and computation overhead. In our protocols, the client need to perform pairing computations frequently, which is impractical due to the high cost of pairings and limited power of clients. Many pairing operations in the protocol can be computed once and stored for future use. Some stored information remains unchanged for all instances of protocol execution. As a result, we need merely take into account the effective storage overhead (i.e., information that is changed and has to be stored at each protocol instance).

VIII.CONCLUSION

In this paper, we propose security architecture consists of ticket-based protocols which resolves the security conflicts namely routing anonymity, traceability and anonymity. Security conflicts can be resolved by utilizing the tickets, selfgenerated pseudonyms and hierarchical identity-based cryptography in the architecture to achieve desired security objectives and efficiency.

REFERENCES

[1] J. Sun, C. Zhang, and Y. Fang, “A Security Architecture Achieving Anonymity and Traceability in Wireless Mesh Networks,” Proc. IEEE INFOCOM, pp. 1687-1695, April. 2011.

[2] D. Boneh and M. Franklin, “Identity-Based Encryption from the Weil Pairings,” Advances in Cryptology-Asiacrypt 2001, pp. 514-532, Springer-Verlag, 2001.

[3] A. Juels, M. Luby, and R. Ostrovsky, “Security of Blind Digital Signatures,” Advances in Cryptology—Crypto ’97, pp. 150-164, Springer-Verlag, 1997.

◈◈◈

Image Enhancement Based on Lazy Wavelet Transform Using Adaptive Wiener With Threshold

G. M. Rajathi[1], M.Sarathkumar[2] & R. Rangarajan[3]

[1&2]Dept of ECE, Sri Ramakrishna Engineering college, Coimbatore, Tamil Nadu, India
[3]VSB Engineering College, Karur Coimbatore, Tamil Nadu, India
E-mail : gmrajathi@yahoo.com[1], sarath.samy@gmail.com[2]

Abstract **-** Image enhancement techniques are to emphasize and sharpen the features of image for better display. In this paper, we have proposed an enhanced adaptive wiener filter based on fast lifting wavelet transform using thresholding technique. Image denoising is performed in order to increase the image SNR that improves the image quality and give prominence to the desired features of the image so as to get better denoising effects. The 2-D Lifting based lazy wavelet transform, converts the noisy image pixels into wavelet components. Then the thresholding is applied in wavelet domain using Bayes Shrink thresholding. The thresholding operators like soft and hard contain discontinuities at some point in the output. this paper approaches an improved model of wavelet coefficient evaluation, which is called the semi-soft threshold method which removes such discontinuities.After that, an adaptive Wiener filter is applied to all the sub-band images. Finally, these sub-band images are inversely transformed to the spatial domain, to reconstruct the final improved image. Adaptive weiner performs filtering with different filter coefficients for each sub-band which improves the image quality and since the filtering is performed in the frequency domain leads to efficient removal of noise pixels with better image resolution.

Keywords - *lifting based lazy wavelet transform , bayesian threshold , Lifting based wavelet domain adaptive wiener.*

I. INTRODUCTION

Images are often corrupted by noise owing to channel transmission errors, faulty image acquisition devices, engine sparks, power interference and atmospheric electrical emissions. A fundamental problem in image processing is to remove the additive white Gaussian noise without blurring the fine details of the images. A vast literature has emerged recently on image enhancement using linear techniques or nonlinear techniques. Most of the linear techniques, such as averaging low-pass filters have low-pass characteristics and they tend to blur edges and to destroy lines and other fine image details. The rapid development of wavelet theory and its good time-frequency characteristics lead the wavelet domain denoising into a broader application of the method of image denoising.

During the last decades, a lot of new methods based on wavelet transforms have emerged. Such as: Mallat's proposed wavelet denoising method based on wavelet transform maximum principle[1]; Xu and others put forward wavelet denoising methods based on wavelet transform scale correlation between the wavelet coefficients[2]; Donoho and others put forward soft-threshold and hard- threshold wavelet denoising methods[3],[4],[1]. Due to the simple and effective algorithm, Wavelet denoising methods based on hard-threshold and soft-threshold are widely used. but both methods contain discontinuities at some point in the output. In the additive noise scenario, the hard and soft threshold value does not always lead to satisfactory results since it does not take into account the statistical properties of the image to be enhanced. To overcome the problems, in this method, we proposed a method for image enhancement, which is based on second-generation wavelet transform with optimized thresholding called as semi soft threshold and an adaptive Wiener filtering. The second generation wavelet transform is constructed by the lifting scheme. A lifting scheme is a new method for constructing wavelets, which is entirely spatial and therefore ideally suited for building second generation wavelets.

The lifting scheme allows a faster implementation of the wavelet transform (WT) and a fully in-place calculation of the WT. Moreover, no extra memory is needed and the original signal can be replaced with its WT. Bayesian threshold applies different threshold value for each sub-band. The adaptive Wiener filter utilizes statistical properties of the image estimated from a local neighbourhood of each pixel in wavelet domain.

In the implementation of the proposed algorithm, the experimental results have shown that the performance of our proposed image enhancement method is noticeably better than wavelet domain wiener filter (WDWF) and bayesianshrink threshold in wavelet domain.

II. FAST LIFTING WAVELET TRANSFORM

A. *Lifting scheme:*

The basic principle of the lifting scheme is to factorize thepolyphase matrix of a wavelet filter into a sequence of alternating upper and lower triangular matrices and a diagonal matrix[4~7]. This leads to the wavelet implementation by means of banded-matrix multiplications.

Let *h~(z) and g~*(*z*) be the lowpass and highpass analysis filters, and let *h*(*z*) *and g*(*z*) be the lowpass and highpass synthesis filters. The corresponding polyphase matrices are defined as

$$\tilde{P}(z)=\begin{bmatrix}\tilde{h}_e(z) & \tilde{h}_o(z)\\ \tilde{g}_e(z) & \tilde{g}_o(z)\end{bmatrix} \text{ and } P(z)=\begin{bmatrix}h_e(z) & h_o(z)\\ g_e(z) & g_o(z)\end{bmatrix}$$

It has been shown that if (*h~*, *g~*) is a complementary filter pair, then *P~*(*z*) can always be factored into lifting steps as

$$\tilde{P}(z)=\begin{bmatrix}1/K & 0\\ 0 & K\end{bmatrix}\prod_{i=1}^{m}\begin{bmatrix}1 & \tilde{s}_i(z)\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ \tilde{t}_i(z) & 1\end{bmatrix}$$

where *K* is a constant. The lifting schemes are shown in Fig.1(a) and 1(b)

Lifting wavelet transform scheme consist of three steps:

1) *Split* step, where the input samples are split into two even and odd samples by the time domain;

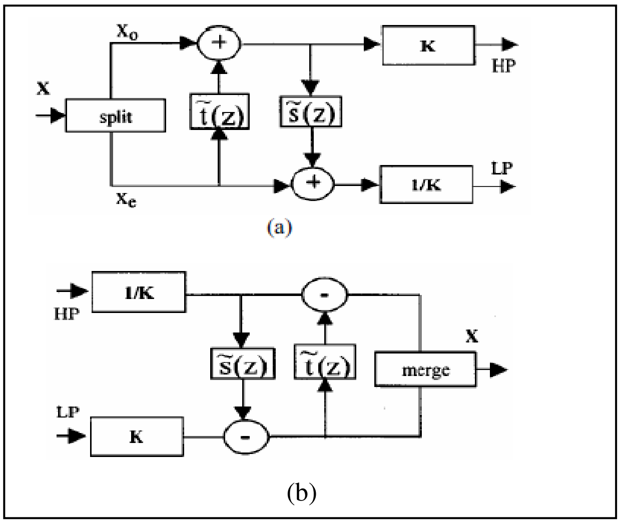
$$X_e(n)=X(2n) \text{ and } X_o(n)=X(2n+1)$$

2) *Predict* step, where the even samples are multiplied by the time domain equivalent of *~t* (*z*) and are added to the odd samples and the results are multiplied by scaling factor *K* .

$$X^{HP}(z)=K[\tilde{t}(z)X_e(z)+X_o(z)]$$

3) *Update* step, where updated odd samples are multiplied by the time domain equivalent of and are added to the even samples and the results are multiplied by scaling factor1/ *K* .

$$X^{LP}(z)=1/K\cdot\{X_e(z)+\tilde{s}(z)[\tilde{t}(z)X_e(z)+X_o(z)]\}$$



(a)

(b)

Fig.1(a) and 1(b).

The inverse WT is obtained by traversing in the reverse direction, changing the factor *K* to 1/ *K* , factor 1/ *K* to *K* , and reversing the signs of coefficients in *~t* (*z*) and *~s* (*z*) .

$$f_L(x,y)=\sum_x f(x,y)\tilde{h}_x,\ f_H(x,y)=\sum_x f(x,y)\tilde{g}_x$$

The follow illustrates the relations between Mallet algorithm and lifting algorithm when input samples are filtered.

B. *Lazy wavelet transform:*

Wavelets are capable of quickly capturing the essence of a data set with only a small set of coefficients. This is based on the fact that most data sets have correlation both in time (or space) and frequency.

Because of the time-frequency localization of wavelets, efficient representations can be obtained.,where the wavelets are not necessarily translates and dilates of each other but still enjoy all the powerful properties of first generation wavelets. These wavelets are referred to as second generation wavelets. The reason is that translation and dilation become algebraic operations in the Fourier domain. Many real life problems require algorithms adapted to irregular sampled data, while first generation wavelets imply a regular sampling of the data.

In Lazy wavelet[7], the general index sets K(j) and M(j) are the generalization of the even and odd indices. E and D are used to denotes the two sub sampling operators for even ,odd respectively.

$$E:\ell^2(K(j+1))\rightarrow\ell^2(K(j)), \text{ where } b=Ea \text{ means that } b_k=a_k \text{ for } k\in K(j).$$

$$D:\ell^2(K(j+1))\rightarrow\ell^2(M(j)), \text{ where } c=Da \text{ means that } c_m=a_m \text{ for } m\in M(j).$$

Although the operator depends on the level j no extra subscript are supplied These operators provide a trivial orthogonal splitting, as

$$\begin{bmatrix} E \\ D \end{bmatrix} \begin{bmatrix} E^* & D^* \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} E^* & D^* \end{bmatrix} \begin{bmatrix} E \\ D \end{bmatrix} = 1 .$$

now decompose any operator

$$\cdot W : \ell^2(\mathcal{K}(j)) \to \ell^2(\mathcal{K}(j)) \text{ as}$$

$$W = W_e E + W_d D, \quad \text{with} \quad W_e = W E^* \quad \text{and} \quad W_d = W D^* .$$

The filter operators of the Lazy wavelet are precisely these sub sampling operators

$$H_j^{\text{Lazy}} = \widetilde{H}_j^{\text{Lazy}} = E \quad \text{and} \quad G_j^{\text{Lazy}} = \widetilde{G}_j^{\text{Lazy}} = D .$$

The Lazy wavelet transform thus is an orthogonal transform that essentially does nothing. It only resample's the coefficients into two groups each step and thus can be seen as the generalization of the polyphase transform to the second generation setting. However, it is important to consider since it is connected with interpolating scaling functions. The operators E and D are crucial when implementing the lifting scheme. With data structure, the implementation of the lifting scheme is straightforward.

C. Wavelet decomposition:

Let $f(x,y)$ represents an image. Implementing the lifting wavelet filters $\tilde{h}_x$ and $\tilde{g}_x$ to the image $f(x,y)$ in the x -direction can acquire low and high frequency constituents in the x -direction. After down sampling by 2, images have reduced to half on the x-direction. $f_L(x,y)$ provides low and high frequency components of the image in the x-direction, correspondingly.

$$f_L(x,y) = \sum_x f(x,y)\tilde{h}_x \,, f_H(x,y) = \sum_x f(x,y)\tilde{g}_x$$

where $\tilde{h}_x$ and $\tilde{g}_x$ represents low-pass and high-pass decomposition filters. Correspondingly as above implementing the lifting wavelet filters $\tilde{h}_x$ and $\tilde{g}_x$ to the two sub images (f_L, f_H) in the y-direction four sub images $(f_{LL}, f_{LH}, f_{HL}, f_{HH})$ are obtained. $f_{LL}(x,y)$ shows decomposition of the image into four sub images.

$$f_{LL}(x,y) = \sum_y f_L(2x,2y)\tilde{h}_y \,, f_{LH}(x,y) = \sum_y f_L(2x,2y)\tilde{g}_y$$

$$f_{HL}(x,y) = \sum_y f_H(2x,2y)\tilde{h}_y \,, f_{HH}(x,y) = \sum_y f_L(2x,2y)\tilde{g}_y$$

The 2-D lifting-based wavelet transform (LBWT) generates one set of approximation coefficients (f_{LL}) and three sets of detail coefficients (f_{LH}, f_{HL}, f_{HH}). Additional decomposition can be accomplished by substituting upon the f_{LL} subband consecutively and the resulting image is divided into multiple bands. The reconstruction of the image can be performed by the following process. Initially, all four sub-bands at coarsest scale are up sampled by a factor of two, and the sub-bands are filtered with low-pass h and high-pass g synthesis filters in all dimension. Then four filtered sub-bands are added to accomplish the low sub-band at the subsequent finer scale. The process is continued until the image is completely reconstructed.

III. BAYESSHRINK ALGORITHM

BayesShrink was proposed by Chang, Yu and Vetterli . The goal of this method is to minimize the Bayesian risk, and hence its name, BayesShrink. It is a subband – dependent which means that threshold level is selected at each band of resolution in the wavelet decomposition.. The Bayes threshold, t_b , is defined as

$$t_b = \frac{\sigma^2}{\sigma_s}$$

where σ^2 is the noise variance and

σ_s^2 is the signal variance without noise

The noise variance σ^2 can be estimated from the subband HH1 in the decomposition of wavelet by the median estimator.

From the definition of additive noise we have

$$w(x,y) = s(x,y) + \eta(x,y)$$

Since the signal and noise are independent of eachother it can be stated that

$$\sigma_w{}^2 = \sigma_s{}^2 + \sigma^2$$

σ_w^2 can be calculated as shown below,

$$\sigma_w{}^2 = \frac{1}{n^2} \sum_{x,y-1}^{n} w^2\,(x,y)$$

The variance of the signal σ^2 is computed as shown below

$$\sigma_s = \sqrt{\max\,(\sigma_w{}^2 - \sigma^2, 0)}$$

With these σ^2 and σ_w^2 the Bayes threshold is computed from the below equation

$$t_b = \frac{\sigma^2}{\sigma_s}$$

the wavelet coefficients are thresholded at each band.

A threshold value for each resolution level in the wavelet transform which is referred to as level dependent thresholding.

The main advantage of bayesianShrink is to minimize the mean squared error, unlike Visu Shrink.

IV. THRESHOLDING OPERATORS

4.4 General thresholding operators are

I. Hard thresholding.

II. Soft thresholding.

III. Semi-soft thresholding.

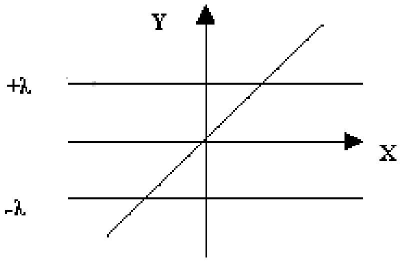


Fig. 2(a) : original image

4.4.1 Hard thresholding:

The hard-thresholding T_H can be defined as

Here t is threshold value.plot for this is as shown below

$$T_H = \begin{cases} x & |X|>\lambda \\ 0 & \text{otherwise} \end{cases}$$

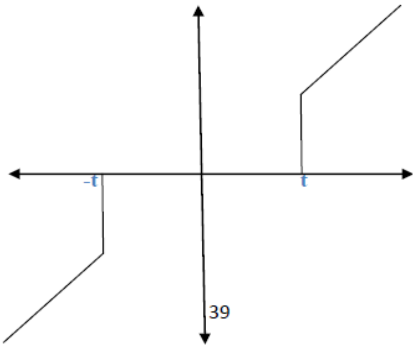


Fig. 2(b) : Hard thresholding

In this, all coefficients whose magnitude is greater than the selected threshold value *t* remain same and the others whose magnitude is smaller than *t* are set to zero. It creates a region around zero where the coefficients are considered negligible.

4.4.2 Soft thresholding:

In Soft thresholding, the coefficients whose magnitude is greater than the selected threshold value are become shrinks towards zero and others set to zero.

The Soft-thresholding T_s can be defined as

$$T_s = \begin{cases} \text{sign}(x)(|x|-\lambda) & \text{for } |x|>\lambda \\ 0 & \text{otherwise} \end{cases}$$

Plot as shown below,

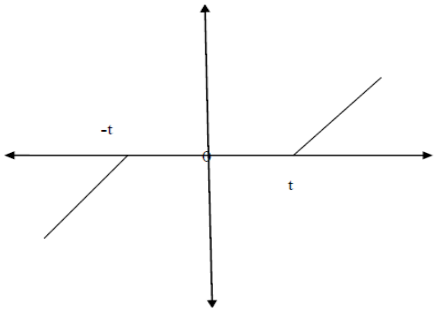


Fig. 2(c) : Soft thresholding

In practice, it can be seen that the soft method is much better and yields more visually pleasant images. This is because the hard method is discontinuous and yields abrupt artifacts in the recovered images. Also, the soft method yields a smaller minimum mean squared error compared to hard form of thresholding.

4.4.3 Semi-soft thresholding:

Hard and soft thresholding are two specific non-linear diagonal estimator, but one can optimize the non-linearity to capture the distribution of wavelet coefficient of a class of images. Bruce and Gao (1997) showed that hard thresholding would cause a bigger variance, while soft thresholding will tend to have a bigger bias because all larger coefficients are reduced by l. To prevent the drawback of hard and soft thresholding, they proposed a semi-soft thresholding approach as given by

$$T_{\text{semi-soft}}(x,y) = \begin{cases} 0 & |X| \le \lambda \\ \text{sign}\{x\}\, \lambda'(|x|-\lambda)/(\lambda'-\lambda) & T < |X| \le \lambda' \\ x & |X| > \lambda' \end{cases}$$

Semi soft thresholding is a family of non-linearities that interpolates between soft and hard thresholding. The aim of semi-soft is to offer a compromise between hard and soft thresholding by changing the gradient of the slope. This scheme requires two thresholds, a lower threshold λ and an upper threshold λ' where λ' is estimated to be twice the value of lower threshold λ. There is no attenuation for inputs beyond λ'.

For inputs below or equal to λ’, the output is forced to zero. For inputs that lie between λ and λ’ the output depends on the gradient formula,

Gradient= sign{x} λ’ (|x| - λ) / (λ’ - λ).

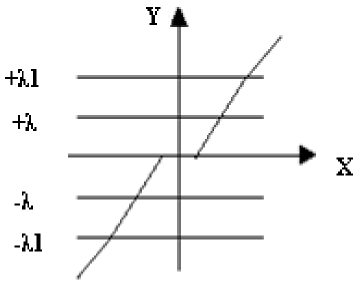


Fig. 2(d) :Semi-soft thresholding

V. LIFTING-BASED WAVELET TRANSFORM ADAPTIVE WIENER

Spatial wiener utilizes the same filter coefficients for the entire image leads to loss of information.Spatial domain adaptive wiener for image de-noising is well known for its high PSNR but the visual performance is not acceptable.

Lifting based adaptive wiener utilizes different filter coefficients for each resolution band of the image leads to both increased PSNR as well as the visual quality of the image.

Applying adaptive wiener to the thresholded image in the transform domain for all sub-bands.

By transforming the observed noisy image *y*(*i*, *j*) into wavelet domain, we obtain four sub-images as $A_{y(i,j)}$,$W_{dy(i,j)}$ (d = 1,2,3) which denote approximation coefficients and three sets of detail coefficients, respectively. Each sub-band can be supposed as a band limited spatial signal and the noise in each sub-band can be supposed as white noise. Wiener filtering can be applied to each sub-band to suppress the corresponding white noise. The lifting based wavelet domain wiener filter can be constructed as

$$\hat{m}_{W_{dy(i,j)}} = \frac{1}{(2m+1)(2n+1)} \sum_{k=i-m}^{i+m} \sum_{l=j-n}^{j+n} W_{dy(k,l)}$$

$$\hat{\sigma}^2_{W_{dy(i,j)}} = \frac{1}{(2m+1)(2n+1)} \sum_{k=i-m}^{i+m} \sum_{l=j-n}^{j+n} [W_{dy(k,l)} - \hat{m}_{W_{dy(i,j)}}]^2$$

$$\hat{\sigma}^2_{W_{dx(i,j)}} = \max\{0, \sigma^2_{W_{dy(i,j)}} - \sigma^2_n\}$$

$$W_{d\hat{x}(i,j)} = \hat{m}_{W_{dy(i,j)}} + \frac{\hat{\sigma}^2_{W_{dx(i,j)}}}{\hat{\sigma}^2_{W_{dx(i,j)}} + \sigma^2_n} (W_{dy(i,j)} - \hat{m}_{W_{dy(i,j)}})$$

$m\hat{}_{wdy(i,j)}$ and $\sigma\hat{}^2_{W(i,j)}$ are local statistics updated at each pixel in each sub-band of the original images, but they can be estimated from each sub-band of the observed image.

VI. SIMULATION RESULTS

Simulations have been performed on a grey-scale image, of size 512.512 of ‘Lena’. For natural images contaminated with computer generated additive Gaussian white noise, we adopt the peak signal-to-noise ratio (PSNR) to test the denoising effect.

The definition of PSNR is:

$$PSNR = -20\log_{10} \frac{\sqrt{\| x(m,n) - \hat{x}(m,n) \|^2}}{M \times N}$$

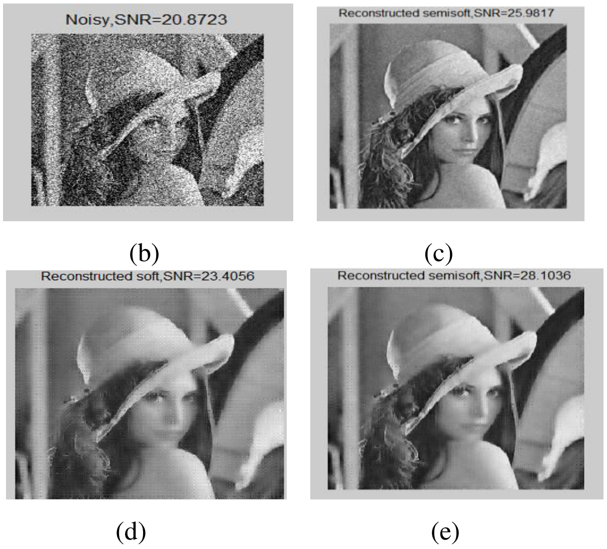
where *m* = 1,2,… ,*M* . *n* = 1,2,…., *N* are positive integers.

x(*m*, *n*) is the original image, *x*ˆ(*m*, *n*) is the reconstruction image or noised image.

To evaluate the performance of the proposed method we compared our image enhancement method to widely used image enhancement methods. the lifting based wavelets in this study belong to lazy. The first method was lifting-based wavelet domain adaptive Wiener filter(LBWDAW). The second method was lifting-based wavelet domain thresholding technique(LBWDT). . The proposed method was lifting-based wavelet domain thresholding with adaptive wiener.

The performance was evaluated in terms of the peak signal-to-noise ratio (PSNR) and visual inspection. The image was corrupted with the white Gaussian noise of various power levels. The image corrupted by the white Gaussian noise is shown in the figure 3(b). Figures 3(c),3(d) and 3(e) plot the result of LBWDT and LBWDAW , and the proposed method, respectively. It is obvious from Fig.2 that our method shows a superior performance to SDWF and CWDWF in the image enhancement.

(a)


Noisy,SNR=20.8723
Reconstructed semisoft,SNR=25.9817
Reconstructed soft,SNR=23.4056
Reconstructed semisoft,SNR=28.1036

(b) (c)

(d) (e)

Fig. 3 (a) : The original Lena image (b) noisy image (c) Result of LBWDT (d) Result of LBWDAW (e) Result of the proposed method

Table.1 shows the performance comparison of the proposedmethod, LBWDAW, and LBWDT for the image ‘Lena’ corrupted with additive white Gaussian noise with zero mean under various input PSNRs. The noise level was adjusted to specific values so as to obtain the different PSNRs. As seen from Table.1, the proposed method gave the best results.

Table 1

VARIANCE	**Proposed method**	**LBWDAW**	**LBWDT**
0.02	26.52	22.43	24.31
0.04	26.26	21.36	23.83
0.06	26.25	21.20	23.50
0.08	26.86	19.84	22.16
0.10	24.93	17.98	21.82
0.12	23.21	16.58	20.14

Performance comparison of LBWDAW, LBWDT and proposed method

VII. CONCLUSION

We have proposed an image enhancement method that relies on an adaptive Wiener filter and bayesian threshold in lifting-based wavelet domain for an image corrupted by AWGN. The proposed method utilizes the multi-scale characteristics of lifting based wavelet transform and the local statistics of each sub-band. In the proposed method, a noisy image is wavelet transformed up the third scale and an adaptive Wiener filter and threshold is applied to each sub-band. The effects of increasing the number of decomposition levels and the different kinds of wavelet filters on the proposed method have been investigated. The proposed method shows a better performance with lazy wavelet filter.

Each of sub-bands is processed independently in the wavelet domain by the threshold and Wiener filter. The simulation results have shown substantial improvement in the enhancement performance by the proposed method over some of recent methods both in visual quality and PSNR.

REFERENCES

[1]. D.L. Donoho, “Denoising by soft thresholding,” IEEE Trans. Inf. Theory, 1995, 41, pp.613–627.

[2] Xu Y, Weaver B, Healy D M, et al. ” Wavelet transform domain filters: A spatially selective noise filtration technique ”. IEEE Transactions on Image Processing, 1994, 3 (6) : 217～237

[3]. Donoho D L, John stone IM”. Ideal spatial adaptation via wavelet shrinkage . ” Biometrika, 1994, 81 (3) : 425～455.

[4]. DONOHO D L. ” De-Noising by Soft-Threshold. IEEE Transactions on Information Theory,” 1995, 41(3):613-627.

[5]. W. Sweldens, “The lifting scheme: a new philosophy in biorthogonal wavelet constructions,” Proc. SPIE-Int. Soc. Opt. Eng., 1995, 2569, pp.68–79

[6]. R.L. Claypoole, RG. Baraniuk, R.D. Nowak "Lifting Construction of Non-Linear Wavelet Transforms", IEEE -SP Symposium on Time-Frequency and Time-Scale Analysis TFTS-98, pp.49-52, Pittsburgh 1998.

[7]. R.L. Claypoole, RG. Baraniuk, R.D. Nowak "Adaptive Wavelet Transforms via Lifting," IEEE Conf. on Acoustics, Speech and Signal Processing, Phoenix, 1999.

[8]. A book on “Image processing MATLAB toolbox”.

[9]. “The lifting scheme: A construction of second generation wavelets” by wim sweldons. May 1995, Revised November 1996.To appear in SIAM Journal on Mathematical Analysis.

[10] Wenbing Fan, Zheng Ge, and Yao Wang, 2008. “Adaptive Wiener Filter based on Fast Lifting Wavelet Transform for Image Enhancement”, 7th World Congress on Intelligent Control and Automation (WCICA), pp. 3633 – 3636.

Performance Evaluation of Distribution Localization Techniques for Mobile under Water Sensor Networks

Mohan V & Mithun T P

Dept.of Telecommunication Engineering, R V College of Engineering, Bangalore-58, Karnataka, India.
E-mail : mhnv45@gmail.com, mithuntp@rvce.edu.in

***Abstract* -** Underwater Wireless Sensor Networks (UWSNs) are expected to provide variety of military and civilian applications. Sensed date can be interpreted meaningfully when referred to location of the sensor, making localization an important problem. In terrestrial WSNs, this can be achieved through a series of message exchanges (via RF communications) between each sensor and Global Positioning System(GPS) receivers. However, this is infeasible in UWSNs as GPs signals do not propagate through water.

In this paper a prediction-based localization scheme is proposed for mobile underwater sensor networks. We study a multi-stage AUV-aided localization technique for underwater wireless sensor networks (UWSN)s. We show that while improved performance with multiple stages is traded off with higher communication costs in general, the latter can be minimized while maintaining good performance with an appropriate choice of the acoustic communication range.

***Keywords* -** *Acoustic Communication, Autonomous Underwater Vehicle(AUV), Localization, Sensor Networks, Underwater Wireless Sensor Networks (UWSNs).*

I. INTRODUCTION

Sensor networks are becoming highly involved in our daily lives as they continuously collect data and monitor the surrounding environment. Raw sensor data are meaningful with the context, i.e., the knowledge of where and when the data is collected. This is known as data tagging. In addition, localization is required for node tracking, target detection and position-based routing algorithms. Sensor networks that operate outdoors are able to benefit from the GPS with some extra cost. Indoor, underground or underwater sensor networks need some specialized solutions for localization. Underwater Sensor Networks (USNs) can improve ocean exploration, allowing a list of new applications that are presently not possible or very costly to perform, including: oceanographic data collection, ecological applications (e.g. pollution, water quality and biological monitoring), public safety (e.g. disaster prevention, seismic and tsunami monitoring), military underwater surveillance, industrial (offshore exploration), etc. However, before USNs become commercially available or widely used, the networking of sensor nodes in underwater has to be addressed. Medium access and packet forwarding are still active research areas in USNs.Alternative GPS-less positioning schemes have been proposed for terrestrial sensor networks but they have to be revised due to acoustic channel properties. Acoustic communications is currently the most viable mode of wireless communications underwater. The acoustic channel has low bandwidth, high propagation delay and high bit error rate. Therefore, localization protocols need to work with minimum possible message exchange. This is also dictated by the limited battery power of the sensor nodes and the difficulty of recharging or replacing batteries of the underwater nodes. In Mobile Underwater Sensor Networks (MUSNs), the mobility of free-floating nodes brings up another challenge in localization. In this paper, we address the localization issue for MUSNs.

These unique characteristics pose severe challenges towards designing localization schemes that fulfil the following desirable qualities:

- **Accurate**- The location of the sensor for which sensed data is derived should be accurate and unambiguous for meaningful interpretation of data.
- **Fast**- Since nodes may drift due to water currents, the localization procedure should be fast so that it reports the actual location when data is sensed.
- **Wide Coverage**- The localization scheme should ensure that all nodes in the network can be localized.
- **Low Communication Costs**- Since the nodes are battery-powered and may be deployed for long durations, it should not waste energy unnecessary transmissions during the localization procedure.

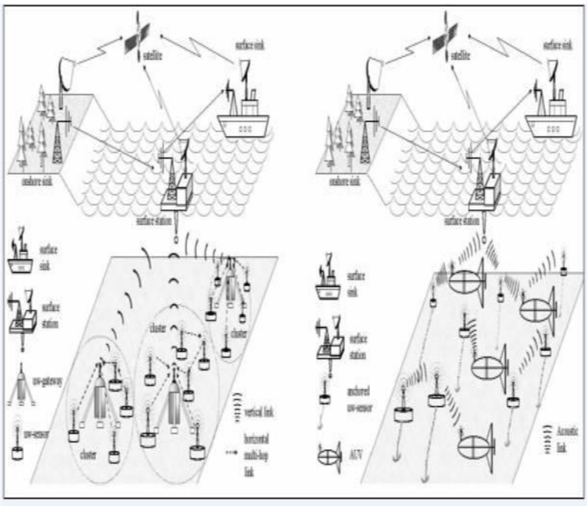


Fig. 1 : A 2-D (left) and 3-D (right) communications architecture for UWSNs [2].

II. RELATED WORK

Recently, a multitude of range-based localization schemes have been proposed specifically for UWSNs that rely on the presence of reference nodes with known position coordinates. Such reference nodes could be fixed (e.g., deployed on surface buoys or on the seabed) or mobile (e.g., AUVs). Each ordinary node estimates its distance from each reference node by ex-changing beacons with the reference nodes (or single reference node at various locations) and measuring the time (or time difference) of arrival. It then runs some localization algorithm(e.g., using multi-lateration or bounding box method) using the distance estimates, where d+1 independent measurements are needed to localize a d-dimensional space.

In the "AUV-Aided" localization technique proposed in [3],the sensor nodes can be dropped into the ocean and will move with the water currents while an AUV will traverse the UWSN periodically. The AUV obtains position updates by rising to the surface to use GPS, and then dives to a predefined depth and starts exchanging three types of messages with the ordinary nodes: wakeup, request and response. "wakeup" messages are sent by the AUV as it enters the network to declare its presence. Ordinary nodes that receive this message will respond with a "request" message to commence range measurement. "request/response" messages are exchanged between the AUV and ordinary nodes to estimate their positions according to the round trip time. This scheme does not assume any fixed infrastructure or time synchronization. In certain cases, simulation results show that 100% localization can be achieved with only 3% position error. However, the localization time required (up to 2 hours) and the message exchange phase to localize the nodes can be improved.

The "AUV-Using Directional Beacons" scheme (UDB)[4] is similar to [3] except for the following differences: (i) it proposes more accurate and efficient ways for localization based on simple calculations using directional instead of omni directional beaconing; and (ii) it reduces energy consumption by integrating "Silent Localization" [10] for the localization process. However, it takes more time to localize all the nodes using directional beacons because the AUV needs to traverse the network at least twice, and the impact of node mobility on its accuracy could be significant.

Instead of AUVs, the "Dive'N'Rise" (DNR)[11] localization scheme uses mobile beacons whose diving/rising is controlled by a weight/bladder mechanism. These beacons update their positions at the surface, and broadcast them when they dive to a certain depth. This is a low-cost scheme that can localize 100% of the nodes with relative small positioning error and can reduce communication costs and energy using "Silent Localization". However, the scheme uses 25 DNR beacons for 1km × 1km × 1km underwater column, which is extremely expensive because it requires 25 GPS and 25 moving devices. Moreover, under actual operating conditions, 909the DNR beacons will be strongly affected by the surface currents, which will degrade the localization accuracy.

In [12], the authors present a multi-stage enhancement to DNR − termed "Multi-stage DNR". When an ordinary node receives at least three messages from the mobile reference at non-collinear locations, it computes its own location. After that, it becomes a reference node and helps to localize the remaining ordinary nodes, provided it lies below the maximum dive depth of DNR mobile beacons.

III. MULTI-STAGE AUV-AIDED UNDERWATER LOCALIZATION

In this section, we describe our proposed Multi-stage AUV-aided localization technique for UWSNs, aimed at improving the "Multi-stage AUV" scheme. This expands the coverage of the mobile beacon in the first stage while utilizing the multi-stage concept to localize the remaining (un-localized) nodes. We consider an UWSN that comprises an AUV and ordinary nodes that dive to a known depth (provided by pressure sensors) and remain static (by fixing with anchors) during the localization process. Moreover, all nodes can communicate (omni-directionally) with the AUV and other nodes by sending or receiving acoustic signals. The AUV can surface to obtain its coordinates using GPS signals, and can be pre-programmed to dive to a given depth (provided by pressure sensors) and traverse a given path. As with the ordinary nodes, the AUV is equipped with an omni-directional antenna and communicate with nodes via acoustic signals. We assume that the AUV as well as the ordinary nodes are all time synchronized.

A. Procedure for AUV

We begin by describing the procedure for the AUV, as illustrated in Fig. 1. Initially, the AUV floats on the surface and can obtain its coordinates from GPS. After that, it will dive to a pre-programmed depth with the help of a pressure sensor and start traversing the sensor network following a pre-programmed path. However, its actual path will deviate due to underwater currents. To correct for / minimize its positional errors, the AUV can (i) surface periodically to obtain GPS updates; or (ii) be equipped with high precision navigation tools (e.g., Doppler Velocity Log) that limits this deviation to an acceptable level. In this study, we assume that the AUV (i) follows a sinusoidal path in the X-Y plane; (ii) is subject to underwater currents in the Y direction; and (iii) is equipped with navigation tools to limit its positional error to 5m.

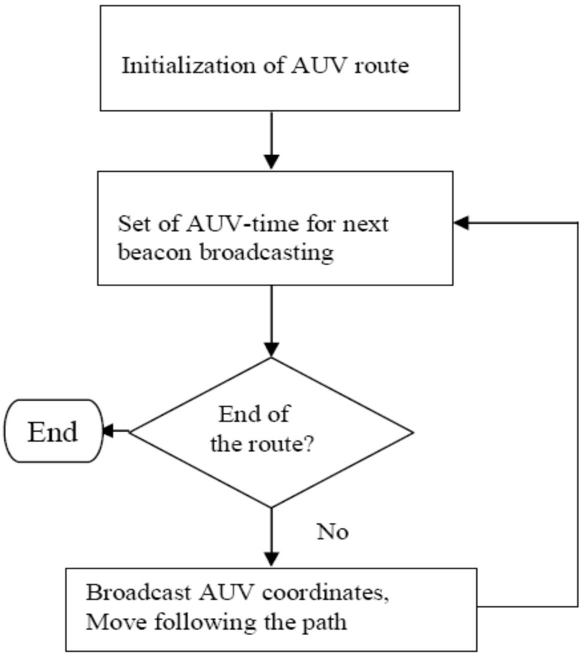

Initialization of AUV route
Set of AUV-time for next beacon broadcasting
End of the route?
End
No
Broadcast AUV coordinates, Move following the path

Fig. 2. Procedure for AUV in Multi-stage, AUV-aided Localization.

The three-stage message exchange (wakeup, request and response) between ordinary nodes and the AUV proposed in the "AUV-Aided" localization technique [2] incurs high energy consumption, which is undesirable as it is difficult to replenish the power source in UWSNs once they are deployed. We propose to reduce this energy consumption by applying some concepts of "Silent Positioning" as proposed in [4], where ordinary nodes remain silent and do not need to transmit at all during the first stage We define the beacon structure to comprise three components: time stamp, coordinates and identifier. The time stamp indicates the time the beacon is created at the AUV and is used by the ordinary node to estimate its distance from the AUV using the Time of Arrival approach (ToA). The second component comprises the coordinates of the AUV at the instant of beaconing, which together with the distance estimates, are used to estimate the ordinary node's coordinates. Finally, the identifier, ID, indicates if the beacon originated from the AUV (ID = 1) or from a localized ordinary node (ID = 0).

B. Procedure for Ordinary Nodes

Initially, each ordinary node sets a beacon counter, m to 0,initializes timer, t, set its status to "unlocalized" (Node stat= UNLOC) and set the variable IDsum to 0. Each ordinary node at position (x,y) listens to the beacons broadcasted by the AUV. When it receives a message transmitted at t1 when the AUV is at location (x1, y1), it updates m and estimates its distance, d1, from the AUV using the average speed of sound underwater and the difference in the arrival time and time stamp. It then stores (˜ x1,˜ y1), which are the AUV's estimated coordinates, the identifier as well as the estimated distance d1.

This is repeated when it receives the next message transmitted at t2 and so on, as illustrated in Figure 2 until it receives three beacons from non-collinear locations. The node can then solve for (x,y) using triangulation. As stated before, z and zi are known via pressure sensors.Using (x,y,z) and the stored coordinates (˜xi, ˜yi, zi), we can recompute the distance between the AUV and the node,

˜di, i=1,2,3 as follows:

˜d2i =(x − ˜xi)2+(y − ˜yi)2+(z − zi)2

Then, the condition for the node to be localized and become a reference node is given as follows:

maxi=1:3 |di − ˜ di|≤ ε. (1)

When a node becomes localized (Node stat =LOC),it updates the variable IDsum by summing up the ID of the received beacons used to estimate its location to determine if it was AUV-localized (IDsum = 3 in this case). After a certain time-out (for the AUV to complete its beacon broadcast), all reference nodes will broadcast their (estimated) coordinates once from which (some of) the remaining ordinary nodes can localize themselves using the same technique as before (k=1). In this case, only the AUV-localized nodes will broadcast once.

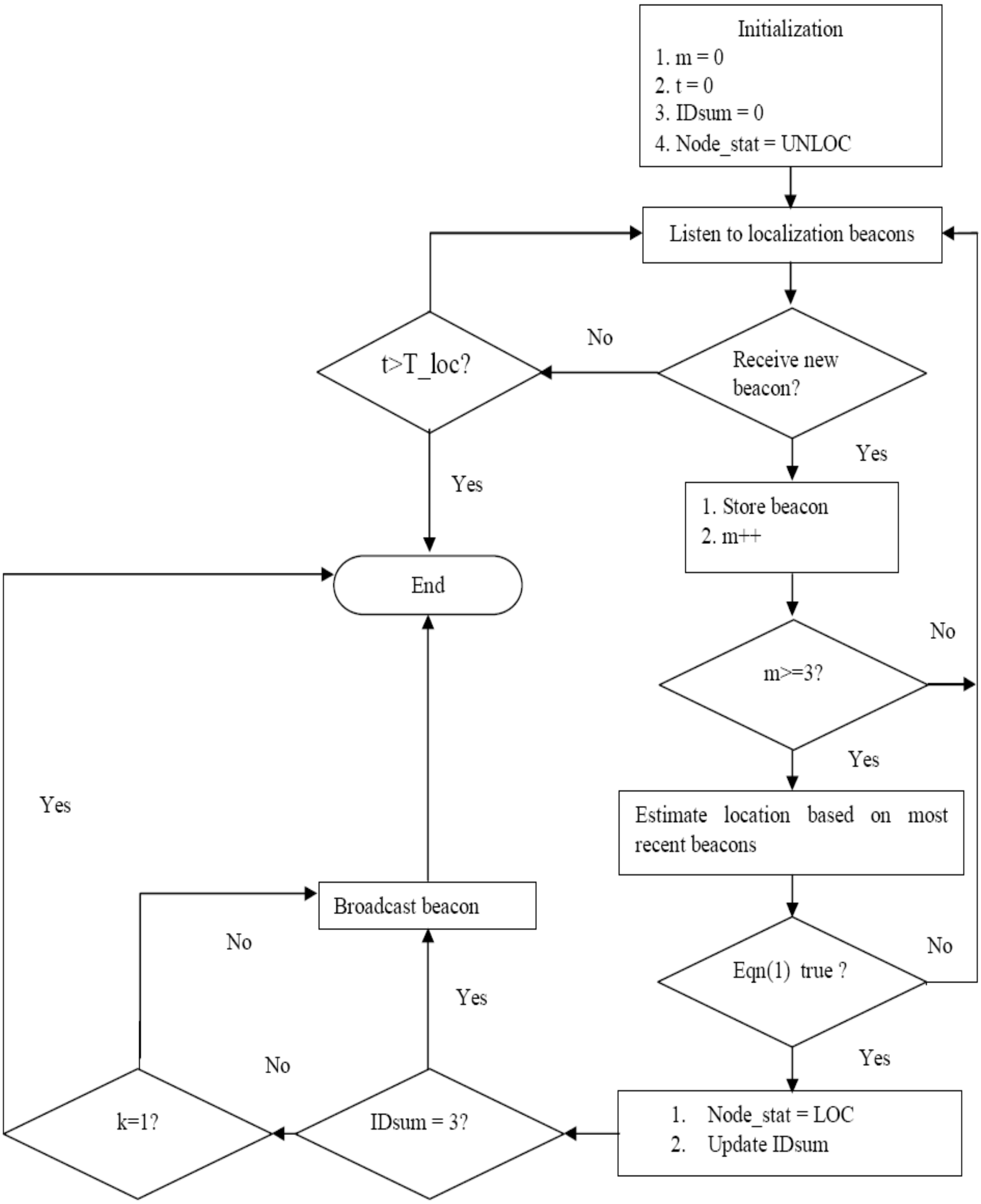


Fig. 3 : Procedure for ordinary node in Multi-stage, AUV-aided Localization

This procedure can be repeated (k>1) until no more ordinary node becomes localized to improve the coverage, albeit at the expense of higher communication costs, as shown in Figure 3. Figure 4 illustrates the interaction between the AUV and the ordinary nodes.

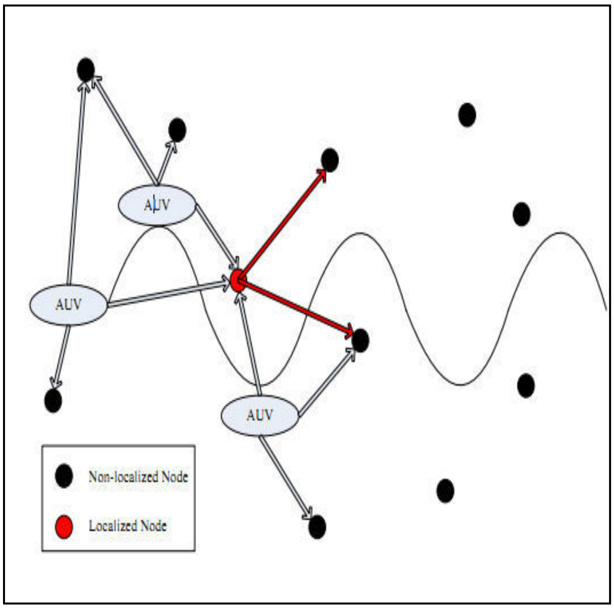
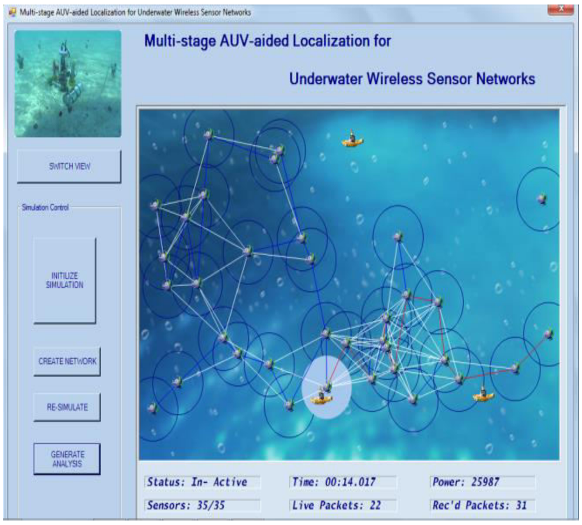


Fig. 4 :Interaction between the AUV and sensor nodes.

IV. SIMULATION RESULTS

We validate our proposed localization scheme using the Microsoft Visual Studio simulator [5]. We deploy between 100 to 200 ordinary nodes (in increments of 10) randomly in a 1000m × 1000m × 100m water volume, and we assume that they remain in a fixed position throughout the simulation duration. The AUV traverses the UWSN at a constant depth and constant speed (5m/s), following its sinusoidal path within a relative small error (up to 5m) in the Y direction due to underwater current. The communication range for the AUV and nodes varies between 50m, 130m and 300m. We use CSMA as the medium access control protocol and model the acoustic channel according to [9].We evaluate and compare the performance of our proposed algorithm for k =1and k> 1 according to three criterion: i) coverage, described as the ratio of the localized nodes to the total number of nodes; ii) localization error, described as the average Euclidean distance between the estimated and real location of each node; and iii) communication costs, described as the total number of messages sent by the nodes, excluding the messages sent by the AUV. The energy consumption to localize a UWSN is directly linked with the total number of beacons sent by the nodes. The UWSN simulator and the coverage analysis of UWSN localization is shown in Figure 5 and Figure 6.


Multi-stage AUV-aided Localization for Underwater Wireless Sensor Networks

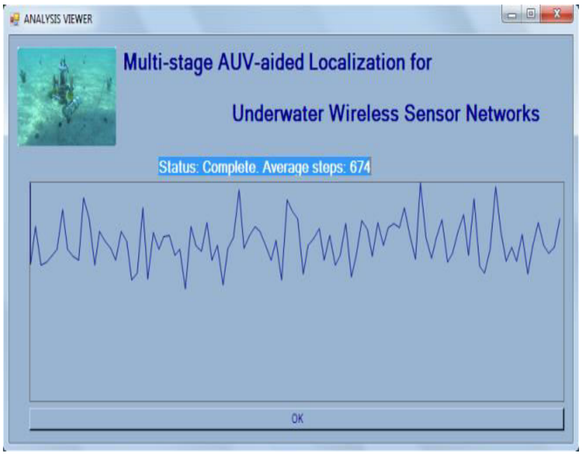
Multi-stage AUV-aided Localization for

Underwater Wireless Sensor Networks

Status: In- Active | Time: 00:14.017 | Power: 25987

Sensors: 35/35 | Live Packets: 22 | Rec'd Packets: 31

Fig. 5: Simulator Interface for UWSN localization.


ANALYSIS VIEWER

Multi-stage AUV-aided Localization for

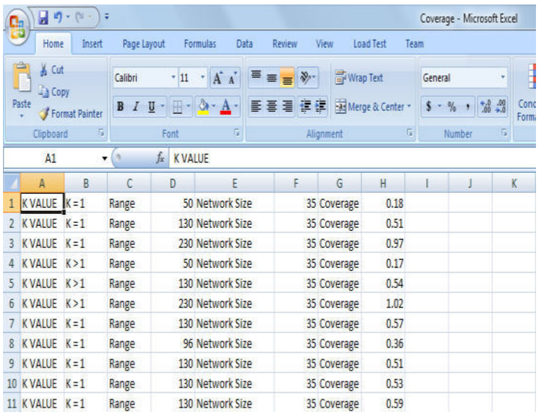
Underwater Wireless Sensor Networks

Status: Complete. Average steps: 674

Fig. 6: Coverage Analysis of UWSN localization.

A. Coverage

In Figure 7, we compare the coverage achieved for k =1 and k> 1. As expected, the coverage for k> 1 is clearly higher than k = 1, with a gain of up to 30%. However, the achievable coverage is still quite low (up to 70%) with a short communication range (50m). At a higher communication range of 230m, the coverage remains over 90% regardless of the network size, and the difference in performance between k =1and k> 1 becomes marginal. This is due to the fact that the AUV and the first set of reference nodes can localize almost the entire UWSN with acoustic range of 300m and thus, extending the multi-stage to k> 1 does not increase the performance as significantly as with smaller communication ranges.



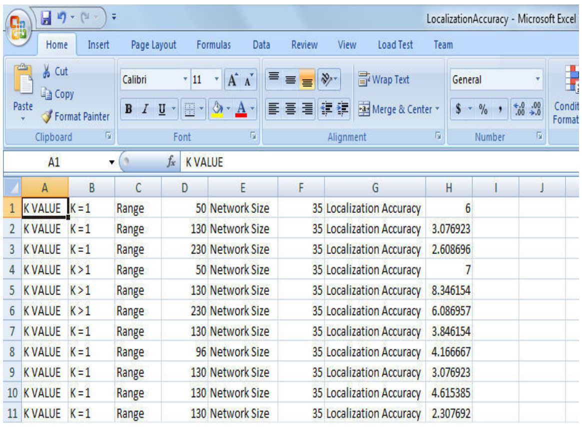
	A	B	C	D	E	F	G	H
1	K VALUE	K=1	Range	50	Network Size	35	Coverage	0.18
2	K VALUE	K=1	Range	130	Network Size	35	Coverage	0.51
3	K VALUE	K=1	Range	230	Network Size	35	Coverage	0.97
4	K VALUE	K>1	Range	50	Network Size	35	Coverage	0.17
5	K VALUE	K>1	Range	130	Network Size	35	Coverage	0.54
6	K VALUE	K>1	Range	230	Network Size	35	Coverage	1.02
7	K VALUE	K=1	Range	130	Network Size	35	Coverage	0.57
8	K VALUE	K=1	Range	96	Network Size	35	Coverage	0.36
9	K VALUE	K=1	Range	130	Network Size	35	Coverage	0.51
10	K VALUE	K=1	Range	130	Network Size	35	Coverage	0.53
11	K VALUE	K=1	Range	130	Network Size	35	Coverage	0.59

Fig. 7: Coverage achieved with k> 1and k =1

B. Localization Accuracy

We show the mean localization error in the estimated positions of the localized nodes for k =1and k> 1in Figure 8. For k = 1, we can observe that, regardless of the communication range, the mean error in the nodes' position is approximately 6-7 meters. This is expected because all the localized nodes get their coordinates from the AUV (with a maximum position deviation of 5m) or from nodes that were localized using the AUV positions (which contributes cumulatively to the position error).

For k> 1, the mean localization error increases considerably especially for the lower acoustic range of 50m and 130m. This is due to the fact that the AUV and the first stage of reference nodes do not provide sufficient coverage to localize all nodes. Hence, the remaining ordinary nodes will use coordinates from other reference nodes and once their position is estimated, they will help to localize other nodes by broadcasting their coordinates, which will propagate the location error.

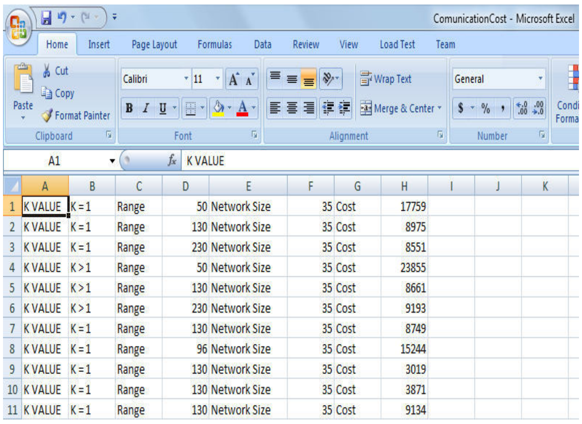


	A	B	C	D	E	F	G	H
1	K VALUE	K=1	Range	50	Network Size	35	Localization Accuracy	6
2	K VALUE	K=1	Range	130	Network Size	35	Localization Accuracy	3.076923
3	K VALUE	K=1	Range	230	Network Size	35	Localization Accuracy	2.608696
4	K VALUE	K>1	Range	50	Network Size	35	Localization Accuracy	7
5	K VALUE	K>1	Range	130	Network Size	35	Localization Accuracy	8.346154
6	K VALUE	K>1	Range	230	Network Size	35	Localization Accuracy	6.086957
7	K VALUE	K=1	Range	130	Network Size	35	Localization Accuracy	3.846154
8	K VALUE	K=1	Range	96	Network Size	35	Localization Accuracy	4.166667
9	K VALUE	K=1	Range	130	Network Size	35	Localization Accuracy	3.076923
10	K VALUE	K=1	Range	130	Network Size	35	Localization Accuracy	4.615385
11	K VALUE	K=1	Range	130	Network Size	35	Localization Accuracy	2.307692

Fig. 8 : Localization accuracy achieved with k> 1 and k =1.

C. Communication Costs

The third measure of our simulations is the communication costs, whose results are represented in Figure 9. These results are directly related to the coverage performances because of the fact that a node will not broadcast any beacon as long as it has not been localized. The communication costs increase linearly as we increase the total number of nodes in our UWSN. This linearity is explained by the fact that each sensor node is limited to broadcast only one beacon once localized and because the coverage ratio stays stable while increasing the total number of nodes. Finally, if the augmentation of the costs between k =1and k> 1 for 50m and 130m ranges results in higher coverage, in the case of a 300m range, the wastage of energy with k> 1 is less justifiable as the gain in coverage is less insignificant.



	A	B	C	D	E	F	G	H
1	K VALUE	K=1	Range	50	Network Size	35	Cost	17759
2	K VALUE	K=1	Range	130	Network Size	35	Cost	8975
3	K VALUE	K=1	Range	230	Network Size	35	Cost	8551
4	K VALUE	K>1	Range	50	Network Size	35	Cost	23855
5	K VALUE	K>1	Range	130	Network Size	35	Cost	8661
6	K VALUE	K>1	Range	230	Network Size	35	Cost	9193
7	K VALUE	K=1	Range	130	Network Size	35	Cost	8749
8	K VALUE	K=1	Range	96	Network Size	35	Cost	15244
9	K VALUE	K=1	Range	130	Network Size	35	Cost	3019
10	K VALUE	K=1	Range	130	Network Size	35	Cost	3871
11	K VALUE	K=1	Range	130	Network Size	35	Cost	9134

Fig. 9: Communication costs with with k> 1and k =1.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a multi-stage AUV-aided localization technique for underwater wireless sensor networks(UWSN)s. The proposed method combines the flexibility and localization accuracy of an AUV-aided localization, the energy efficiency of "Silent Localization" and improved localization coverage with k-stage localization based on sensor nodes. We evaluated our proposed method for k =1and k> 1 with three different values of acoustic transmission ranges: 50m, 130m and 300m according to three criterions: coverage, accuracy and communication costs for a network comprising 100 to 200 fixed nodes deployed randomly in an underwater column measuring 1000m × 1000m × 100m. With k> 1, the localization process by the (non-AUV localize) reference nodes continue until no new ordinary node can be localized. The whole localization process can be completed in less than 10 minutes with approximately 7 meters error in the positioning and can cover more than 95% of the whole network. In addition, we observe that with the lower acoustic ranges, the increase in coverage with k> 1 is achieved at the expense of higher localization error and communication costs compared to k = 1.

However, with a 300m acoustic communications range, additional stages do not achieve a significant gain in terms of coverage and accuracy while incurring higher communication costs. Basically, this indicates that by employing an acoustic antenna with a reasonably long range (300m), a single-level of multi-stage is sufficient to achieve the best coverage and localization accuracy while preventing wastage of energy by broadcasting beacons unnecessarily.

For future work, we plan to (i) extend our scheme for three-dimensional localization; (ii) consider more realistic mobility current effects such as the meandering current mobility model10]; and (iii) compare various AUV paths in evaluating our proposed scheme. In the long term, we hope to implement and evaluate the proposed scheme in an actual underwater environment.

VI. ACKNOWLEDGEMENT

We would first like to extend our gratitude to T.P.Mithun, lecturer, for his constant guidance and encouragement throughout the endeavour. His support, motivation and innovative ideas helped us tremendously whenever we met a roadblock. We acknowledge the support from the whole of our Telecommunication Department and our head of department Dr H V Kumarswamy for his invaluable suggestions and help.

REFERENCES

[1] Kai Chen, Yi Zhou, Jianhua He ,"A Localization Scheme for Underwater Wireless Sensor Networks"

[2] M. Erol, L. F. M. Vieira, and M. Gerla, "AUV-Aided Localization for Underwater Sensor Networks," Proc. of the WASA, pp. 44–51, August 2007.

[3] Vijay Chandrasekhar, Winston KG Seah, Yoo Sang Choo, How Voon Ee "Localization in Underwater Sensor Networks — Survey and Challenges".

[4] Pavlos Papageorgiou, "Literature Survey on Wireless Sensor Networks", July 2003

[5] "Microsoft Visual Studio , programmer's guide," Scalable Network Technologies Inc,

[6] A. Caruso, F. Paparella, L. F. M. Vieira, M. Erol, and M. Gerla, "The meandering current mobility model and its impact on Underwater Mobile Sensor Networks," Proc. of the IEEE INFOCOM, pp. 772–779,April 2008.

[7] H.Luo,Y.Zhao,Z.Guo,S.Liu,P.Chen,andL.M.Ni, "UDB:UsingDirectional Beacons for Localization in Underwater Sensor Networks," Proc. of the ICPADS, pp. 551–558, December 2008.

[8] M. Stojanovic, "On the relationship between capacity and distance in an underwater acoustic communication channel," Proc. of the ACM WUWNet, September 2006

[9] Zaihan Jiang, "Underwater Acoustic Networks – Issues and Solutions", International journal of intelligent control and systems, Vol.13, No 13, September-2008

[10] X. Cheng, H. Shu, Q. Liang, and D. Du, "Silent Positioning in Underwater Acoustic Sensor Networks," IEEE Trans. Veh. Technol.,vol.57, no. 3, pp. 1756–1766, May 2008.

[11] M. Erol, L. F. M. Vieira, and M. Gerla, "Localization with Dive'N'Rise (DNR) Beacons for Underwater Acoustic Sensor Networks," Proc. Of the WUWNet, pp. 97–100, September 2007.

[12] M. Erol, L. F. M. Vieira, A. Caruso, F. Paparella, M. Gerla, and S. Oktug, "Multi Stage Underwater Sensor Localization using Mobile Beacons," Proc. of the IEEE SENSORCOMM, pp. 710–714, August 2008.

◈◈◈

Exypnos Office

Maruthi Naik R.K & M. Nagaraja

Dept of E&C, Sri Siddhartha University, Sri Siddhartha institute of Technology, Tumkur, Karnataka, India
E-mail : Maruti.group1@gmail.com

Abstract - The project is aimed to automate the office environment. The Office systems are among the newest and most rapidly expanding computer based information systems. The backbone of office automation is the LAN which allows users to transmit data. The project identifies the employee entering the office greets the user with his name and the login time , intimates the schedules for the day via SMS, automatically starts the computer for the employee, turns-off the monitor at the employee absence and will turn on the monitor on the employee presence. The attendance for all employees is done automatically. For the new users the photograph can be taken using web cam. All these activities mentioned have been made very simple and effective by the use of computers and Sensors

No longer are a sharp pencil and a legal pad enough to make you competitive as a business organization. As computers have infiltrated and entrenched themselves in the world of big business so also have they trickled down to companies and made themselves a necessity in every business setting.wireless communication technology, Internet technology, and standardized transducer interfaces are serving to shape the landscape of next-generation's remote distributed control system. The paper proposes originally to build the DMCS integration use of GPRS and Internet technology. Based on the analysis of function demands of the market for the measurement and control system, the thesis makes sure the function goals of the system, builds framework structure of it and determines the application model of the system and software architecture, which is the combination of the C/S structure and B/S structure.

Keywords - *GPRS(general packet radio service), JDBC(java data base connectivity), AWT(abstract window tool kit), C/S(client/server),B/S(webbandwidth/server).*

I. INTRODUCTION

Data transmission through wireless and mobile networks became very attractive for many business agents, in the purpose of promoting and offering electronic services to their clients and also for remote process control. This paper presents a client server wireless system, which offers to the mobile users the possibility of GPRS and WLAN access to electronic services, through mobile data terminals like SMARTPHONES and PDAs. As a representative example of business environment electronic service, installed on the server, a stock exchange e-service was realized. The client server functionality can be extended for a larger number of information services and applications: e-commerce, mobile-banking, e-health, e-learning, e-government, but also for processes automation and remote control, etc. An informational service is a software application, whose main component is a database, which can be accessed by the users via Internet or mobile communication networks. The data transmission through mobile telephony networks gives to the users the mobility advantage. The 3G networks offers significant channel bandwidth for services access and providing and processes control from the distance.

The remote distributed control system， based on GPRS and Internet， is a new concept. In the past few years, this concept has become a hot research topic in the field of industrial control. This new control system, supporting open standards for IEEE 1451， is the standardization of control systems. Universal Smart Transducer Interface Module (STIM), open network communications (TCP / IP) and GPRS wireless communication technology make this control system the next generation of remote distributed control system prototype. This paper describes the design of the overall program based on GPRS and Internet-based remote distributed control system.

1.1 Objective

To design and develop a user friendly technology which can be use of maximum to a particular place using the latest technology in the real time environment. Here we are using multiple technologies for a single application.

1.2 Motivation

Wanted to work on a project which can use multiple technologies, as it will give me an Opportunity not only to learn them but also to implement the same in the real time environment.

1.3 Problem statement

To design a product which can boot from remote place , send SMS to the user with the works to be done and use the symbian / android technology in the cell phone to put them into the silence mode and also to use the wireless technology to save the power problems.

1.4 Solution to the problem

The project is aimed to automate the office environment. The Office systems are among the newest and most rapidly expanding computer based information systems. The backbone of office automation is the LAN which allows users to transmit data. The project identifies the employee entering the office greets the user with his name and the login time , intimates the schedules for the day via SMS, automatically starts the computer for the employee, turns-off the monitor at the employee absence and will turn on the monitor on the employee presence. The attendance for all employees is done automatically. For the new users the photograph can be taken using web cam. All these activities mentioned have been made very simple and effective by the use of computers and Sensors.

1.5 Methodology

We design the micro controller and interface the same to the RFID and the PC . Develop the PC to work as master and slave in the real time. To design the wireless technology using the RF and develop the software in the micro controller and the PC . Finally to club all the same and make it work in the real time environment.

II. LITERATURESURVEY

Typically, the system is to achieve a common goal by interrelated set of components. For the control system, this common goal is monitoring and controlling the production of industrial process, and making the whole system of input and output to maintain the inherent relationship in unusual circumstances. The input is treated with the function module, analysis with the information module and communication module, judged by behavior module. The remote distributed control system, which based on GPRS and Internet, is not just the traditional control system for easy replacement, but using the full range of advanced technologies to enhance system functionality and promote the development of control systems to meet the increasing demands of control system. The features of the new control system as shown below:

- Able to network-based measure the key parameters of industrial processes automatically.
- Able to do network-based control subjects.
- Able to achieve online programming and sensor node controlling.
- User-friendly design interface.

Data transmission through wireless and mobile networks became very attractive for many business agents, in the purpose of promoting and offering electronic services to their clients and also for remote process control. This paper presents a client server wireless system which offers to the mobile users the possibility of GPRS and WLAN access to electronic services, through mobile data terminals like SMARTPHONES and PDAs. As a representative example of business environment electronic service, installed on the server, a stock exchange e-service was realized. The client server functionality can be extended for a larger number of information services and applications: e-commerce, mobile-banking, e-health, e-learning, e-government, but also for processes automation and remote control, etc. An informational service is a software application, whose main component is a database, which can be accessed by the users via Internet or mobile communication networks. The data transmission through mobile telephony networks gives to the users the mobility advantage. The 3G networks offers significant channel bandwidth for services access and providing and processes control from the distance.

III. SYSTEM DESIGN

- Design the RFID
- Deign the micro controller board
- Develop the program
- Develop the program in PC to send SMS
- Develop program in the cell phone
- Design the PCB for all the above
- Design the wireless RF TX and RF RX boards
- Integrate the whole system
- Final testing.

The employee enters, dips his RF-card containing the user identification into the RF- reader the information is read using the Communication API's .Using the identification the users schedules are collected and sent as a SMS to the Mobile using JSMS Engine API.

Java socket program is used to turn on the computer automatically as the employee enters. The employee is greeted using the Java Speech API as he logs-in. The employee presence is detected using the TACTILE SWITCHES as sensors which are fixed to the employee chairs. The data transition of sensor to the client computer is done via RF Transmitter (AKS-433 MHZ)

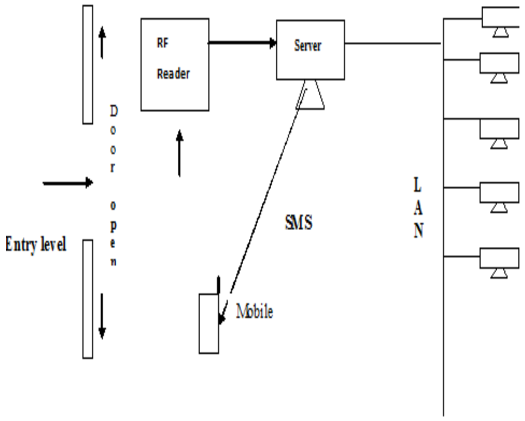


Fig.1: Block Diagram

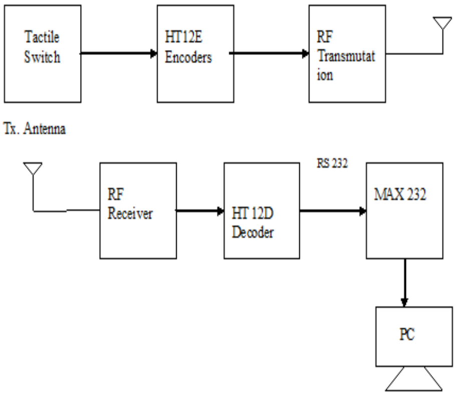


Fig. 2: Client's Block Diagram

IV. STEPPER MOTOR

In this project we are using the stepper motor as per the specification mentioned below. The stepper motor is a 4 pole where in this poles are connected to the relay and is controlled by the relay driver IC ULN 2003. At any given instant of time only 1 relay is activated such that depending upon the particular relay activated that particular pole in stepper motor is energized and accordingly the stepper motor moves to that particular pole due to excitation. Depending on the type of relays to be activated and the particular order, stepper motor accordingly moves in either clock wise or anti-clock wise direction

Stepper Motor Specification :

1.	Step Angle	0.5°
2.	Step angle accuracy	5%
3.	Rate phase current	0.22 A
4.	Phase resistance	23 Ω
5.	Phase inductance	30 mH
6.	Holding torque	20 Ncm
7.	Detent torque	2 Ncm
8.	Rotor inertia	70grcm^2
9.	Weight	0.2 kg
10.	Insulation	class B
11.	Voltage	+12V.

V. RFID READER:

An **RFID** reader is a device that is used to interrogate an RFID tag. The reader has an antenna that emits radio waves; the tag responds by sending back its data. A number of factors can affect the distance at which a tag can be read (the read range). The frequency used for identification, the antenna gain, the orientation and polarization of the reader antenna and the transponder antenna, as well as the placement of the tag on the object to be identified will all have an impact on the RFID system's read range.

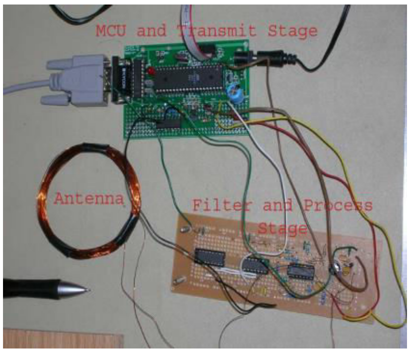


Fig. 3: RFID READER

Radio-frequency identification (**RFID**) is a technology that uses radio waves to transfer data from an electronic tag, called RFID tag or label, attached to an object, through a reader for the purpose of identifying and tracking the object. Some RFID tags can be read from several meters away and beyond the line of sight of the reader. The application of bulk reading enables an almost-parallel reading of tags. The tag's information is stored electronically. The RFID tag includes a small RF transmitter and receiver. An RFID reader transmits an encoded radio signal to interrogate the tag. The tag receives the message and responds with its identification information.

Many RFID tags do not use a battery. Instead, the tag uses the radio energy transmitted by the reader as its energy source. The RFID system design includes a method of discriminating several tags that might be within the range of the RFID reader.

A number of organizations have set standards for RFID, including the International Organization for Standardization (ISO), the International Electro technical Commission (IEC). There are also several specific industries that have set guidelines including the Financial Services Technology Consortium (FSTC) has set a standard for tracking IT Assets with RFID, the Computer Technology Industry Association CompTIA has set a standard for certifying RFID engineers and the International Airlines Transport Association IATA set tagging guidelines for luggage in airports.

RFID can be used in many applications. A tag can be affixed to any object and used to track and manage inventory, assets, people, etc. For example, it can be affixed to cars, computer equipment, books, mobile phones, etc. The Healthcare industry has used RFID to reduce counting, looking for things and auditing items. Many financial institutions use RFID to track key assets and automate compliance. Also with recent advances in social media RFID is being used to tie the physical world with the virtual world.

RFID is a superior and more efficient way of identifying objects than manual system or use of bar code systems that have been in use since the 1970s. Furthermore, passive RFID tags (those without a battery) can be read if passed within close enough proximity to an RFID reader. It is not necessary to "show" the tag to the reader device, as with a bar code. In other words it does not require line of sight to "see" an RFID tag, the tag can be read inside a case, carton, box or other container, and unlike barcodes RFID tags can be read hundreds at a time. Bar codes can only be read one at a time.

RF modules are normally divided into three groups, RF transmitter module, RF receiver module and RF transceiver module. Transmitter module is an electronic component using a variety of radio signals to remote control the target device which has a receiver module built-in. The remote distance can be very long and you don't need a line-of-sight remote controlling compared to remote controls using infrared technology. And RF modules are widely used in garage door openers, wireless alarm systems, industrial remote controls and wireless home automation systems. The RF module, as the name suggests, operates at Radio Frequency. The corresponding frequency range varies between 30 kHz & 300 GHz. In this RF system.

VI. SERVER SYSTEM

In the context of client-server architecture, a **server** is a computer program running to serve the requests of other programs, the "clients". Thus, the "server" performs some computational task on behalf of "clients". The clients either run on the same computer or connect through the network.

In most common use, **server** is a physical computer (a hardware system) dedicated to running one or more such services (as a host), to serve the needs of users of the other computers on the network. Depending on the computing service that it offers it could be a database server, file server, mail server, print server, web server, or other. In the context of Internet Protocol (IP) networking, a **server** is a program that operates as a socket listener. Servers often provide essential services across a network, either to private users inside a large organization or to public users via the Internet. For example, when you enter a query in a search engine, the query is sent from your computer over the internet to the servers that store all the relevant web pages. The results are sent back by the server to your computer The term *server* is used quite broadly in information technology. Despite the many server-branded products available (such as server versions of hardware, software or operating systems), in theory any computerized process that shares a resource to one or more client processes is a server. To illustrate this, take the common example of file sharing. While the existence of files on a machine does not classify it as a server, the mechanism which shares these files to clients by the operating system is the server.

DECODER & ENCODER

An **encoder** is a device, that converts information from one format or code to another, for the purposes of standardization, speed, secrecy, security, or saving space by shrinking size.

A **decoder** is a device which does the reverse operation of an encoder, undoing the encoding so that the original information can be retrieved. The same method used to encode is usually just reversed in order to decode. It is a combinational circuit that converts binary information from n input lines to a maximum of 2^n unique output lines.

In digital electronics, a decoder can take the form of a multiple-input, multiple-output logic circuit that converts coded inputs into coded outputs, where the input and output codes are different. e.g. n-to-2^n, binary-coded decimal decoders. Enable inputs must be on for the decoder to function, otherwise its outputs assume a single "disabled" output code word. Decoding is necessary in applications such as data multiplexing, 7 segment display and memory address decoding.

INTERFACE UNIT

RS-232 (Recommended Standard 232) is the traditional name for a series of standards for serial binary single-ended data and control signals connecting between a *DTE* (Data Terminal Equipment) and a *DCE* (Data Circuit-terminating Equipment). It is commonly used in computer serial ports. The standard defines the electrical characteristics and timing of signals, the meaning of signals, and the physical size and pin out of connectors. The current version of the standard is TIA-232-F Interface Between Data Terminal Equipment and Data Circuit-Terminating Equipment.

The MAX232 from Maxim was the first IC which in one package contains the necessary drivers (two) and receivers (also two), to adapt the RS-232 signal voltage levels to TTL logic. It became popular, because it just needs one voltage (+5V) and generates the necessary RS-232 voltage levels (approx. -10V and +10V) internally. This greatly simplified the design of circuitry. Circuitry designers no longer need to design and build a power supply with three voltages (e.g. -12V, +5V, and +12V), but could just provide one +5V power supply.

VII. ADVANTAGES & LIMITATIONS

- An affordable Technology with high end advantage
- Time consumption is reduced
- Automatic Authentication
- Paper work reduced
- Human errors are minimized

Limitations:

- Technical knowledge is a must to operate.
- Depends on the network

APPLICATIONS

- Used in Govt. offices
- Used in companies
- Used in call centers
- Used in customer support offices

VIII. CONCLUSION

The data transmission through wireless and mobile networks became very attractive for many business purpose offering electronic services to their clients for remote process control this presents client server wireless system through LAN. The client server functionally can be extended for larger number of information services and application the data transmission through mobile telephony networks use the user mobility advantages.

This project can provide automate of office environment . the office system are among the newest and most rapidly expanding computer based information all these activities mentioned have been made very simple and effective by the use of computer and sensors.

ACKNOWLEDGEMENT

I am deeply indebted and I would like to express my sincere thanks to our beloved principal Dr.K.A.Krishnamurthy, for providing me on opportunity to do this project.

My special gratitude and sincere thanks to my guide Dr.M.Nagaraja asst.proffessor, Department of E&C,SSIT,TUMKUR for his guidance ,constant encouragement and whole hearted support.

REFERENCES

[1] Mo Guan Guangije Han,Hai Zhao, the embedded internet technology based on a real time kernel for non pc devices, networking, sensing and control,2004 IEEE international conference on volume1,march 2004.

[2] Wei Zhang,Branicky M.S,Philips SM stability of networked control system.IEEE control systems Magazine,2001, 21(1):8499

[3] Rovetta A,sala R.Remote control in telerobotic suegery.IEEEtrascaction on system,1996, 26(4):438-444

[4] Freeman RL., telecommunication transmission hand book,John Willey,1975

[5] Jakes,W,C. microwave mobile communications, John Willey 1975[9] Ziemer,R.E,Peterson R.,L., Digital communications and spread spectrum systems, Newyork, Macmillian,1985

[6] Hirsch F.,& kemp, J."mobile web services",john willey&sons, ltd.,2006

◈◈◈

Provably Secure and Higher Efficient Auditing for Data Storage Security in Cloud Computing

Maragatham. S[1] & Ravikumar. P[2]

Department of ISE[1], Department of CSE[2]
[1&2]SJBIT Institute of Engineering
Email: smgmaragatham@gmail.com, ravikumar.pe@gmail.com

Abstract - Cloud Computing is the long dreamed vision of computing as a utility, where users can remotely store their data into the cloud so as to enjoy the on-demand high quality applications and services from a shared pool of configurable computing resources. By data outsourcing, users can be relieved from the burden of local data storage and maintenance. However, the fact that users no longer have physical possession of the possibly large size of outsourced data makes the data integrity protection in Cloud Computing a very challenging and potentially formidable task, especially for users with constrained computing resources and capabilities. Thus, enabling public auditability for cloud data storage security is of critical importance so that users can resort to an external audit party to check the integrity of outsourced data when needed. To securely introduce an effective third party auditor (TPA), the following two fundamental requirements have to be met: 1) TPA should be able to efficiently audit the cloud data storage without demanding the local copy of data, and introduce no additional on-line burden to the cloud user; 2) The third party auditing process should bring in no new vulnerabilities towards user data privacy. In this paper, we utilize and uniquely combine the public key based homomorphic authenticator with random masking to achieve the privacy-preserving public cloud data auditing system, which meets all above requirements. To support efficient handling of multiple auditing tasks, we further explore the technique of bilinear aggregate signature to extend our main result into a multi-user setting, where TPA can perform multiple auditing tasks simultaneously. Extensive security and performance analysis shows the proposed schemes are provably secure and highly efficient.

I. INTRODUCTION

Cloud Computing has been envisioned as the nextgeneration architecture of IT enterprise, due to its long list of unprecedented advantages in the IT history: on-demand self-service, ubiquitous network access, location independent resource pooling, rapid resource elasticity, usage-based pricing and transference of risk [1]. As a disruptive technology with profound implications, Cloud Computing is transforming the very nature of how businesses use information technology. One fundamental aspect of this paradigm shifting is that data is being centralized or outsourced into the Cloud. From users' perspective, including both individuals and IT enterprises, storing data remotely into the cloud in a flexible on-demand manner brings appealing benefits: relief of the burden for storage management, universal data access with independent geographical locations, and avoidance of capital expenditure on hardware, software, and personnel maintenances, etc [2].

While Cloud Computing makes these advantages more appealing than ever, it also brings new and challenging security threats towards users' outsourced data. Since cloud service providers (CSP) are separate administrative entities, data outsourcing is actually relinquishing user's ultimate control over the fate of their data. As a result, the correctness of the data in the cloud is being put at risk due to the following reasons. First of all, although the infrastructures under the cloud are much more powerful and reliable than personal computing devices, they are still facing the broad range of both internal and external threats for data integrity. Examples of outages and security breaches of noteworthy cloud services appear from time to time [3]–[5]. Secondly, for the benefits of their own, there do exist various motivations for cloud service providers to behave unfaithfully towards the cloud users regarding the status of their outsourced data. Examples include cloud service providers, for monetary reasons, reclaiming storage by discarding data that has not been or is rarely accessed, or even hiding data loss incidents so as to maintain a reputation [6]–[8]. In short, although outsourcing data into the cloud is economically attractive for the cost and complexity of long-term large-scale data storage, it does not offer any guarantee on data integrity and availability. This problem, if not properly addressed, may impede the successful deployment of the cloud architecture.

As users no longer physically possess the storage of their data, traditional cryptographic primitives for the purpose of data security protection can not be directly adopted. Thus, how to efficiently verify the correctness of outsourced cloud data without the local copy of data files becomes a big challenge for data storage security in Cloud Computing. Note that simply downloading the data for its integrity verification is not a practical solution due to the expensiveness in I/O cost and transmitting the file across the network. Besides, it is often insufficient to detect the data corruption when accessing the data, as it might be too late for recover the data loss or damage. Considering the large size of the outsourced data and the user's constrained resource capability, the ability to audit the correctness of the data in a cloud environment can be formidable and expensive for the cloud users [8], [9]. Therefore, to fully ensure the data security and save the cloud users' computation resources, it is of critical importance to enable public auditability for cloud data storage so that the users may resort to a third party auditor (TPA), who has expertise and capabilities that the users do not, to audit the outsourced data when needed. Based on the audit result, TPA could release an audit report, which would not only help users to evaluate the risk of their subscribed cloud data services, but also be beneficial for the cloud service provider to improve their cloud based service platform [7]. In a word, enabling public risk auditing protocols will play an important role for this nascent cloud economy to become fully established, where users will need ways to assess risk and gain trust in Cloud.

Recently, the notion of public auditability has been proposed in the context of ensuring remotely stored data integrity under different systems and security models [6], [8], [10], [11]. Public auditability allows an external party, in addition to the user himself, to verify the correctness of remotely stored data. However, most of these schemes [6], [8], [10] do not support the privacy protection of users' data against external auditors, i.e., they may potentially reveal user data information to the auditors, as will be discussed in Section III-C. This severe drawback greatly affects the security of these protocols in Cloud Computing. From the perspective of protecting data privacy, the users, who own the data and rely on TPA just for the storage security of their data, do not want this auditing process introducing new vulnerabilities of unauthorized information leakage towards their data security [12]. Moreover, there are legal regulations, such as the US Health Insurance Portability and Accountability Act (HIPAA) [13], further demanding the outsourced data not to be leaked to external parties [7]. Exploiting data encryption before outsourcing [11] is one way to mitigate this privacy concern, but it is only complementary to the privacy-preserving public auditing scheme to be proposed in this paper. Without a properly designed auditing protocol, encryption itself can not prevent data from "flowing away" towards external parties during the auditing process. Thus, it does not completely solve the problem of protecting data privacy but just reduces it to the one of managing the encryption keys. Unauthorized data leakage still remains a problem due to the potential exposure of encryption keys. Therefore, how to enable a privacy-preserving third-party auditing protocol, independent to data encryption, is the problem we are going to tackle in this paper. Our work is among the first few ones to support privacy-preserving public auditing in Cloud Computing, with a focus on data storage. Besides, with the prevalence of Cloud Computing, a foreseeable increase of auditing tasks from different users may be delegated to TPA. As the individual auditing of these growing tasks can be tedious and cumbersome, a natural demand is then how to enable TPA to efficiently perform the multiple auditing tasks in a batch manner, i.e., simultaneously. To address these problems, our work utilizes the technique of public key based homomorphic authenticator [6], [8], [10], which enables TPA to perform the auditing without demanding the local copy of data and thus drastically reduces the communication and computation overhead as compared to the straightforward data auditing approaches. By integrating the homomorphic authenticator with random masking, our protocol guarantees that TPA could not learn any knowledge about the data content stored in the cloud server during the efficient auditing process. The aggregation and algebraic properties of the authenticator further benefit our design for the batch auditing. Specifically, our contribution in this work can be summarized as the following three aspects:

1) We motivate the public auditing system of data storage security in Cloud Computing and provide a privacy-preserving auditing protocol, i.e., our scheme supports an external auditor to audit user's outsourced data in the cloud without learning knowledge on the data content.
2) To the best of our knowledge, our scheme is the first to support scalable and efficient public auditing in the Cloud Computing. In particular, our scheme achieves batch auditing where multiple delegated auditing tasks from different users can be performed simultaneously by the TPA.
3) We prove the security and justify the performance of our proposed schemes through concrete experiments and comparisons with the state-of-the-art.

The rest of the paper is organized as follows. Section II introduces the system and threat model, our design goals, notations and preliminaries. Then we provide the detailed description of our scheme in Section III. Section IV gives the security analysis and performance evaluation, followed by Section V which overviews the related work. Finally, Section VI gives the concluding remark of the whole paper.

II. PROBLEM STATEMENT

A. *The System and Threat Model*

We consider a cloud data storage service involving three different entities, as illustrated in Fig. 1: the *cloud user* (U), who has large amount of data files to be stored in the cloud; the *cloud server* (CS), which is managed by *cloud service provider* (CSP) to provide data storage service and has significant storage space and computation resources (we will not differentiate CS and CSP hereafter.); the *third party auditor* (TPA), who has expertise and capabilities that cloud users do not have and is trusted to assess the cloud storage service security on behalf of the user upon request.

Users rely on the CS for cloud data storage and maintenance. They may also dynamically interact with the CS to access and update their stored data for various application purposes. The users may resort to TPA for ensuring the storage security of their outsourced data, while hoping to keep their data private from TPA. We consider the existence of a semi-trusted CS as [14] does. Namely, in most of time it behaves properly and does not deviate from the prescribed protocol execution. However, during providing the cloud data storage based services, for their own benefits the CS might neglect to keep or deliberately delete rarely accessed data files which belong to ordinary cloud users. Moreover, the CS may decide to hide the data corruptions caused by server hacks or Byzantine failures to maintain reputation. We assume the TPA, who is in the business of auditing, is reliable and independent, and thus has no incentive to collude with either the CS or the users during the auditing process. TPA should be able to efficiently audit the cloud data storage without local copy of data and without bringing in additional on-line burden to cloud users. However, any possible leakage of user's outsourced data towards TPA through the auditing protocol should be prohibited.

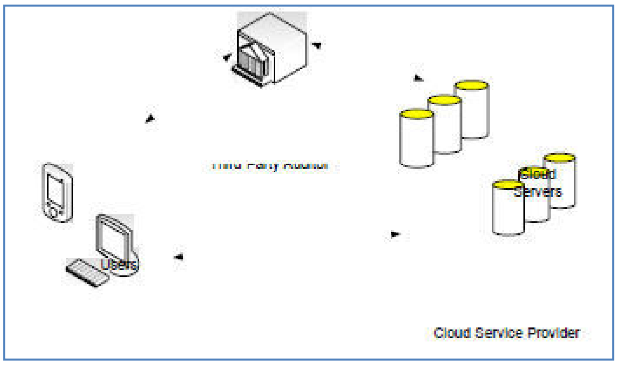


Fig. 1: The architecture of cloud data storage service

Note that to achieve the audit delegation and authorize CS to respond to TPA's audits, the user can sign a certificate granting audit rights to the TPA's public key, and all audits from the TPA are authenticated against such a certificate. These authentication handshakes are omitted in the following presentation.

B. *Design Goals*

To enable privacy-preserving public auditing for cloud data storage under the aforementioned model, our protocol design should achieve the following security and performance guarantee: 1) Public auditability: to allow TPA to verify the correctness of the cloud data on demand without retrieving a copy of the whole data or introducing additional on-line burden to the cloud users; 2) Storage correctness: to ensure that there exists no cheating cloud server that can pass the audit from TPA without indeed storing users' data intact; 3) Privacy-preserving: to ensure that there exists no way for TPA to derive users' data content from the information collected during the auditing process; 4) Batch auditing: to enable TPA with secure and efficient auditing capability to cope with multiple auditing delegations from possibly large number of different users simultaneously; 5) Lightweight: to allow TPA to perform auditing with minimum communication and computation overhead.

C. *Notation and Preliminaries*

- F – the data file to be outsourced, denoted as a sequence of n blocks $m_1, \ldots, m_n$ □ Z_p for some large prime p.
- $f_{key}(\cdot)$ – pseudorandom function (PRF), defined as: $\{0, 1\}^{□} \times key \rightarrow Z_p$.
- $\pi_{key}(\cdot)$ – pseudorandom permutation (PRP), defined as: $\{0, 1\}^{\log\ _2(n)} \times key \rightarrow \{0, 1\}^{\log\ _2(n)}$.
- $MAC_{key}(\cdot)$ – message authentication code (MAC) function, defined as: $\{0, 1\}^{□} \times key \rightarrow \{0, 1\}^{l}$.
- $H(\cdot)$, $h(\cdot)$ – map-to-point hash functions, defined as: $\{0, 1\}^{□} \rightarrow G$, where G is some group.

We now introduce some necessary cryptographic background for our proposed scheme.

Bilinear Map Let G_1, G_2 and G_T be multiplicative cyclic groups of prime order p. Let g_1 and g_2 be generators of G_1 and G_2, respectively. A bilinear map is a map $e : G_1 \times G_2 \rightarrow G_T$ with the following properties [15]: 1) Computable: there exists an efficiently computable algorithm for computing e; 2) Bilinear: for all u □ G_1, v □ G_2 and a, b □ Z_p, $e(u^a, v^b) = e(u, v)^{ab}$; 3) Non-degenerate: $e(g_1, g_2) = 1$; 4) for any u_1, u_2 □ G_1, v □ G_2, $e(u_1 u_2, v) = e(u_1, v) \cdot e(u_2, v)$.

III. THE PROPOSED SCHEMES

In the introduction we motivated the public auditability with achieving economies of scale for cloud computing. This section presents our public auditing scheme for cloud data storage security. We start from the overview of our public auditing system and discuss two straightforward schemes and

their demerits. Then we present our main result for privacypreserving public auditing to achieve the aforementioned design goals. Finally, we show how to extent our main scheme to support batch auditing for TPA upon delegations from multi-users.

A. Definitions and Framework of Public Auditing System

We follow the similar definition of previously proposed schemes in the context of remote data integrity checking [6], [10], [11] and adapt the framework for our privacy-preserving public auditing system.

A public auditing scheme consists of four algorithms (KeyGen, SigGen, GenProof, VerifyProof). KeyGen is a key generation algorithm that is run by the user to setup the scheme. SigGen is used by the user to generate verification metadata, which may consist of MAC, signatures, or other related information that will be used for auditing. GenProof is run by the cloud server to generate a proof of data storage correctness, while VerifyProof is run by the TPA to audit the proof from the cloud server.

Our public auditing system can be constructed from the above auditing scheme in two phases, Setup and Audit:

- Setup: The user initializes the public and secret parameters of the system by executing KeyGen, and preprocesses the data file F by using SigGen to generate the verification metadata. The user then stores the data file F at the cloud server, deletes its local copy, and publishes the verification metadata to TPA for later audit. As part of pre-processing, the user may alter the data file F by expanding it or including additional metadata to be stored at server.

- Audit: The TPA issues an audit message or challenge to the cloud server to make sure that the cloud server has retained the data file F properly at the time of the audit. The cloud server will derive a response message from a function of the stored data file F by executing GenProof. Using the verification metadata, the TPA verifies the response via VerifyProof.

Note that in our design, we do not assume any additional property on the data file, and thus regard error-correcting codes as orthogonal to our system. If the user wants to have more error-resiliency, he/she can first redundantly encode the data file and then provide us with the data file that has error correcting codes integrated.

B. The Basic Schemes

Before giving our main result, we first start with two warmup schemes. The first one does not ensure privacypreserving guarantee and is not as lightweight as we would like. The second one outperforms the first one, but suffers from other undesirable systematic demerits for public auditing: bounded usage and auditor statefulness, which may pose additional on-line burden to users as will be elaborated shortly. The analysis of these basic schemes will lead to our main result, which overcomes all these drawbacks.

Basic Scheme I The cloud user pre-computes MACs $\sigma i = MACsk(i||mi)$ of each block mi (i □ $\{1, \ldots, n\}$), sends both the data file F and the MACs $\{\sigma i\}1\leq i\leq n$ onto the cloud server, and releases the secret key sk to TPA. During the Audit phase, the TPA requests from the cloud server a number of randomly selected blocks and their corresponding MACs to verify the correctness of the data file. The insight behind this approach is that auditing most of the file is much easier than the whole of it. However, this simple solution suffers from the following severe drawbacks: 1) The audit from TPA demands retrieval of users' data, which should be prohibitive because it violates the privacy-preserving guarantee; 2) Its communication and computation complexity are both linear with respect to the sampled data size, which may result in large communication overhead and time delay, especially when the bandwidth available between the TPA and the cloud server is limited.

Basic Scheme II To avoid retrieving data from the cloud server, one may improve the above solution as follows: Before data outsourcing, the cloud user chooses s random message authentication code keys $\{sk\tau\}1\leq\tau\leq s$, pre-computes s MACs,$\{MACsk\ (F\ )\}1\leq\tau\leq s$ for the whole data file F , and publishes these verification metadata to TPA. The TPA can each time reveal a secret key $sk\tau$ to the cloud server and ask for a fresh keyed MAC for comparison, thus achieving privacy-preserving auditing. However, in this method: 1) the number of times a particular data file can be audited is limited by the number of secret keys that must be a fixed priori. Once all possible secret keys are exhausted, cloud user then has to retrieve data from the server in order to re-compute and re-publish new MACs to TPA. 2) The TPA has to maintain and update state between audits, i.e., keep a track on the possessed MAC keys. Considering the potentially large number of audit delegations from multiple users, maintaining such states for TPA can be difficult and error prone.

Note that another common drawback of the above basic schemes is that they can only support the case of static data, and none of them can deal with data dynamics, which is also of paramount importance for cloud storage systems. For the reason of brevity and clarity, we will focus on the static data, too, though our auditing protocol can be immediately adapted to support data dynamics, based on our previous work [8].

C. The Privacy-Preserving Public Auditing Scheme

To effectively support public auditability without having to retrieve the data blocks themselves, we resort to the homomorphic authenticator technique [6], [8], [10]. Homomorphic authenticators are unforgeable verification metadata generated from individual data blocks, which can be securely aggregated in such a way to assure an auditor that a linear combination of data blocks is correctly computed by verifying only the aggregated authenticator. However, the direct adoption of these techniques is not suitable for our purposes, since the linear combination of blocks may potentially reveal user data information, thus violating the privacy-preserving guarantee. Specifically, if enough number of the linear combinations of the same blocks are collected, the TPA can simply derive the user's data content by solving a system of linear equations.

Overview To achieve privacy-preserving public auditing, we propose to uniquely integrate the homomorphic authenticator with random masking technique. In our protocol, the linear combination of sampled blocks in the server's response is masked with randomness generated by a pseudo random function (PRF). With random masking, the TPA no longer has all the necessary information to build up a correct group of linear equations and therefore cannot derive the user's data content, no matter how many linear combinations of the same set of file blocks can be collected. Meanwhile, due to the algebraic property of the homomorphic authenticator, the correctness validation of the block-authenticator pairs will not be affected by the randomness generated from a PRF, which will be shown shortly. Note that in our design, we use public key based homomorphic authenticator, specifically, the BLS based signature [10], to equip the auditing protocol with public auditability. Its flexibility in signature aggregation will further benefit us for the multi-task auditing.

Scheme Details Let G_1, G_2 and G_T be multiplicative cyclic groups of prime order p, and $e : G_1 \times G_2 \rightarrow G_T$ be a bilinear map as introduced in preliminaries. Let g be the generator of G_2. $H(\cdot)$ is a secure map-to-point hash function: $\{0, 1\}^{\square} \rightarrow G_1$, which maps strings uniformly to G_1. Another hash function $h(\cdot) : G_1 \rightarrow Z_p$ maps group element of G_1 uniformly to Z_p. The proposed scheme is as follows:

`Setup Phase:`

1) The cloud user runs `KeyGen` to generate the system's public and secret parameters. He chooses a random $x \leftarrow Z_p$, a random element $u \leftarrow G_1$, and computes $v \leftarrow g^x$ and $w \leftarrow u^x$. The secret parameter is $sk = (x)$ and the public parameters are $pk = (v, w, g, u)$. Given data file $F = (m_1, \ldots, m_n)$, the user runs `SigGen` to compute signature σ_i for each block m_i: $\sigma_i \leftarrow (H(i) \cdot u^{m_i})^x \square G_1$ $(i = 1, \ldots, n)$. Denote the set of signatures by $\Phi = \{\sigma_i\}_{1 \leq i \leq n}$. The user then sends $\{F, \Phi\}$ to the server and deletes them from its local storage.

`Audit Phase:`

2) During the auditing process, to generate the audit message "*chal*", the TPA picks a random c-element subset $I = \{s_1, \ldots, s_c\}$ of set $[1, n]$, where $s_q = \pi_{k_{prp}}(q)$ for $1 \leq q \leq c$ and k_{prp} is the randomly chosen permutation key by TPA for each auditing. We assume that $s_1 \leq \cdots \leq s_c$.

For each element $i \square I$, the TPA also chooses a random value v_i (of a relative small bit length compared to $|p|$). The message "*chal*" specifies the positions of the blocks that are required to be checked in this `Audit` phase. The TPA sends the $chal = \{(i, v_i)\}_{i \square I}$ to the server.

3) Upon receiving challenge $chal = \{(i, v_i)\}_{i \square I}$, the server runs `GenProof` to generate a response proof of data storage correctness. Specifically, the server chooses a random element $r \leftarrow Z_p$ via $r = f_{k_{prf}}(chal)$, where k_{prf} is the randomly chosen PRF key by server for each auditing, and calculates $R = (w)^r = (u^x)^r \square G_1$. Let μ denote the linear combination of sampled blocks specified in $chal$: $\mu = {}_{i \square I} v_i m_i$. To blind μ with r, the server computes: $\mu = \mu + rh(R) \square Z_p$. Meanwhile, the server also calculates an aggregated signature $\sigma = {}_{i \square I} \sigma_i^{v_i} \square G_1$. It then sends $\{\mu, \sigma, R\}$ as the response proof of storage correctness to the TPA. With the response from the server, the TPA runs `VerifyProof` to validate the response by checking the verification equation

$$e(\sigma \cdot (R^{h(R)}), g) \overset{?}{=} e(\prod_{i=s_1}^{s_c} H(i)^{v_i} \cdot u^{\mu}, v) \qquad (1)$$

The correctness of the above verification equation can be elaborated as follows:

$$\begin{aligned} e(\sigma \cdot R^{h(R)}, g) &= e(\prod_{i=s_1}^{s_c} \sigma_i^{v_i} \cdot (u^x)^{r \cdot h(R)}, g) \\ &= e(\prod_{i=s_1}^{s_c} (H(i) \cdot u^{m_i})^{x \cdot v_i} \cdot (u^{r \cdot h(R)})^x, g) \\ &= e(\prod_{i=s_1}^{s_c} (H(i)^{v_i} \cdot u^{m_i v_i}) \cdot (u^{r \cdot h(R)}), g)^x \\ &= e(\prod_{i=s_1}^{s_c} (H(i)^{v_i}) \cdot u^{\ +r \cdot h(R)}, g^x) \\ &= e(\prod_{i=s_1}^{s_c} (H(i)^{v_i}) \cdot u^{\mu}, v) \end{aligned}$$

It is clear that the random mask *R* has no effect on the *k* validity of the checking result. The security of this protocol will be proved in Section IV.

Discussion As analyzed at the beginning of this section, this approach ensures the privacy of user data content during the auditing process. Meanwhile, the homomorphic authenticator helps achieve the constant communication overhead for server's response during the audit: the size of *{σ, μ, R}* is fixed and has nothing to do with the number of sampled blocks *c*.

Note that there is no secret keying material or states for TPA to keep or maintain between audits, and the auditing protocol does not pose any potential on-line burden toward users. Since the TPA could "re-generate" the random *c*-element subset *I* = *{s*1 , . . . , *sc }* of set [1, *n*], where *sq* = *πk* (*q*), for 1 ≤ *q* ≤ *c*, unbounded usage is also achieved.

Previous work [6], [8] showed that if the server is missing a fraction of the data, then the number of blocks that needs to be checked in order to detect server misbehavior with high probability is in the order of *O*(1). For example, if the server is missing 1% of the data *F* , the TPA only needs to audit for *c* = 460 or 300 randomly chosen blocks of *F* so as to detect this misbehavior with probability larger than 99% or 95%, respectively. Given the huge volume of data outsourced in the cloud, checking a portion of the data file is more affordable and practical for both TPA and cloud server than checking all the data, as long as the sampling strategies provides high probability assurance. In Section IV, we will present the experiment result based on these sampling strategies.

D. Support for Batch Auditing

With the establishment of privacy-preserving public auditing in Cloud Computing, TPA may concurrently handle multiple auditing delegations upon different users' requests. The individual auditing of these tasks for TPA can be tedious and very inefficient. Given *K* auditing delegations on *K* distinct data files from *K* different users, it is more advantageous for TPA to batch these multiple tasks together and audit at one time. Keeping this natural demand in mind, we propose to explore the technique of bilinear aggregate signature [15], which supports the aggregation of multiple signatures by distinct signers on distinct messages into a single signature and thus provides efficient verification for the authenticity of all messages. Using this signature aggregation technique and bilinear property, we can now aggregate *K* verification equations (for *K* auditing tasks) into a single one, as shown in equation 2, so that the simultaneous auditing of multiple tasks can be achieved.

The details of extending our main result to this multi-user setting is described as follows: Assume there are *K* users in the system, and each user *k* has a data file *Fk* = (*mk,*1 , . . . , *mk,n*) to be outsourced to the cloud server, where *k* □ *{*1, . . . , *K}*. For a particular user *k*, denote his secret key as *xk* ← *Zp* , and the corresponding public parameter as (*vk, wk, gk, uk*) = (*gx* , *ux* , *g, uk*). In the Setup phase, each user *k* runs SigGen and computes signature *σk,i* ← [*H*(*k*||*i*) · *um*]*x* □*G*1 for block *mk,i* (*i* □ *{*1, . . . , *n}*). In the Audit phase, the TPA sends the audit challenge *chal* = *{*(*i, vi*)*}i*□*I* to theserver for auditing data files of all *K* users. Upon receiving *chal*, for each user *k* (*k* □ *{*1, . . . , *K}*), the server randomly picks *rk* □ Z*p* and computes

$$\mu_k = \mathop{}\limits_{i=s_1}^{s_c} v_i m_{k,i} + r_k h(R_k) \square Z_p \quad \text{and} \quad \sigma = \mathop{}\limits_{k=1}^{K} (\mathop{}\limits_{i=s_1}^{s_c} \sigma_{k,i}^{v_i}),$$

where *Rk* = (*wk*)*r* = (*ux*)*r* . The server then responses the TPA with *{σ, {μk}*1≤*k*≤*K, {Rk}*1≤*k*≤*K}*. Similar as the single user case, using the properties of the bilinear map, the TPA can check if the following equation holds:

$$e(\sigma \cdot \mathop{}\limits_{k=1}^{K} R_k^{h(R_k)}, g) \overset{?}{=} \mathop{}\limits_{k=1}^{K} e(\mathop{}\limits_{i=s_1}^{s_c} [H(k\|i)]^{v_i} \cdot (u_k)^{\mu_k}, v_k) \quad (2)$$

The left-hand side (LHS) of the equation expands as:

$$\begin{aligned} \text{LHS} &= e(\mathop{}\limits_{k=1}^{K} \mathop{}\limits_{i=s_1}^{s_c} \sigma_{k,i}^{v_i} (u^{x}_{k})^{r\; h(Rk)}, g) \\ &= e(\mathop{}\limits_{k=1}^{K} \mathop{}\limits_{i=s_1}^{s_c} [H(k\|i) \cdot u_k^{m_{k,i}}]^{x v_i} (u_k^{r_k h(Rk)})^{x_k}, g) \\ &= \mathop{}\limits_{k=1}^{K} e(\mathop{}\limits_{i=s_1}^{s_c} [H(k\|i) \cdot u_k^{m_{k,i}}]^{v_i} (u_k^{r_k h(Rk)}), g)^{x_k} \\ &= \mathop{}\limits_{k=1}^{K} e(\mathop{}\limits_{i=s_1}^{s_c} [H(k\|i)]^{v_i} \cdot (u_k)^{m_{k,i} + r_k h(Rk)}, g^{x_k}) \\ &= \mathop{}\limits_{k=1}^{K} e(\mathop{}\limits_{i=s_1}^{s_c} [H(k\|i)]^{v_i} \cdot (u_k)^{\mu_k}, v_k), \end{aligned}$$

which is the right hand side, as required.

Discussion As shown in equation 2, batch auditing not only allows TPA to perform the multiple auditing tasks simultaneously, but also greatly reduces both the communication cost on the server side and the computation cost on the TPA side. For saving communication cost, the bilinear aggregate signature ensures that the server only needs to send one group element σ in the response to TPA, instead of a set of $\{\sigma k\}1\leq k\leq K$ as required by individual auditing, where $\sigma k = \sigma k,i$ denotes the aggregation signature supposed to be returned for each user k. Meanwhile, aggregating K verification equations into one helps reduce the number of expensive pairing operations from $2K$, as required in the individual auditing, to $K + 1$. Thus, a considerable amount of auditing time is expected to be saved.

Note that the verification equation 2 only holds when all the responses are valid, and fails with high probability when there is even one single invalid response in the batch auditing. In many situations, a response collection may contain invalid responses, especially $\{\mu k\}1\leq k\leq K$, caused by accidental data corruption, or possibly malicious activity by a cloud server. The ratio of invalid responses to the valid could be quite small, and yet a standard batch auditor will reject the entire collection. To further sort out these invalid responses in the batch auditing, we can utilize a recursive divide-and-conquer approach (binary search). Specifically, if the batch auditing fails, we can simply divide the collection of responses into two halves, and recurse the auditing on halves via equation 2. Note that TPA may now require the server to send back all the $\{\sigma k\}1\leq k\leq K$, as the same in individual auditing. In Section IV-B2, we show through carefully designed experiment that using this recursive binary search approach, even if up to 20% of responses are invalid, batch auditing still performs faster than individual verification.

IV. SECURITY ANALYSIS AND PERFORMANCE EVALUATION

A. Security Proofs

We evaluate the security of the proposed scheme by analyzing its fulfillment of the security guarantee described in Section II, namely, the storage correctness and privacy preserving. We start from the single user case, where our main result is originated. Then we show how to extend the security guarantee to a multi-user setting, where batch auditing for TPA is enabled. Due to space limitation, we only list the proof sketches to gives a rough idea for achieving those guarantees. The detailed formalized proofs of Theorem 1, 2, and 3 are provided in the full version [16]. Note that all proofs are derived on the probabilistic base, i.e., with high probability assurance, which we omit writing explicitly.

1) Storage Correctness Guarantee: We need to prove that the cloud server can not generate valid response toward TPA without faithfully storing the data, as captured by Theorem 1.

Theorem 1: If the cloud server passes the Audit phase, then it must indeed possess the specified data intact as it is.

Proof Sketch: The proof consists of three steps. First, we show that there exists no malicious server that can forge a valid response $\{\sigma, \mu, R\}$ to pass the verification equation 1. The correctness of this statement follows from the Theorem 4.2 proposed in [10]. Note that the value R in our protocol, which enables the privacy-preserving guarantee, will not affect the validity of the equation, due to the hardness of discrete-log and the commutativity of modular exponentiation in pairing. Next, we show that if the response $\{\sigma, \mu, R\}$ is valid, where $\mu = \mu + rh(R)$ and $R = (v2\ )r$, then the underlying μ must be valid too. This can be derived immediately from the collision free property of hash function $h(\cdot)$ and determinism of discrete exponentiation.

Finally, we show that the validity of μ implies the correctness of $\{mi\}i$□I where $\mu = vimi$. Here we utilize the small exponent (SE) test technique of batch verification in [17]. Because $\{vi\}$ are picked up randomly by the TPA in each Audit phase, $\{vi\}$ can be viewed similarly as the random chosen exponents in the SE test [17]. Therefore, the correctness of individual sampled blocks is ensured. All above sums up to the storage correctness guarantee.

2) Privacy Preserving Guarantee: We want to make sure that TPA can not derive users' data content from the information collected during auditing process. This is equivalent to prove the Theorem 2. Note that if μ can be derived by TPA, then $\{mi\}i$□I can be easily obtained by solving a group of linear equations when enough combinations of the same blocks are collected.

Theorem 2: From the server's response $\{\sigma, \mu, R\}$, no information of μ will be leaked to TPA.

Proof Sketch: Again, we prove the Theorem 2 in three steps. First, we show that no information on μ can be learned from μ. This is because μ is blinded by r as $\mu = \mu + rh(R)$ and $R = (v2\ )r$, where r is chosen randomly by cloud server and is unknown to TPA. Note that even with R, due to the hardness of discrete-log assumption, the value r is still hidden against TPA. Thus, privacy of μ is guaranteed from μ.

Second, we show that no information on μ can be learned from σ, where

TABLE I: Notation summary of cryptographic operations

$Hash^t_{\mathrm{G}}$	hash t values into the group G.
Add^t_{G}	t additions in group G.
$Mult^t_{\mathrm{G}}$	t multiplications in group G.
$Exp^t_{\mathrm{G}}(\ )$	t exponentiations g^{a_l}, where g □ G, $\lvert a_l \rvert =$.
$m\text{-}MultExp^t_{\mathrm{G}}(\ )$	t m-term exponentiations $\ {}^{m}_{i=1} g^{a_l}$.
$Pair^t_{\mathrm{G,H}}$	t pairings $e(g_l, h_l)$, where g_l □ G, h_l □ H.
$m\text{-}MultPair^t_{\mathrm{G,H}}$	t m-term pairings $\ {}^{m}_{i=1} e(g_l, h_l)$.

$$\sigma \;=\; \underset{i□I}{\ }\ \sigma_i^{v_i} = \underset{i□I}{\ }\ (H(i)\cdot u^{m_i})^{x\cdot v_i}$$

$$= [\underset{i□I}{\ }\ H(i)^{v_i}\cdot u^{\mu}]^{x} = [\underset{i□I}{\ }\ H(i)^{v}]^{x}\cdot[(u^{\mu})^{x}].$$

This can be analyzed as follows: $(u^{\mu})^{x}$ is blinded by $[\ _{i□I} H(i)^{v_i}]^{x}$. However, to compute $[\ _{i□I} H(i)^{v}]^{x}$ from $[\ _{i□I} H(i)^{v_i}]$ and g^{x}, which is the only information TPA can utilize, is a computational Diffie-Hellman problem, which can be stated as: given g, g^{a}, g^{b}, compute g^{ab}. This problem is hard for unknown a, b □ Z_p. Therefore, on the basis of computational Diffie-Hellman assumption, TPA can not derive the value of $(u^{\mu})^{x}$, let alone μ .

Finally, all that remains is to prove from *{σ, μ, R}*, still no information on μ can be obtained by TPA. Recall that r is a random private value chosen by the server and $\mu = \mu + rh(R)$. Following the same technique of Schnorr signature [18], our auditing protocol between TPA and cloud server can be regarded as a provably secure honest zero knowledge identification scheme [19], by viewing μ as a secret key and $h(R)$ as a challenge value, which implies no information on μ can be leaked. This completes the proof of Theorem 2.

3) Security Guarantee for Batch Auditing: Now we show that extending our main result to a multi-user setting will not affect the aforementioned security insurance, as shown in Theorem 3:

Theorem 3: Our batch auditing protocol achieves the same storage correctness and privacy preserving guarantee as in the single-user case.

Proof Sketch: Due to the space limitation, we only prove the storage correctness guarantee. The privacy-preserving guarantee in the multi-user setting is similar to that of Theorem 2, and thus omitted here. The proposed batch auditing protocol is built upon the aggregate signature scheme proposed in [15]. According to the security strength of aggregate signature [15], in our multi-user setting, there exists no malicious cloud servers that can forge valid $\mu 1$ *, . . . ,* μk in the responses to pass the verification equation 2. Actually, the equation 2 functions as a kind of screening test as proposed in [17]. While the screening test may not guarantee the validity of each individual σk, it does ensure the authenticity of μk in the batch auditing protocol, which is adequate for the rationale in our case. Once the validity of $\mu 1$ *, . . . ,* μk is guaranteed, from the proof of Theorem 1, the storage correctness guarantee in the multi-user setting is achieved.

TABLE II: Performance comparison under different number of sampled blocks c for high assurance auditing.

	Our Scheme		[10]	
Sampled blocks c	460	300	460	300
Sever compt. time (ms)	405.57	273.34	403.69	270.46
TPA compt. time (ms)	525.89	493.25	524.02	491.38
Comm. cost (Byte)	60	40	60	40

B. Performance Analysis

We now assess the performance of the proposed privacypreserving public auditing scheme. We will focus on the extra cost introduced by the privacy-preserving guarantee and the efficiency of the proposed batch auditing technique. The experiment is conducted using C on a Linux system with an Intel Core 2 processor running at 1.86 GHz, 2048 MB of RAM, and a 7200 RPM Western Digital 250 GB Serial ATA drive with an 8 MB buffer. Algorithms use the Pairing-Based Cryptography (PBC) library version 0.4.18. The elliptic curve utilized in the experiment is a MNT curve, with base field size of 159 bits and the embedding degree 6. The security level is chosen to be 80 bit, which means $|vi| = 80$ and $|p| = 160$. All experimental results represent the mean of 20 trials.

1) Cost of Privacy-preserving Guarantee: We begin by estimating the cost in terms of basic cryptographic operations, as notated in Table I. Suppose there are c random blocks specified in the *chal* during the Audit phase. Under this setting, we quantify the extra cost introduced by the support of privacy-preserving into server computation, auditor computation as well as communication overhead. On the server side, the generated response includes an aggregated signature

$\sigma = {}_{i□I}\,\sigma_i^{v_i r}$ □ G_1, a random metadata $R = (w)^r = (u^{\lambda})^r$ □ G_1, and a blinded linear combination of sampled blocks $\mu = {}_{i□I}\, v_i m_i + rh(R)$ □ Z_p. The corresponding computation cost is c-$MultExp^1_G(|v_i|)$, $Exp^1_G(|p|)$, and $Add^c_{Z_p} + Mult^{c+1}_{Z_p} + Hash^1_{Z_p}$, respectively. Compared to the existing homomorphic authenticator based solution for ensuring remote data integrity [10][1], the extra cost for protecting the user privacy, resulted from the random mask R, is only a constant: $Exp^1_G(|p|) + Mult^1_{Z_p} + Hash^1_{Z_p} + Add^1_{Z_p}$, which has nothing to do with the number of sampled blocks c. When c is set to be 460 or 300 for high assurance of auditing, as discussed in Section III-C, the extra cost for privacy-preserving guarantee on the server side would be negligible against the total server computation for response generation.

Similarly, on the auditor side, upon receiving the response $\{\sigma, R, \mu\}$, the corresponding computation cost for response validation is $Hash^1$ + c-$MultExp^1(|v_i|)$ + $Hash^c$ + $Mult^2 + Exp^2(|p|) + Pair_{G,G}$, among which only $Hash^1$ + $Exp^1(|p|)$ + $Mult^1$ account for the additional constant computation cost. For c = 460 or 300, and considering the relatively expensive pairing operations, this extra cost imposes little overhead on the overall cost of response validation, and thus can be ignored. For the sake of completeness, Table II gives the experiment result on performance comparison between our scheme and the state-of-the-art [10]. It can be shown that the performance of our scheme is almost the same as that of [10], even if our scheme supports privacy-preserving guarantee while [10] does not. Note that in our scheme, the server's response *{σ, R, μ}* contains an additional random element R, which is a group element and has the same size of 159 bits as σ does. This explains the extra communication cost of our scheme opposing to [10].

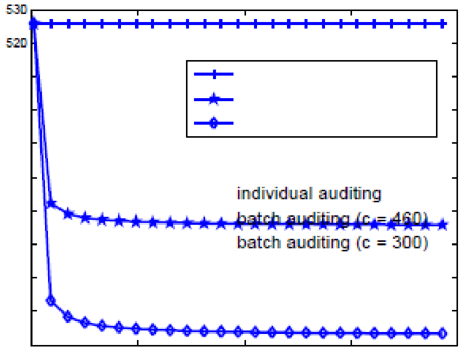


Fig. 2: Comparison on auditing time between batch auditing and individual auditing. Per task auditing time denotes the total auditing time divided by the number of tasks.

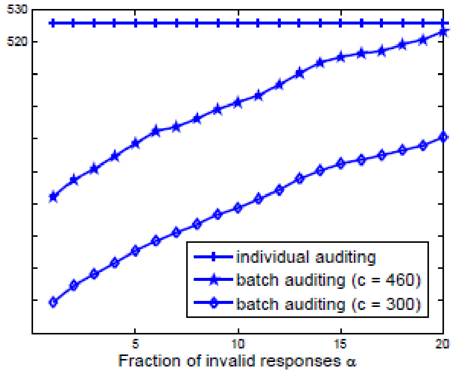


Fig. 3: Comparison on auditing time between batch auditing and individual auditing, when α-fraction of 256 responses are invalid. Per task auditing time denotes the total auditing time divided by the number of tasks.

2) *Batch Auditing Efficiency:* Discussion in Section III-D gives an asymptotic efficiency analysis on the batch auditing, by considering only total number of expensive pairing operations. However, on the practical side, there are additional operations required for batching, such as modular exponentiations and multiplications. Meanwhile, the different sampling strategies, i.e., different number of sampled blocks c, is also a variable factor that affects the batching efficiency. Thus, whether the benefits of removing pairings significantly outweighs these additional operations is remained to be verified. To get a complete view of batching efficiency, we conduct a timed batch auditing test, where the number of auditing tasks is increased from 1 to approximately 200 with intervals of 8. The performance of the corresponding non-batched (individual) auditing is provided as a baseline for the measurement. Following the same experimental setting as c = 460 and 300, the average per task auditing time, which is computed by dividing total auditing time by the number of tasks, is given in Fig. 2 for both batch auditing and the individual auditing. It can be shown that compared to individual auditing, batch auditing indeed helps reduce the TPA's computation cost, as more than 11% and 17% of per-task auditing time is saved, when c is set to be 460 and 300, respectively.

3) *Sorting out Invalid Responses:* Now we use experiment to justify the efficiency of our recursive binary search approach for TPA to sort out the invalid responses when batch auditing fails, as discussed in Section III-D. This experiment is tightly pertained to work in [20], which evaluates the batch verification efficiency of various short signature schemes.

To evaluate the feasibility of the recursive approach, we first generate a collection of 256 valid responses, which implies the TPA may concurrently handle 256 different auditing delegations. We then conduct the tests repeatedly while randomly corrupting an α-fraction, ranging from 0 to 20%, by replacing them with random values. The average auditing time per task against the individual auditing approach is presented in Fig. 3. The result shows that even the number of invalid responses exceeds 15% of the total batch size, the performance of batch auditing can still be safely concluded as more preferable than the straightforward individual auditing. Note that the random distribution of invalid responses within the collection is nearly the worst-case for batch auditing. If invalid responses are grouped together, it is possible to achieve even better results.

VI. CONCLUSION

In this paper, we propose a privacy-preserving public auditing system for data storage security in Cloud Computing. We utilize the homomorphic authenticator and random masking to guarantee that TPA would not learn any knowledge about the data content stored on the cloud server during the efficient auditing process, which not only eliminates the burden of cloud user from the tedious and possibly expensive auditing task, but also alleviates the users' fear of their outsourced data leakage. Considering TPA may concurrently handle multiple audit sessions from different users for their outsourced data files, we further extend our privacy-preserving public auditing protocol into a multi-user setting, where TPA can perform the multiple auditing tasks in a batch manner,i.e.,simultaneously. Extensive analysis shows that the proposed schemes are provably secure and highly efficient.

REFERENCES

[1] P. Mell and T. Grance, "Draft nist working definition of cloud computing," Referenced on June. 3rd, 2009 Online at http://csrc.nist.gov/ groups/SNS/cloud-computing/index.html, 2009.

[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A berkeley view of cloud computing," University of California, Berkeley, Tech. Rep. UCB-EECS-2009-28, Feb 2009.

[3] Amazon.com, "Amazon s3 availability event: July 20, 2008," Online at http://status.aws.amazon.com/ s3-20080720.html, July 2008.

[4] S. Wilson, "Appengine outage," Online at http://www.cio-weblog.com/ 50226711/ appengine outage.php, June 2008.

[5] B. Krebs, "Payment Processor Breach May Be Largest Ever," Online at http://voices.washingtonpost.com/securityfix/2009/01/payment processor breach may b.html, Jan. 2009.

[6] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable data possession at untrusted stores," Cryptology ePrint Archive, Report 2007/202, 2007, http://eprint.iacr.org

[7] M. A. Shah, R. Swaminathan, and M. Baker, "Privacy-preserving audit and extraction of digital contents," Cryptology ePrint Archive, Report 2008/186, 2008, http://eprint.iacr.org/.

[8] Q. Wang, C. Wang, J. Li, K. Ren, and W. Lou, "Enabling public verifiability and data dynamics for storage security in cloud computing," in Proc. of ESORICS'09, Saint Malo, France, Sep. 2009.

[9] Cloud Security Alliance, "Security guidance for critical areas of focus in cloud computing," 2009, http://www.cloudsecurityalliance.org.

[10] H. Shacham and B. Waters, "Compact proofs of retrievability," in Proc. of Asiacrypt 2008, vol. 5350, Dec 2008, pp. 90–107.

[11] A. Juels and J. Burton S. Kaliski, "Pors: Proofs of retrievability for large files," in Proc. of CCS'07, Alexandria, VA, October 2007, pp. 584–597.

[12] M. A. Shah, M. Baker, J. C. Mogul, and R. Swaminathan, "Auditing to keep online storage services honest," in Proc. of HotOS'07. Berkeley, CA, USA: USENIX Association, 2007, pp. 1–6.

[13] 104th United States Congress, "Health Insurance Portability and Accountability Act of 1996 (HIPPA)," Online at http://aspe.hhs.gov/ admnsimp/pl104191.htm, 1996, last access: July 16, 2009.

[14] S. Yu, C. Wang, K. Ren, and W. Lou, "Achieving secure, scalable, and fine-grained access control in cloud computing," in Proc. of IEEE INFOCOM'10, San Diego, CA, USA, March 2010.

[15] D. Boneh and C. Gentry, "Aggregate and verifiably encrypted signatures from bilinear maps," in Proceedings of Eurocrypt 2003, volume 2656 of LNCS. Springer-Verlag, 2003, pp. 416–432.

[16] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for data storage security in cloud computing," Cryptology ePrint Archive, Report 2009/579, 2009, http://eprint.iacr.org/.

[17] M. Bellare, J. Garay, and T. Rabin, “Fast batch verification for modular exponentiation and digital signatures,” in Proceedings of Eurocrypt 1998, volume 1403 of LNCS. Springer-Verlag, 1998, pp. 236–250.

[18] C. Schnorr, “Efficient identification and signatures for smart cards,” in Proceedings of Eurocrypt 1989, volume 435 of LNCS. Springer-Verlag, 1989, pp. 239–252.

[19] D. Pointcheval and J. Stern, “Security proofs for signature schemes,” in Proceedings of Eurocrypt 1996, volume 1070 of LNCS. Springer-Verlag, 1996, pp. 387–398.

[20] A. L. Ferrara, M. Greeny, S. Hohenberger, and M. Pedersen, “Practical short signature batch verification,” in Proceedings of CT-RSA, volume 5473 of LNCS. Springer-Verlag, 2009, pp. 309–324.

[21] G. Ateniese, R. D. Pietro, L. V. Mancini, and G. Tsudik, “Scalable and efficient provable data possession,” in Proc. of SecureComm’08, 2008.

[22] C. Wang, Q. Wang, K. Ren, and W. Lou, “Ensuring Data Storage Security in Cloud Computing,” in Proc. of IWQoS’09, July 2009.

[23] C. Erway, A. Kupcu, C. Papamanthou, and R. Tamassia, “Dynamic provable data possession,” in Proc. of CCS’09, 2009.

◈◈◈

Continuous Integration For Enterprise Systems Using Heterogeneous Techniques

Raksha SV

Department of Information Science Engineering, MSRIT, Bangalore, India
E-mail : raksha.vjkmr@gmail.com

***Abstract* -** Continuous integration is an extensive experience in the Enterprise and system architectures, it is the concept of integrating new code into existing code and then utilizing the testing techniques defined by Extreme Programming. This practice yields units of code that are continually tested during development. Using continuous integration, a programmer integrates new code into existing code after the existing code has been relentlessly built. Therefore, the release of the change is done when it is and everyone knows about the new functionality at the time of the release. While this is dynamic, it should not pose problems since the code is released only after thorough testing. This particular idea generates much discussion among the software practitioners, but the result is that it does advocate the thorough building and testing of modules prior to release as its main premise. On this main premise there is no resistance. It also recognizes that the integration of code in the object-oriented paradigm deserves some focus. One of the key requirement of this paper is to develop a common platform to do automated build environment, as the current environment is emulated from decade old build system based on Make and there is no way to track changes made to the system for detecting dependency flaws.

***Keywords* -** *Continuous Integration, Mirroring, Virtualization, Computing ,Repository.*

I. INTRODUCTION

IT today strives to create an infrastructure that is dynamic, scalable and flexible to satisfy the needs of mission-critical work as well as the development and deployment of new workloads. Business trends are driving a new generation of applications that are diverse in their functions, processing requirements and processing power requirements. Advanced Analytics is fast becoming a staple of business processing trying to take the existing Non-virtualized infrastructure and turn it into a more efficient and highly utilized environment. Without the benefits of virtualization is costing companies an enormous amount of time and money, but still yielding only minimal results. The trend is to use platform-agnostic software. However, that does not always take advantage of the inherent architecture of the platform, such as symmetric multiprocessing systems or workloads that require optimization to meet throughput and processor utilization objectives.

It is currently under a mandate to use resources more efficiently. In the past, IT organizations typically built an infrastructure that tightly coupled application workloads with physical assets, forcing IT administrators to continually perform repetitive manual tasks just to keep the operating environment up and running. In an effort to maintain some control, software is layered alongside the multiple instances of operating systems to enforce security, manage availability, and ensure performance. Then, to further compound the complexity, some applications require unique hardware, software, and skill sets to support specific business requirements.

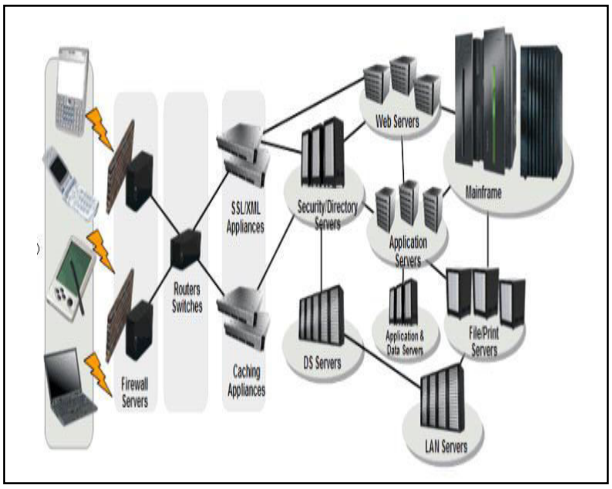


Fig. 1 : Heterogeneous System Infrastructure

The heterogeneous system infrastructure illustrates the IT dilemma in managing the complexity that business solutions require today. This illustration depicts the platforms necessary to process a single transaction with all the components representing islands of computing.

These islands of computing all use different languages and there is little coordination between them.

For business to gain a competitive edge, IT must be able to rapidly deploy servers and server resources to support new multi platform applications. The methodology of sizing the solution, ordering the servers waiting for installation, cabling and testing is changing. Virtualization with a supporting infrastructure can be used to provision virtual servers on existing hardware. Cloning production environment can be very difficult task but virtualization of environments can help.

II. DRAWBACKS OF THE EXISTING SYSTEM

Learning Curve - It is time consuming for any new member to learn the system .All new driver builds will have build issues which have to be resolved manually and no easy way out.

1. Remote Build Machine Storage is generally limited just enough to keep few versions of driver. Increase in team size has impact on storage. Generally there is no Sandbox/Desktop build capability
2. There are no automated tests to run the Junit test cases.
3. No build triggering on atomic commits .No report of code quality/complexity Ex:Cyclomatic, duplicated code, metrices related to LOC.
4. No component level output bundle /artifact

All these concepts are present as independent entities or some together to achieve the desired competency. The idea of this paper is to put together all the above flaws and bring in the concept of continuous Integration

III. CONTINUOUS INTEGRATION

Continuous Integration is a software development practice where changes are made frequently and are updated to the larger code base on a daily basis. The goal of CI is to provide rapid feedback. The process of frequently integrating one's new or revised code with the current code repository should be done frequently enough that no intermediate window remains between commit and build, so that no errors can arise without being noticed by the developers. Rather than periodically scheduling a build, normal process is to trigger these builds by every commit to a repository.

On all the projects, the difficulty of making changes is directly related to specific design qualities. The most obvious is duplication: when a change we wanted was localized in a single class, it was trivial. When we had to modify similar code over and over, the change was tedious and took a lot longer. Other design qualities also affected our ability to make changes. Simplicity is important. When adding features, we were better off when there was no pre-existing design to handle that feature. Adding code that doesn't exist is easy, fixing someone's preconceptions about a feature first is more costly. The side- bar lists a number of other design qualities that have made our projects easier to maintain and change.

Completely script your build so that it can be run by a headless process such as a CI server and that it doesn't rely on an IDE or similar dependency. You "centralize software assets" by putting everything that is needed to build the software in the version control repository. Finally, ensure that you can "perform single command builds" without needing to configure this file or that variable, ensuring a smooth integration build.

Adopt CI by determining on where to focus or what to improve by using self check value stream map. Involve and educate the whole team. Management buy in and support to include CI cost/plans in project cost/plan(unit tests,build/deploy tool and scripts)

The advantages of adopting CI:

- Better views into project health
  - Notice trends and then take action earlier
  - Developers have more confidence to refactor
- Stable, working software available sooner and more often
  - Focus is on getting working builds
  - Less escapes, better product for validation testing
  - Early feedback from stakeholders to validate requirements
  - Greater confidence in the product
  - Deployable software is a tangible in short time boxed iterations.

The various continuous Integration engines include Rational team concert build engine,Jenkins,cruise control and rational build forge.

IV. TECHNICAL SPECIFICATION OF CI ARCHITECTURE

A brief insight into the working of CI

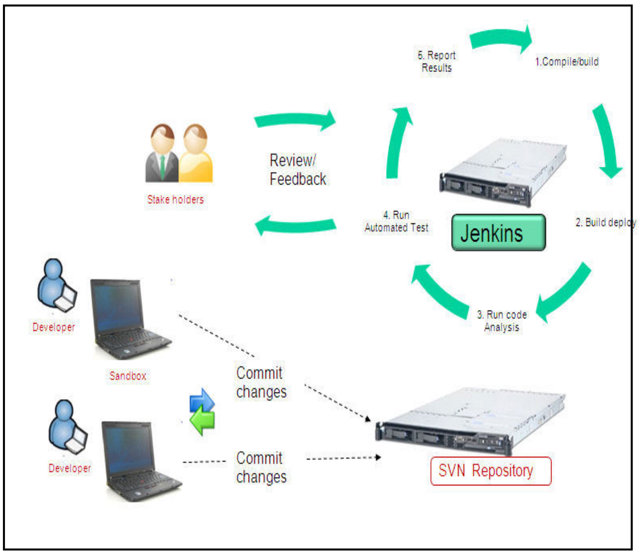


Fig. 2 : CI Life cycle

Manual build approach is similar to the local build that a developer does before a commit into the repository. The developer goes to the integration machine, checks out the head of the mainline (where his last commit resides) and kicks off the integration build. He monitors the progress, and if the build succeeds he is done with his commit. A continuous integration server acts as a monitor to the repository. Every time a commit against the repository finishes the server automatically checks out the sources onto the integration machine, initiates a build, and notifies the committer of the result of the build. The committer isn't done until they get the notification - usually an email. Mostly companies do regular builds on a timed schedule, like a nightly build. This is not the same thing as a continuous build and isn't enough for continuous integration. The whole point of continuous integration is to find problems as soon as you can. Nightly builds mean that bugs lie undetected for a whole day before anyone discovers them. The whole point of working with CI is for developing on a known stable base.

V. TECHNIQUES FOR IMPLEMENTING CI

- Jenkins for build
- Mirroring repositories using Subversion

JENKINS – open source CI server

Jenkins, is an open source continuous integration tool written in Java. Jenkins provides continuous integration services for software development, primarily in the Java programming language. It is a server-based system running in a servlet container such as Apache Tomcat. It supports tools including CVS, Subversion, Git and Clearcase, and can execute Apache Ant based projects including shell scripts and Windows batch commands. Builds can be triggered by commit in a version control system, scheduling mechanisms like cron jobs , build when other builds have completed, by requesting a specific build URL.

A. Why choose Jenkins

enkins improves the Productivity by detecting build breaks sooner, report failing tests more clearly and makes progress more visible. Continuous Integration can be done using Jenkins as it supports convention over configuration:no setup or minimal configuration required to build artifacts,Multi-technology, multi-platform,extensible environment.

1) Implementation

Jenkins provides a user friendly web user interface (UI) where you can access information like build number, last updated files, build artifacts and test reports. You can create continuous integration jobs and configure relative properties, such as SVN source location, schedule, and e-mail notification settings though this web UI. A build process requires a build tool such as Ant ,build invoker and some shell scripts. It can be scheduled or done manually using the invoker. Before Jenkins starts to build and execute the regression bucket, it first checks out the latest version of source programs from the SVN repository. The mechanism can be configured either using Update or Revert from the repository, and it will record current version of sources within every Run. Jenkins tool is installed in which a new job can be created to include the projects artifacts which is stored in the job directory. The build script that performs the build (ant, maven, shell script, batch file) is where the real work happens. The environment variables such as version,build.xml file location has to be set corresponding to Jenkins server. Source code resides in a CVS or SVN repository. The developer checks out the code to the workspace in Jenkins along with the build invoker-build.xml file to Jenkins/jobs/PROJECTNAME/ workspace. Jenkins can poll the revision control system using the same syntax for crontab in linux. If the polling period is shorter, there could multiple builds for each change. Thereby adjust your polling period to be longer than the amount of time it takes to poll your revision control system, or use a post-commit trigger. You can examine the Polling Log for each build. Optional steps to notify other people/systems with the build result, such as sending e-mails, IMs, updating issue tracker.

Frequently categorize Developer Tests, use a Dedicated Integration Build Machine and Use Continuous Feedback Mechanisms. Continuous integration using Jenkins is easier to monitor and can be refactored depending upon the nature of the project and the team size.

B. Mirroring Repositories Using Subversion.

Mirroring

The primary goal of mirroring is to provide a robust, economic replication solution. Traditional replication solutions often require substantial capital investments in infrastructure, deployment, configuration, software licensing, and planning. The solutions that rely on shared resources are susceptible to single point of failure for that shared resource. Combining continuous integration with mirroring provides an additional level of availability, thus greatly minimizing work flow and user disruption. Configuring the two mirror members in separate data centers offers additional redundancy and protection from catastrophic events. By utilizing logical data replication, mirroring reduces the potential risks associated with physical replication, such as out-of-order updates and carry-forward corruption, which are possible with other replication technologies such as SAN-based replication.

In this paper we explore two ideas of mirroring using the subversion.

- Mirroring the multi-site servers
- Mirroring using a dump

1) Mirroring the multi-site servers

Mirroring is necessary for synchronization and backing up of data. Create a master-slave mirror setup of the desired server which houses the organizations data. The primary/master servers configuration will be managed by the repository administrator using the svnadmin, providing the ability to create subversion repositories and perform several maintenance operations on those repositories.

The architectural representation is given below

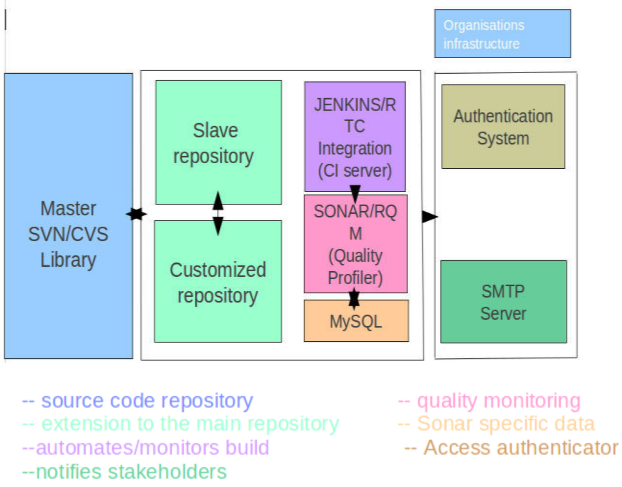


Fig. 3 : Mirroring Repository-Flow of Actions.

The primary/master repository is where all the project related artifacts resides. The latest driver releases lie in the head of the repository. Hierarchy of the repository is given by the svnadmin. Here We can treat the master repository as any other file system and recursively copy it to the mirror location. This is not ideal if the repository is in use -you're copying a moving target - so you will have to take down the subversion server while making the mirror. If this downtime is acceptable, netcat, nc, combined with tar is a neat way to recursively copy a directory across a network connection using TCP/IP onto the slave repository. This code synchronization will happen at regular intervals to update the master with all changes made to the slave. The customized repository is used as a sandbox where all the deliverables can be accessed from the slave and modifications can be made right here without affecting the master repository. Build process can be triggered through Jenkins which is associated with Sonar which is a Quality profiling tool used for code quality management such as unit tests, coding rules, duplication's. MySql is used for querying. The organizations infrastructure consists of an Authentication Engine and SMTP server. Authentication engine allows authenticates user to access the build process and SMTP server to notify stakeholders of successful or failed builds.

The main advantage of mirroring is that the master repository remains untouched and refactoring is made by taking advantage of the slave repository,when changes are made unknowingly it does not affect the primary repository and can be handled at the slave repository level.

2) Mirroring using Dump

Another way to mirror a subversion repository is to combine svnadmin dump with svadmin load.

svnadmin dump PATH_TO_REPOS | svnadmin load PATH_TO_MIRROR

Svnadmin dump is designed to create a portable repository dump. The resulting dumpfile can be loaded into a new subversion repository-even if the new repository is using a different database back-end, or even a different revision of subversion. Svnadmin dump will happily run on a live repository (no need to take the server down). In short, combining svnadmin dump with svnadmin load is probably more powerful than we need if we just want to mirror our repository to a new location. Svnadmin dump - on its own - is the best way to fully backup a repository, since the dumpfile it creates is portable (as described above). If we replicate a repository by piping svnadmin dump to svnadmin load, we lose the dumpfile in the pipeline and do far more work than we need to.svnadmin dump does not dump your repository configuration files and hook scripts. If your backup strategy is based around these commands, you will need separate arrangements for backing up hook scripts and configuration files.

VI. RESULTS

Time/Cost investment => Overall Time/Cost savings and higher product quality

In Continuous Integration executing test is all part of build and build fails if the test fails, XUnit ,JUnit, Nunit – for unit testing purpose. From an efficiency perspective this increase in development time is offset by the by the reduced maintenance costs due to the improvement in quality . For each new deliverable driver received from development team daily, one person hour in configuring testing environment is saved, checking out the latest sources and starting execution. It’s a common understanding that for software development; the earlier a defect/error can be detected and fixed, the lower cost it is. E-mail notification allows timely response to errors and defect. Individual tester is now notified when his or her own programs has affected the whole integrity of automation framework. The effort for fixing errors in the regression bucket is now distributed to each tester, and whole team productivity is increased.

VI. CONCLUSION

Continuous Integration is about delivering value predictably .In terms of software quality, continuous integration can help measure cyclomatic complexity, code duplication, dependencies and coding standards so that developers can proactively refactor code before a defect is introduced. If a defect is introduced into a code base, CI can provide feedback soon after, when defects are less complex and less expensive to fix. Also, when using an effective developer testing regimen, CI provides quick feedback, via regression tests. CI acts as a mirror of your software under development.

Research in pedagogical patterns for implementing pair programming and grading the results of such activities are still emerging. With applying Jenkins to realize the continuous integration approach in Agile process, we are now able to automate and facilitate the process of executing regression testing. The continuous integration approach in Agile gives us better chances to identify programs causing integration errors as early as possible, and secures the product integrity anytime in a regular scheduled time frame in each sprint

REFERENCES

[1] http://en.wikipedia.org/wiki/Hudson_(software)

[2] https://wiki.jenkins-ci.org/display/JENKINS/

[3] http://docs.intersystems.com/cache20121/csp/docbook/DocBook.UI.Page.cls?KEY=GHA_mirror

[4] http://svnbook.redbean.com/en/1.7/svn.reposadmin.maint.html#svn.reposadmin.maint.replication.pre-revprop-change

[5] http://martinfowler.com/articles/agileOffshore.html

[6] http://www.redbooks.ibm.com/redpieces/abstracts/sg247921.html?Open

[7] http://martinfowler.com/articles/continuousIntegration.html

◈◈◈

Maximum Likelihood Factor Analysis Based Speaker Recognition Using Pitch Strength

Rama Koteswara Rao P[1], Srinivasa Rao Y[2] & Vijaya Kumar D[3]

[1&3]ECE Department, UshaRama College of Engineering & Technology, Telaprolu, AP, India
[2] Instrument Technology Department, AU College of Engineering, Andhra University, Visakhapatnam, AP, India, [2&3]Member, IEEE
E-mail : prkr74@gmail.com[1], srinniwasarau@gmail.com[2], vijaykumarurs@ieee.org[3]

Abstract - Speaker recognition is important for successful development of speech recognizers in various real world applications. In this paper, the speaker recognizer was developed using sizable collection of various speakers both male as well as female with pitch strength as the feature. We proposed Maximum Likelihood Factor Analysis (PFA) technique for dimensionality reduction for accurate speaker recognition system. The first module performs feature extraction from speech samples taking pitch strength as the feature. The second module executes dimensionality reduction from the windowing of speech samples, where data samples are normally signified as matrices or higher order tensors. The system was trained by Support Vector Machine (SVM) using dimensionality reduced feature matrix. The implementation results show that the proposed system recognizes whether the given speaker is authorized or not.

Keywords - *Speaker recognition technique, MLFA, PFA,FA, SVM, Pitched, Unpitched, Dimensionality reduction.*

I. INTRODUCTION

Speech recognition is one of the most active areas of present Informatics [5] and together it acquired particular attention in the recent years [2]. Voice dialing, simple data entry, call routing and preparation of structured documents are the applications of speech recognition that have emerged over the recent years [1]. The basic method of speech recognition is to decode the speech signal in a chronological mode on the basis of observed acoustic characteristics of the signal and recognized relations between acoustic features and phonetic symbols [4]. Generally, a speech recognition system comprises of the subsequent components: signal processing, speech decoding, and adaptation [3]. Speech signal processing and feature extraction are the preliminary stages of any speech recognition system, it is by means of this component that the system considers the speech signal itself. Speech signal processing relates to the operations that are executed on the speech signal (namely, filtering, spectral analysis, digitization, etc.). Feature extraction is a pattern recognition term that relates to the exemplifying measurements that are carried out on a pattern (or signal). These features develop the input to the classifier that identifies the pattern [6].

In the work, we deal with one of the major areas of the speech recognition techniques, speaker recognition. Auto-matic speaker recognition is a vital, progressing technology with several possible applications in commerce and business, security, surveillance, etc [7] and it has been an extensive area of research [9]. Recognition systems that associate with obtaining the individuality of the person speaking [10] be different in their approach to speech knowledge. Speaker recognition anticipates differentiating one speaker from the others, irrespective of what they said, and therefore the differences of voice features among speakers must be utilized, while the phonemic dissimilarities for the speech signals are better discarded [8].

In this paper, we propose an effective speaker recognition/identification system that recognizes the authorized user's speech more accurately. Firstly, we propose to extract two main features from an input speech signal, namely, pitch and pitch strength as they heavily differentiate the speech of different users. The extracted feature matrix is in higher dimension that increases the complexity of further processes and so the dimension of the feature matrix is reduced by MLFA. The resultant dimensionality reduced matrix from MLFA, termed as Projection matrix, is used to train SVM for classifying the authorized and unauthorized users from their speech signal. The SVM is chosen for classification because of its ability to learn with very little samples. So, given a speech signal, the well-learned SVM can effectively recognize.

II. PROPOSED TECHNIQUE FOR EFFECTIVE SPEECH BASED RECOGNITION SYSTEM

Here, an effective speaker recognition technique is proposed to make an efficient speech recognition system which is very effective in the associated applications. The proposed technique is comprised of three stages of operation, namely, Feature extraction, Dimensionality reduction and Recognition. In the proposed technique, which is depicted in the Fig 1, extracts pitch and pitch strength as the feature. The extracted feature matrix is subjected to MLFA for dimensionality reduction and finally subjected to SVM. The SVM-based recognition consists of two processes, training and testing.

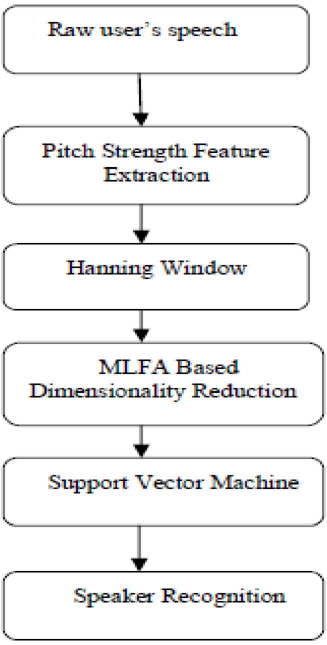


Fig.1. Proposed Speaker Recognition System

In the training process, dimensionality reduced feature samples of the known users are given to the SVM whereas in the testing process, the dimensionality reduced feature samples of the known/unknown users are given to the SVM. The SVM authenticate the user and outputs whether the user is authenticated or not (i.e. known or not). The detailed description of the proposed speaker recognition technique is given in the further sub-sections. Let $A_{ij}(t)$; $0 \le i \le N_u - 1$ and $0 \le j \le N_s^{(i)} - 1$, be the speech signal obtained from different users, where, N_u be the number of users and $N_s^{(i)}$ be the number of samples taken from i^{th} user, provided, $N_s^{(0)} = N_s^{(1)} = \cdots = N_s^{N_u - 1}$.The sample speech signals from different users are subjected to feature extraction that extracts pitch and its strength as the feature for every sample.

III FEATURE EXTRACTION

Pitch, by definition, is a subjective sensation where all the tones perceived by the listener are particularly assigned to relative positions on a musical scale and it is performed mainly based on the frequency of vibration. Though sounds vary in pitch, certainly not all sounds have pitch. While speaking or singing, some of the sounds has a strong pitch sensation (e.g., vowels) and some have not (e.g., most consonants). Hence, the categorization of the sounds into classes, namely, pitched and unpitched, is valuable in applications like music transcription, speech coding and query by humming.

As the pitch differentiates the user's speech greatly, the pitch and its strength are suggested to extract as the feature for the proposed speaker recognition system. For the feature extraction, the continuous speech signals of all the users are converted to discrete speech signals as $A_{ij}(n)$; $n = 0,1,2,\cdots,N_l$.The speech samples of every user are processed one by one. So, the first sample of the first user is subjected to the process of extracting the pitch and pitch strength from the speech sample. The process is initiated by selecting windows of signal instant with different sizes. So, N_c classes of windows are generated in which each class has their own window size. The window of signal sequences are then processed by Hanning window function as follows

$$W_{kl}(m) = 0.5\left(1 - \left(\cos\left(\frac{2\pi\, w_{kl}(m)}{|w_k| - 1}\right)\right)\right) \qquad (1)$$

where, $0 \le m \le |w_k| - 1$, $0 \le k \le N_c - 1$, $0 \le l \le N_w^{(k)} - 1$, N_w is the number of windows belongs to the k^{th} class, $w_{kl}(m)$ is the m^{th} instant of time in the l^{th} window of the k^{th} class, $W_{kl}(m)$ are the Hanning window coefficients determined from the Eq. (1). The size of the windows that belongs to each class is calculated as $|w_{kl}| = 2^{k+2}$. A pitch vector $[P_{kl}]$ $P_{kl}(m) \in \{0,1\}$ is generated with the same size of $W_{kl}(m)$ in which each element of the vector $P_{kl}(m)$ is arbitrarily generated from {0, 1} (i.e. $P_{kl}(m) \in \{0,1\}$). For every window element, centroids for pit-ched and unpitched classes are determined as follows

$$G_{P_{kl}} = \frac{1}{\sum_{m=0}^{|W_{kl}|-1} W_{kl}(m) P_{kl}(m)} \sum_{m=0}^{|W_{kl}|-1} a_{00}(W_{kl}(m))\, W_{kl}(m) P_{kl}(m) \qquad (2)$$

$$G_{UP_{kl}} = \frac{1}{\sum_{m=0}^{|W_{kl}|-1} W_{kl}(m)\overline{P_{kl}(m)}} \sum_{m=0}^{|W_{kl}|-1} a_{00}(W_{kl}(m))\, W_{kl}(m)\overline{P_{kl}(m)} \tag{3}$$

The centroids determined for pitched and unpitched class using Eq. (2) and Eq. (3) respectively are based on the pitch vector, window coefficients and the speech signal. In Eq. (2) and Eq. (3), $a_{00}(W_{kl}(m))$ is the magnitude of the speech signal at a specific time interval indicated by the window element $W_{kl}(m)$.Once the centroids are calculated, the time instant at which the pitch is present $\{P_I\}$ and the strength of the pitch $\{P_s\}$ are determined as $\{P_I\} = \{P_I^{'}\} - \phi$ and $\{P_s\} = \{P_s^{'}\} - \phi$ respectively, where, the set $\{P_I^{'}\}$ and $\{P_s^{'}\}$ are calculated as

$$\{P_I^{'}\} << \begin{cases} n \;; & \dfrac{2(a_{00}(n) - G_{UP}^{\max})}{G_P^{\max} - G_{UP}^{\max}} > 1 \\ \varphi \;; & otherwise \end{cases} \tag{4}$$

$$\{P_s^{'}\} << \begin{cases} a_{00}(n) \;; & if \;\; n \in \{P_I\} \\ \varphi \;; & otherwise \end{cases} \tag{5}$$

In Eq. (4), $G_P^{\max}$ and $G_{UP}^{\max}$ is a centroid pair that exhibits maximum distance among all the centroid pairs. The $G_P^{\max}$ and $G_{UP}^{\max}$ is determined by firstly calculating the distance between every centroid pairs as $d_{kl} = G_{P_{kl}} - G_{UP_{kl}}$. Then, $G_{P_{kl}}$ and $G_{UP_{kl}}$ that contributes for maximum d_{kl} is obtained as $G_P^{\max}$ and $G_{UP}^{\max}$, respectively. The determined feature set, $\{P_I\}$ and $\{P_s\}$ is stored for the first sample of the first user. The process is repeated for all the speech samples of the same user and the obtained feature set is stored. The similar process is performed for all the N_u users and stored in the feature database. The obtained feature set for all the speech samples of every user is combined to form feature matrix. In the feature matrix P_{ab}; $0 \le a \le P_{\max} - 1$, $0 \le b \le N_T - 1$, each column is comprised of the elements of the feature set of each speech sample of a single user. Thus obtained feature matrix is of size $P_{\max} \times N_T$, where, $P_{\max}$ is the feature set for a sample which has maximum elements and $N_T = N_u . N_s$. All the remaining feature sets are filled up with zeros to attain the size of $P_{\max}$.

IV. DIMENSIONALITY REDUCTION

A. *Factor analysis*

Factor analysis (FA) is a linear method, based on the secondorder data summaries. First suggested by psychologists, FA as-sumes that the measured variables depend on some unkn-own, and often unmeasurable, common factors.Typical examples include variables defined as various test scores of individuals, as such scores are thought to be related to a com-mon intelligence" factor. The goal of FA is to uncover such relations, and thus can be used to reduce the dimension of da-tasets following the factor model.

The zero-mean p-dimensional random vector x_{px1} with covariance matrix Σ satisfies the k-factor model if

$$x = \Lambda f + u \tag{6}$$

where Λ_{pxk} is a matrix of constants, and f_{kx1} and u_{px1} are the random common factors and specific factors, respectively. In addition, the factors are all uncorrelated and the common factors are standardized to have variance one:

$$E(f) = 0, \; Var(f) = I, \tag{7}$$

$$E(u) = 0, \; Cov(u_i, u_j) = 0 \text{ for } i \neq j, \tag{8}$$

$$Cov(f, u) = 0. \tag{9}$$

Under these assumptions, the diagonal covariance matrix of u can be written as

$$Cov(u) = \Psi = diag(\Psi_{11}, \ldots ; \Psi_{pp}).$$

If the data covariance matrix can be decomposed as

$$\Sigma = \mathbf{\Lambda\Lambda^T} + \Psi, \tag{10}$$

then it can be shown that the k-factor model holds. Since x_i can be written as

$$x_i = \sum_{j=1}^{k} \lambda_{ij} f_j + u_i \;, \quad i=1,\ldots,p, \tag{11}$$

its variance may be decomposed as

$$\sigma_{ii} = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_{ii} \;, \tag{12}$$

where the first part $h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2$ is called the communality and represents the variance of x_i common to all variables, while the second part Ψ_{ii} is called the specific or unique variance and it is the contribution in the variability of x_i due to its specific u_i part, not shared by the other variables. The term λ_{ij}^2 measures the magnitude of the dependence of x_i on the common factor f_j. If several variables xi have high loadings $\boldsymbol{\lambda_{ij}}$ on a given factor f_j, the implication is that those variables measure the same unobservable quantity, and are therefore redundant.

The factor model does not depend on the scale of the variables. However, the factor model also holds for orthogonal rotations of the factors. Given the orthogonal matrix G, given the model (13), the new model

$$x = (\Lambda G)(G^T f) + u, \qquad (13)$$

also holds, with new factors $G^T f$ and corresponding loadings ΛG. Therefore, the factors are generally rotated to satisfy some additional constraints, such as

$$\Lambda^T \Psi^{-1} \Lambda \text{ is diagonal, or} \qquad (14)$$

$$\Lambda^T D^{-1} \Lambda \text{ is diagonal, } D = \mathrm{diag}(\sigma_{11}, \ldots\ldots, \sigma_{pp}), \qquad (15)$$

where the diagonal elements are in decreasing order. There are techniques, such as the varimax method, to rotate the factors to obtain a parsimonious representation with few significantly non-zero loadings (i.e. sparse matrix **Λ**). In many cases, a k-order factor model in (10) provides a better explanation for the data than the alternative full covariance model Var(x) = Σ. In such cases, it is possible to derive parameter estimates Λ and ψ.

Let $\bar{x}$, R, and S denote the sample mean, covariance matrix, and correlation matrix, respectively, of the observed data matrix X. Then, starting with

$$\sigma_{ii} = s_{ii}, \; i=1,\ldots\ldots,p, \qquad (16)$$

and using

$$\Sigma = \Lambda \Lambda^T + \psi, \qquad (17)$$

we obtain

$$\sigma_{ii} = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_{ii} \qquad (18)$$

B. Principal factor analysis

Suppose the data is standardized, so that its covariance matrix is equal to the correlation matrix. To obtain estimates Λ and ψ for the standardized variables, first estimate h_i^2 for i = 1,….., p. Common estimates h_i^2 include the square of the multiple correlation coefficients of the i_{th} variable with all the other variables, and the largest correlation coefficient between the ith variable and one of the other variables. Next, form the reduced correlation matrix $R-\psi$, where the diagonal ele-ments of 1 in R are replaced by the elements $h_i^2 = 1-\psi_{ii}$.

Then, decompose the reduced correlation matrix in terms of the eigenvalues a1≥….≥a_p and orthonormal eigenvectors $\gamma_{(1)},\ldots,\gamma_{(p)}$ as

$$R-\psi = \sum_{i=1}^{p} a_{(i)} \gamma_{(i)} \gamma_{(i)}^T \qquad (19)$$

If the first k eigen values are positive, estimate the i_{th} column of Λ by

$$\lambda_{(i)} = a_i^{1/2} \gamma_{(i)}, \; i=1,\ldots,k. \qquad (20)$$

Equivalently,

$$\Lambda = \Gamma_1 A_1^{1/2}, \qquad (21)$$

where $\Gamma_1 = (\gamma_{(1)},\ldots,\gamma_{(p)})$, and $A_l = \mathrm{diag}(a_1,\ldots a_k)$.The eigenvectors are orthogonal, so the constraint in (22) holds.

Finally, the specific variance estimates are updated as

$$\psi_{ii} = 1 - \sum_{i=1}^{k} \lambda_{ij}^2, \qquad i=1,\ldots p, \qquad (22)$$

The k-factor model is permissible if all the p terms are non-negative.

In practice, the number of factors may be determined by looking at the eigen values a_i of the reduced correlation matrix, and choosing k as the index where there is a sharp drop in the eigen value magnitudes.

As its name suggests, principal factor analysis (PFA) is related to principal component analysis. When the specific variances are all zero, ψ = 0, comparing Equations (10) and (19) indicates that PFA is equivalent to PCA.

C. Maximum likelihood factor analysis

If, in addition to the factor model specified in (6)-(9), we also assume that the factors f and u are distributed as multivariate normal variables, then parameters of the model can also be estimated by maximizing the likelihood. In such cases, one can also test the hypothesis that the k-factor model describes the data more accurately than the un-constrained variance model. The log-likelihood function can be written as

$$l = -\frac{1}{2} n \log|2\Pi\Sigma| - \frac{1}{2} n \; tr\Sigma^{-1} S \qquad (23)$$

and the goal is to maximize it with respect to the parameters Λ and Ψ subject to the constraint in (22) on Λ.

Under the factor model, $\Sigma = \mathbf{\Lambda\Lambda^T} + \Psi$

The optimization is carried out by noting that the function

$$F(\Lambda,\Psi)=F(\Lambda,\Psi;S)=tr\Sigma^{-1}S-\log|\Sigma^{-1}S|-p \qquad (24)$$

is a linear function of the log-likelihood *l*, with a maximum in *l* corresponding to a minimum in F. Also, in terms of the arithmetic mean 'a' and the geometric mean 'g' of the eigenvalues of $\Sigma^{-1}S$, we have

$$F=p(a-\log g-1) \qquad (25)$$

Minimizing $F(\Lambda,\Psi)$ proceeds in two stages: first, the minimization over Λ for a fixed Ψ has an analytical solution, then, the minimization over Ψ is carried out numerically.

V. SVM-BASED RECOGNITION

SVM, a special case of Tikhonov regularization, belongs to the family of general linear classifier, which has a property that the classifier minimizes the classification and improves the geometric margin. In the training process, an error function is to be minimized which can be given as

$$e=\frac{\alpha^T\alpha}{2}+p\sum_{a=0}^{N^{red}-1}\beta_a \qquad (26)$$

with the constraints,

$$\beta_a\geq 1-L_a\left(\alpha^T\phi(M_a)+\gamma\right) \qquad (27)$$

$$\beta_a\geq 0 \qquad (28)$$

In Eq. (26) and Eq. (28), α is a coefficient matrix, p is the penalty factor and β is a parameter that handles the data. In the constraints given in Eq. (27) and Eq. (28), L is the class label of the a^{th} dataset, φ is the kernel that transforms the input data to the feature space and γ is a constant. Once the error function given in the Eq. (26) gets minimized, the training process for the SVM is completed and so the SVM can be applied for recognizing the user's speech.

In recognizing the user's speech, the well-trained SVM is given by speech samples of some users. When a speech sample of user is given to proposed technique, firstly, the pitch and pitch strength for the sample is extracted as feature set (as performed above). The obtained feature set is subjected to MLFA-based dimensionality reduction and then it is given to SVM. The trained SVM recognizes well that the given speech sample corresponds to the valid user or not. When the given speech sample belongs to the trained sample, the SVM outputs that the given speech is from authenticated user, otherwise, the speech is from invalid user.

VI. RESULTS & DISCUSSION

The proposed speaker recognition technique is implemented in the research tool, MATLAB of version 7.10. The technique is tested using a speech database obtained from the Neurosciences Institute through online. In the database, both male and female users with different speech samples are available. For training, twenty speech samples for every user are used and ten samples are used for testing the recognition technique. Also, for testing, an invalid user with forty speech samples is used. Once, the technique has been tested, the performance of the technique is evaluated by determining the performance metrics, accuracy, false rejection rate (FRR) and false acceptance rate (FAR). The performance of the proposed speaker recognition technique is compared against the speaker recognition technique using MPCA and PFA. The perfo-rmance measures of the speaker recognition using proposed technique, MPCA and PFA are given in Table I.

Table I: Performance comparison between the proposed speaker recognition technique with MPCA-based recognition technique and PFA-based recognition technique

Recognition Technique	Speakers		% of Recognition Accuracy	
	Male	Female	Male	Female
MLFA	30	30	98.7001	98.7113
	20	20	98.6781	98.6991
	10	10	98.6715	98.6766
PFA	30	30	98.6001	98.6635
	20	20	98.4956	98.5403
	10	10	98.3551	98.4226
MPCA	30	30	98.1779	98.2779
	20	20	97.2551	97.6224
	10	10	95.5661	95.9111

From the obtained results, it can be seen that the performance of the proposed technique is better than the performance of the other two techniques. This shows that the proposed technique effectively recognizes the speaker whether he/she is authorized or not.

VII. CONCLUSION

In this paper, we have proposed an effective speaker recognition technique. It uses the pitch strength of the speech signal as features for identifying the user. The obtained high dimensional feature set was reduced with the aid of MLFA and the recognition was performed using SVM. First, the SVM was trained well by the low dimensional feature set, which was obtained from the training dataset. After the training was completed, we tested the SVM with a few authorized and unauthorized user speeches. The test results showed that the proposed technique correctly recognizes the authorized and unauthorized users from their speech. The proposed technique exhibits remarkably high correct classification rate, low false rejection rate and low false acceptance rate. A comparison has also been made between the proposed recognition technique, the MPCA-based recognition technique and the PFA-based recognition technique. From the comparison result, it was seen that the proposed technique outperforms the other recognition techniques.

ACKNOWLEDGEMENT

The authors would like to thank Sri S.Ramabrahmam and Dr. G.V.K.S.V.Prasad from URCE for making Matlab demo available.

REFERENCES

[1] T. Santhanam, M. Nachamai, M. Muthuraman and C.P. Sumathi, "Stressed/ Neutral Speech Classification using Gaussian Support Vector Machines", in proceeding of International Journal on Soft Computing, vol. 2, no. 2, pp: 335- 338, 2007

[2] Saeid Rahati, Quchani and Kambiz Rahbar, “Discrete Word Speech Recognition using Hybrid Self-adaptive HMM/SVM Classifier”, Journal of Technical Engineering, Vol.1, No. 2, p.p. 79-90, 2007

[3] TI Modipa, HJ Oosthuizen and MJD Manamela, "Automatic Speech Recognition of Spoken Proper Names", in proc. of Southern African Telecommunication Networks and Applications Conference, 9- 13 September 2007.

[4] A.Revathi, R.Ganapathy and Y.Venkataramani, "Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach", in proc. of International Journal on Computer science & Information Technology (IJCSIT), vol. 1, no. 2, p.p. 30-42, November 2009.

[5] Dimo Dimov and Ivan Azmanov, "Heuristic Improvements of the HMM Use in Isolated Word Speech Recognition", in proc. of Cybernetics and Information Technologies, vol. 7, no. 3, pp. 73-88, 2007

[6] Akram M. Othman, and May H. Riadh,"Speech Recognition using Scaly Neural Networks", in proc. of Intl. Journal on Computer Systems Science and Engineering, Vol. 3, no. 2, p.p. 253-258, 2008

[7] J.E.Higgins, R.I. Damper, “An HMM-based subband processing approach to speaker identification”, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Audio- and Video-Based Biometric Person Authentication, Pages169-174, 2001.

[8] Nick J.-C. Wang, Wei-Ho Tsai1, and Lin-Shan Lee, “Eigen-MLLR Coeffcients as New Feature Parameters for Speaker Identification”,

[9] Muzhir Shaban Al-Ani, Thabit Sultan Mohammed and Karim M. Aljebory, “Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform”, Journal of Computer Science 3 (5): 304-309, 2007, ISSN 1549-3636

[10] Ching-Tang Hsieh, Eugene Lai and You-Chuang Wang, “Robust Speaker Identification System Based on Wavelet Transform and Gaussian Mixture Model”, Journal of information Science and Engineering, Vol. 19, 267-282, 2003.

◈◈◈

Computer Aided Analysis of Lung CT Based on Transform Domain Features

V.Minnal

Department of Information Technology, MIT, Anna University, Chennai, India
E-mail : minnal.v@gmail.com

Abstract **-** As Many CADx systems have been developed to detect lung cancer based on spatial domain features that process only the pixel intensity values, the proposed scheme applies frequency transform to the lung images to extract frequency domain features and they are combined with spatial features so that the features that are not revealed in spatial domain will be extracted and the classification performance can be tuned up. The proposed CADx comprises of four stages. In the first stage, lung region is segmented using Convexity based active contour segmentation. At second stage ROIs are extracted using spatially constrained KFCM clustering. Followed by standard wavelet transforms is applied on ROI so that transform domain features are extracted with shape and haralick histogram features. Finally neural network is trained by combined feature set to identify the cancerous nodules. Our proposed scheme has shown sensitivity of 95% and specificity of 96%.

Keywords **-** *convexity; active contour; KFCM; wavelet features;neural network.*

I. INTRODUCTION

Recently computer programming plays a vital role in solving most of the problems in medical applications such as health monitoring, intensive surgery due to their fastness and accuracy. Especially image processing techniques are considered to be mostly preferred in diagnosing abnormalities on medical images like X-ray, CT. Computer Aided Diagnosis is the procedure that involves imaging technologies followed by the image processing techniques where as the various images of human body obtained by imaging technologies are processed to identify the illness or abnormalities if present. If CADx developed devoid of drawbacks then it would be a perfect second opinion to the radiologists in medical image interpretation. In last two decades research on development of CADx has grown exponentially, and review of that survey is shown in next section.

As various technologies emerging, several works have been done to identify lung cancer from CT images. Lung cancer diagnosis has been an active research domain since 1980's. In literature, very recent advanced works in diagnosing lung cancer has been reviewed. The development of new schemes for image acquisition, such as high resolution CTs, has improved the detection and diagnosis of cancer significantly.

Even if the research on automated computer aided diagnosis of lung cancer has been started since 1990's list of approaches have been proposed since 2006. In 2006, Antonelli *et al* [1] have proposed Lung Nodule Detection scheme based on an Anatomical Model and a Fuzzy Neural Network. He used anatomical model of lung to segment lung parenchyma from the CT image and fuzzy Neural Network to classify by using the fuzzy set of features. Adrien *et al* [2] in 2007 have done a scheme of Lung Tissue Classification Using Wavelet Frames which has achieved multiclass accuracy of 92.5%. In 008 Liu lu *et al* have proposed to use SVM to detect the pulmonary nodules with the sensitivity of 90% and specificity of 70%. During 2009 jaffar *et al* put forward Fuzzy morphology and FCM classification to identify nodules from CT images. In 2010, lafari *et al* has proposed Automated Detection of Pulmonary Nodules in 3D Thoracic CT Images with the FP of 10.3 and sensitivity of 88%. In same era Zhang *et al* [] have published the diagnosis system of lung cancer by using different classes of features by the sensitivity of 99% and specificity of 86%.those classes of features included 1st, 2nd and higher order features along with the structural features. Tong *et al* have proposed nodule detection scheme based on Rule Based Classification that achieved Overall detection rate 85%. In 2011, Hua *et al* [3] have developed Graph-Search Algorithm to segment the lung parenchyma and KNN classifier for lung tissue classification which have given Sensitivity of 98.6% and Specificity of 99.8%. In 2011, ying *et al* has proposed Autonomous Detection of Solitary Pulmonary Nodules in which Quantification of feature parameter is done using SVM with the 94.6% of sensitivity. In same year, Aravind Kumar *et al* have developed Robust and Automated Lung Nodule Diagnosis from CT Images based on fuzzy inference system (FIS). This achieved Classification accuracy of 90% and sensitivity of 86%.

In this proposed approach, CADx consisting of four phases as follows. The lung parenchyma is segmented from background by using convexity based Active Contour segmentation. ROIs are then identified and extracted using spatially constrained Kernel FCM clustering. Followed by shape based and histogram features are extracted from ROIs. Transform domain features are extracted by applying daubechies wavelets which are known as standard wavelets and then finally neural network is trained by combined features to identify the cancer nodules.

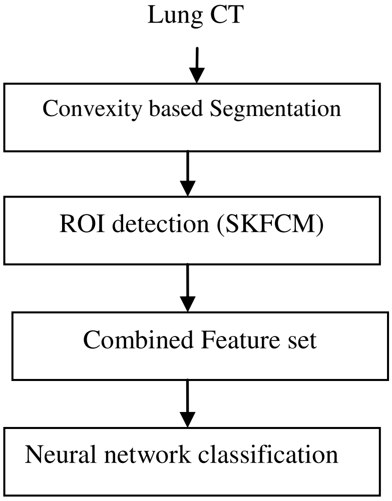


Fig.1: The proposed architecture

II. LUNG SEGMENTATION

Segmentation is considered to be the essential first step in Computer analysis. Lung is a container of air in human body and hence it visible as darker in CT scan. This helps in segmenting out the lung region from its background. There are various image segmentation techniques available such as Threshold based, region based and edge based methods. Here region based segmentation method ‘snakes’ which are termed as Active Contour Model is used to segment both the lungs.

A. Localized Active Contour segmentation

The traditional snakes are described as the structure that enlarge or shrinks so that they can exactly overlap the object contour by minimizing its energy. The snake reshapes according to the minimization of global energy which works well when the object is of homogeneous intensity. Here in lung CT, as lung will have heterogeneous intensity, standard global energy method will not produce good segmentation results.

In Localized Active Contour Segmentation [4], foreground and background are described in terms of local regions. These regions are analyzed to construct local energies. In order to achieve localized region segmentation global energies can be localized such as mean separation energy, and histogram separation energy. In our work mean separation energy is included to the energy terms of region.

$$E_{MS} == \int_{\Omega y} (u-v)^2 \qquad (1)$$

u, v are the means of foreground and background. Respective F is calculated as

$$F_{MS} = (u_x - v_x)^2 \qquad (2)$$

In some cases where the lung nodules are at the chest wall, as the nodules and chest wall regions are having nearly same intensity values, snakes are intended to segment the lung with elimination of nodules located near chest so that those nodules will be left out in ROI extraction phase. This drawback can be overcome by using the convexity of lung region as follows.

B. Convexity Based segmentation

Hence the human’s left and right lungs are symmetric in shape they will have almost same convexity. Normally snakes are shrinking or extending until the specified iteration threshold. Where as in convexity based segmentation, after segmented the normal size lung and then the area based convexity is calculated using the equation

$$convexity = 1 - \frac{Area_{diff}}{Area_{lung}} \qquad (3)$$

$Area_{diff}$ is the area of region that the lung differs from its equivalent convex hull.

From the analysis of various normal lung CT slices, it confirmed that the maximum difference in area between left and right lung is 20%.when this difference exceeds this limit, then it is assumed that the smaller lung must have nodules that are attached to the chest wall. Then the snakes are implemented on these small lungs until the convexity of the region covered by active contour match the convexity of larger one.

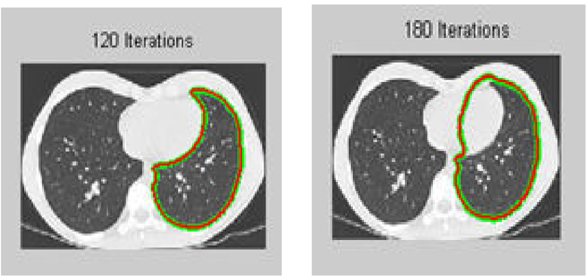


Fig. 2 : Difference between traditional and convexity based Contour Segmentation

Figure 1 shows how the convexity based Contour differently segment the small lungs when they considered as their nodules have been attached to the chest walls.

III. SPATIALLY CONSTRAINED KFCM

ROI detection has to be carried out to reduce the participation of irrelevant data for nodule identification and classification. Fuzzy Clustering method is mostly preferred in ROI detection since its capability to process multi dimensional information.FCM technique is also posses low sensitive to noise. It is a centroid based clustering in which the image pixels having intensities same or close to the centroid of one cluster will be associated with higher membership level than that to the other cluster.

A. Kernelized FCM

KFCM [5] is the enhancement of classical Fuzzy C-Means clustering by the incurring input dataset. This method maps the data from data or feature space $\Xi \subseteq Rp$ to the much higher dimensional space *H* (Hilbert or Kernel space).this mapping is done by the transform function □: $\Xi \rightarrow H$.

The kernel function here is used to simulate the distances that would be obtained by transferring the points to a higher dimensional space. As the data in higher dimensional space exhibiting clear and simple structures it will be resulted in effortless clustering by FCM. The well known kernel function is Gaussian radial basis function.

$$K(x,y) = \exp(\frac{\|x-y\|^2}{\sigma}) \tag{4}$$

Kernel functions convert non linear problems in input domain to linear problems in frequency domain.

B. Spatial FCM

This approach [6] utilizes the spatial coefficient information which is a sum of neighborhood information of each pixel. Considering the spatial function beneficial in improving the noise sensitivity and even more homogeneous regions can be figured out. At first step, initial membership function for each pixel is computed as follows

$$h_{ij} = \sum_{k \in NB(x_j)} u_{jk} \tag{5}$$

Where NB (x_j) represents the resizable neighborhood window of pixel x_j. After that the new membership function is formed by using this spatial function.

C. Advanced SKFCM

This advanced technique [7] combines both the kernel method and spatial function. This method applies the spatial functions to the data in higher dimensional space for identifying new centroid of cluster. Initially the Kernel membership function is computed as follows

$$u_{jk} = \frac{(1/d^2(x_j,V_k))^{\frac{1}{m-1}}}{\sum_{j=1}^{C}(1/d^2(x_j,V_k))^{\frac{1}{m-1}}} \tag{6}$$

Where

$$d^2(x_j,V_k) = K(x_j,x_j) - 2K(x_j,V_k) + K(V_k,V_k) \tag{7}$$

m is the fuzzy index which decides the fuzziness of the clusters. In next step spatial function is applied to find the membership factor.

$$K(x_j,\widehat{V}_k) = \frac{\sum_{i=1}^{N}(u_{ik})^m K(x_i,x_j)}{\sum_{i=1}^{N}(u_{ik})^m} \tag{8}$$

The iteration process will be continued until the error lies above the determined threshold. Or else a new iteration will be started.

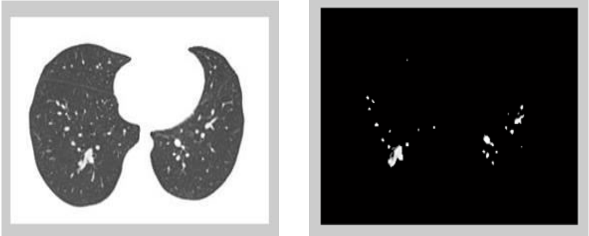


Fig. 3 : ROI Extraction using SKFCM

IV. FEATURE EXTRACTION

The working performance of computer aided diagnosis system is mainly depends on the feature set used in it. In this proposed approach, both the spatial and frequency domain features are extracted for each region of interest.

A. Spatial Features

These are the features computed by the pixel values in the image. In many cases the spatial feature set is enough to identify the required object. There are various classes of features in spatial domain as follows.

1. *Geometric Features*: - These features are describing the shape and structure of the regions directly. From this category, area, perimeter, equivalent diameter, major axis length, minor axis length, eccentricity such features are calculated for each of the region.
2. *Histogram Feature:* - These features are derived from the histogram information that presents in image regions. There are various features involved with histogram information. Haralick features such as mean, variance, standard deviation, skewness, energy and kurtosis are computed by using the histogram information.

B. Wavelet Features

Wavelets transform has obtained a huge importance in image processing methodology for its multi-resolution representation of image data and its transformation of data in both time and frequency. When the image is subjected to the wavelet transform, it produces multi resolution versions of input image. In practical wavelet transform decomposes the image into Approximation and detailed coefficients.

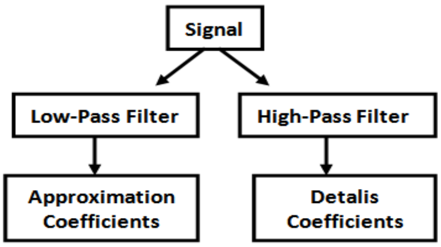


Fig. 4 : Wavelet Decomposition

Approximation coefficients will include low frequency components and detailed coefficients will include high frequency components. Wavelet transform gets the input image and transform into multiple frequency bands [10]. The standard wavelet mostly used in image feature extraction is daubechies wavelet. From db2 to db20 are the mostly used from daubechies wavelet family.

By mentioning the level of decomposition the approximation coefficients can be further decomposed into LL, LH, HL, HH frequency components. Here db4 decomposition coefficients have been given.

TABLE 1 : DB4 COEFFICIENTS

NO	LOW PASS	HIGH PASS
1	-0.129409	-0.48296
2	0.22414	0.83651
3	0.83651	-0.22414
4	0.48296	-0.12940

Thus the wavelet transform develops four matrices of various frequency bands. From that low and high frequency components the features such as vertical mean, horizontal mean and energy are computed and they are combine with the spatial features to train the neural network.

V. LUNG NODULE CLASSIFICATION

The main objective of CADx is to recognize the nodules by their features and it cannot be done efficiently without prior knowledge about the distinct features of the cancerous nodules from the linear tissue structures. There are so many nodule detection schemes and classification methods have been developed.

There are two Major categories in classification as supervised, unsupervised and regression classification techniques. Regression method is the most advanced and have better efficiency. Regression technique includes linear regression, Gaussian regression and neural networks.

A. Artificial Neural Network

Neural networks resemble the circuit of biological neurons which are composed of artificial neurons. ANN is composed of interconnecting neurons. Both networks are similar in their tasks. The biological neurons have the aim of solving particular tasks, while the ANN aims to build mathematical models of biological neural systems. A simple neural network will be constituted of three parts as input layer, output layer and hidden layer.

Back propagation is the most widely used technique to train neural networks. It is a supervised learning method, and is a generalization of the delta rule. It requires a dataset of the desired output for many inputs, making up the training set. It is most useful for feed-forward networks which doesn't have connection that loop. Here in this project all importance have been given to the feature extraction which is sufficient for nodule recognition the basic back propagation algorithm [8] is used to train our neural network.

Back Propagation Algorithm:

```
Initialize random weights to the networks
Do
  For each val in the training set
       Out=nn_output (network, val);
       T=teacher output for val;
       Find error (T-out) at output units;
       Compute del_wh for all weights from
       hidden layer to output layer; (bwd pass)
       Compute del_wh for all weights from
       input layer to hidden layer ;( bwd pass)
       Update weights in n/w;
Until all values classified or criterion have met
Return the network
```

Back propagation computes the gradient of the error of the network regarding the network's random weights. This gradient is almost always used in a simple stochastic gradient descent algorithm to find weights that minimize the error. Often the term back propagation is used in a more general sense, to refer to the entire procedure encompassing both the calculation of the gradient and its use in stochastic gradient descent. After the input pattern was presented to the network and processed by all layers, we have errors.

Our testing set contains 400 samples in which 200 samples include nodules in it and the remaining are normal. The output of this network is 1 if the input image contains nodule, otherwise 0.

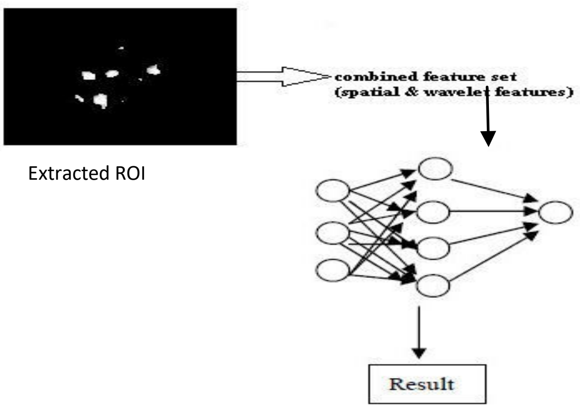


Fig. 5 : Nodule classification

VI. RESULTS AND DISCUSSION

The efficiency of CADx system is described by its sensitivity and specificity measures. The well designed system should identify the true positive regions more that analogous to the expert radiologists. True positive means recognize the affected region and mark it as affected. And whereas the false positive is known as mark wrongly the normal region as the affected one. The FP rate should be very less to provide better diagnosis results.

In our project, our analysis is based on high resolution lung CT images taken from Lung Image Database Consortium (LIDC), which are accessed through the National Biomedical Imaging Archive (NBIA). Each image was annotated by four experts, at first during a blind revision, then by communicating the discrepancies that were found and asking for their correction. The studies are stored in the DICOM standard, with size 512x512 pixels and a grey scale of 12 bits in Hounsfield Units (HU). Each case is associated with the XML document which contains the details about nodules present.

A neural network with 500 neurons and input layer of 25 neurons which accommodate 25 values of combined feature set and corresponding two output layers (to classify nodules either affected or normal) has been used at training stage.

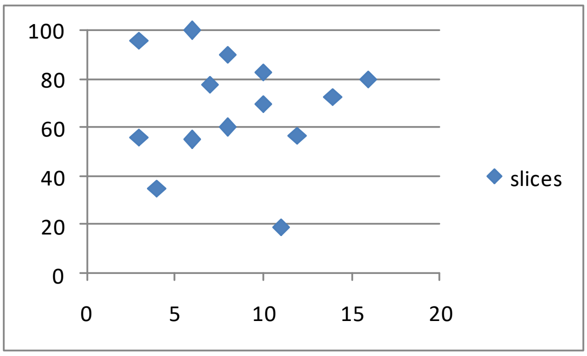


Fig. 6 : ROC plot for classification nodules

The above figure 6 shows that the SKFCM technique gives the low index of false positive. In graph, slice 13 got high FP with low TP rate. It should be because of the main part of nodule might be removed during preprocessing.

The proposed scheme has been experimented by randomly selected 15 slices from LIDC dataset (nodule's information known). The result shown that proposed system detecting nodules with the sensitivity of 94%. Micro nodules of less than 3 mm have been challenge and are not marked as nodules. The regions which are with the diameter of less than 5 mm will posses very less possibility of malignancy [9] and hence they can be avoided as non-nodules.

VII. CONCLUSION

In this paper, we have proposed multi stage Computer Aided Diagnosis scheme which segment not even the lung but as well as the nodules that attached to the chest wall by using convexity information of lung region into the Active contour segmentation. ROI's have been extracted by advanced spatially constrained Kernel FCM clustering. This technique processes the spatial neighborhood information in higher dimensional space. The combined feature set was extracted and used in neural network training by back propagation algorithm. This feature set included geometrical, histogram based and wavelet features. The ROI detection scheme in this work works well when we are using three clusters to separate Region of our interest. As 20 cases selected from LIDC collection were used to test the system and it achieved sensitivity of 94% with high TP rate. In future more advanced cancer diagnosing systems can be designed to recognize nodules with much more accuracy with the help of optimal feature set contains both spatial and frequency domain features.

REFERENCES

[1] M. Antonelli, G. Frosini, B. Lazzerini, "A CAD system for Lung Nodule Detection based on Anatomical model and Fuzzy Neural Network," Fuzzy Information Processing Society, NAFIPS 2006, pp. 448-453, June 2006.

[2] Depeursinge A, Saga D, Hidki, "Lung Tissue Classification Using Wavelet Frames", 29th Annual International Conference on Engineering in Medicine and Biology Society , pp. 6259 – 6262, Aug 2007

[3] Panfang Hua, Qi Song, Sonka M, Hoffma E, Reinhardt M, "Segmentation Of Pathological And Diseased Lung Tissue In CT Images Using A Graph-Search Algorithm", International Symposium on Biomedical Imaging: From Nano to Macro, pp. 2072 – 2075, April 2011.

[4] Shawn Lankton, Allen Tannenbaum," Localizing Region-Based Active Contours", IEEE Transactions On Image Processing, Vol. 17, NO. 11, pp. 2029-2039, November 2008.

[5] Zhong-dong Wu, Wei-xin Xie, Jian-ping Yu, "Fuzzy C-means clustering algorithm based on kernel method", International Conference on Computational Intelligence and Multimedia Applications, pp. 49-54, September 2003.

[6] Keh-Shih Chuang, Hong-Long Tzeng, Sharon Chen, "Fuzzy c-means clustering with spatial information for image segmentation", Computerized Medical Imaging and Graphics, pp. 9-15, September 2005.

[7] Alberto Rey, Bernardino Arcay, Alfonso Castro, "Study on Various Defuzzification Methods for Fuzzy Clustering Algorithms to Improve ROIs Detection in Lung CTs", pp. 2123-2130, June 2011.

[8] http://en.wikipedia.org/wiki/Backpropagation

[9] Jin Mo Goo, "Computer-Aided Detection of Lung Nodules on Chest CT: Issues to be Solved before Clinical Use", Korian J Radiol, pp. 62-63, June 2005.

[10] Domingues Barros, Gonçalves da Silva, "Characterization of Bone Tissue by Microwaves Using Wavelets and KNN", Journal of Microwaves, Optoelectronics and Electromagnetic Applications,Vol. 10, No. 1, pp. 217-231, June 2011.

◈◈◈

Improve the Processing Performance of Top-K Spatial Keyword Query on Road Networks

M. Sreenath Yadav & K.R.K.Satheesh

Dept. of CSE, Madanpalle Institute of Technology and Science, Madanpalle, Andhra Pradesh, India
E-mail : rg.mech@rmkec.ac.in, sridhar0410@yahoo.com, kesavanchandrasekaran@gmail.com

Abstract - In this paper we address, for the first time, the challenging problem of processing top-k spatial keyword queries on road networks where the distance between the query location and the spatial object is the shortest path. For example, consider a road network database with available paths for two points. A customer may want to rank the paths with respect to their distance defined after aggregating the qualities of other features (road location, road condition, travelling time, and traffic monitoring and road length) within a distance range. A top-k spatial keyword query returns the k best objects ranked in terms of both distance to the query location and textual relevance to the query keywords. We formalize the new query type, and present novel indexing structures and algorithms that are able to process such queries efficiently.

Keywords - *Top-k spatial keyword queries, Group Probing, indexing, Query processing, spatial databases,*

I. INTRODUCTION

Top-k spatial keyword queries return the k best spatio-textual objects ranked in terms of both spatial proximity to the query location and textual relevance to the query keywords. Despite the wide range of location-based applications that can benefit from these queries, the current approaches for processing top-k spatial keyword queries are restricted to the Euclidean distance [3, 12, 18]. In this paper, we address, for the first time, the challenging problem of processing top-k spatial keyword queries on road networks. Given a set of spatio-textual objects (e.g., restaurants annotated with a text), a query location (latitude and longitude), and a set of query keywords, a top-k spatial keyword query on road networks returns the k best objects in terms of both 1) shortest path to the query location, and 2) textual relevance to the query keywords.

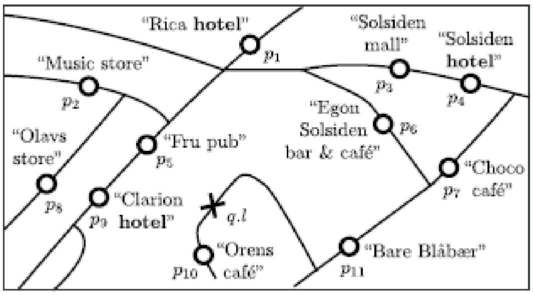


Figure 1: Top-k spatial keyword query on road networks.

For example, Figure 1 illustrates the road networks and spatiotextual objects in a tourist area of Trondheim, Norway. The circles represent spatio-textual objects p with a textual description, and the cross mark *q.l* represents the query location. Assume a tourist in q.l with a GPS-enabled mobile phone.

The tourist poses a top-k spatial keyword query looking for "hotel" (his spatial location is automatically sent by the mobile phone). If a traditional query (Euclidean distance) is considered, the top-1 hotel is p9 on the left side of the figure. However, when road networks are considered, the top-1 hotel is p4 on the right side of the figure. In top-k spatial keyword queries on road networks both shortest path and textual relevance are considered. For example, for the query "bar café" posed in q.l, the spatio-textual object p6 may appear better ranked than p7 because the description of p6 ("Egon Solsiden bar & café") is more textually relevant to the query keywords than the description of p7 ("Choco café"), and p6 is only slightly more distant to *q.l* than p7. The top-1 object, however, is p10 because it is very near to *q.l* and is also relevant to the query keywords.

Note that p11 is not returned as a result of this query, since none of the terms in the description of p11 appear in the query keywords. Top-k spatial keyword queries on road networks can be employed by location-based applications to provide a more precise and realistic result. However, processing these queries is costly, since it requires computing several shortest paths. To the best of our knowledge, processing top-k spatial keyword queries on road networks has never been proposed before.

In this paper, we formalize the concepts of this new query type and describe

How to rank objects considering both the network distance and the textual relevance. We also propose a basic approach for processing top-k spatial keyword queries on road networks combining. The state-of-the-art approaches for road network and spatiotextual indexing. Finally, we describe how to improve the query processing performance In order to identify the relevant regions; we compute an upper bound score for any object in the region in terms of both minimum network distance to the query location and maximum textual score. The maximum textual score of any object in a region is obtained through an abstract textual representation that is maintained for each region.

In summary, the main contributions of this paper are:

- We introduce top-k spatial keyword queries on road networks.
- We describe a basic approach for processing top-k spatial keyword queries on Road networks combining state-of-the art techniques.
- We propose an enhanced approach that indexes the description of the objects on

A segment of the road network for efficient query processing.

II. LITERATURE REVIEW:

Object ranking is a popular retrieval task in various applications. In relational databases, we rank tuples using an aggregate score function on their attribute values [3]. For example, a real estate agency maintains a database that contains information of flats available for rent. A potential customer wishes to view the top 10 flats with the largest sizes and lowest prices.

In this case, the score of each flat is expressed by the sum of two qualities: size and price, after normalization to the domain [0, 1]. In spatial databases, ranking is often associated to nearest neighbor (NN) retrieval. Given a query location, we are interested in retrieving the set of nearest objects to it that qualify a condition. Assuming that the set of interesting objects is indexed by an R-tree [4], we can apply distance bounds and traverse the index in a branch-and-bound fashion to obtain the answer [5].

Nevertheless, it is not always possible to use multidimensional indexes for top-k retrieval. First, such indexes break down in high-dimensional spaces [6], [7]. Second, top-k queries may involve an arbitrary set of user-specified attributes from possible ones and indexes may not be available for all possible attribute combinations because they are too expensive to create and maintain. Third, information for different rankings to be combined for different attributes could appear in different databases in a distributed database scenario and unified indexes may not exist for them.

Solutions for top-k queries [8], [3], [9], [10] focus on the efficient merging of object rankings that may arrive from different sources. Their motivation is to minimize the number of accesses to the input rankings until the objects with the top-k aggregate scores have been identified. To achieve this, upper and lower bounds for the objects seen so far are maintained while scanning the sorted lists. In the following sections, we first review the R-tree, which is the most popular spatial Access method and the NN search algorithm of [5]. Then, survey our feature-based spatial queries.

2.1 Spatial Query Evaluation On R-Trees

The most popular spatial access method is the R-tree [4], which indexes minimum Bounding rectangles (MBRs) of objects. Fig. 2 shows a set D ={p1 ... p8} of spatial objects and an R-tree that indexes them. R-trees can efficiently process main spatial query types, including spatial range queries, nearest neighbor queries, and spatial joins. Given a spatial region W, a spatial range query retrieves from D the objects that intersect W.

For instance, consider a range query that asks for all objects within the shaded area in Fig. 2. Starting from the root of the tree, the query is processed by recursively following entries, having MBRs that intersect the query region. For instance, e1 does not intersect the query region, thus the sub tree pointed by e1cannot contain any query result. In contrast, e2 is followed by the algorithm and the points in the corresponding node are examined recursively to find the query result p7. A nearest neighbor query takes as input a query object q and returns the closest object in D to q.

For instance, the nearest neighbor of q in Fig. 2 is p7. Its generalization is the k-NN query, which returns the k closest objects to q, given a positive integer k. NN (and k-NN) queries can be efficiently processed using the best-first (BF) algorithm of [5], provided that D is indexed by an R-tree. A min-heap H which organizes R-tree entries based on the (minimum) distance of their MBRs to q is initialized with the root entries. In order to find the NN of q in Fig. 2, BF first inserts to H entries e1, e2, e3, and their distances to q. Then, the nearest entry e2 is retrieved from H and objects p1, p7, p8 are inserted to H. The next nearest entry in H is p7, which is the nearest neighbor of q. In terms of I/O, the BF algorithm is shown to be no worse than any NN algorithm on the same R-tree [5].

The aggregate R-tree (aR-tree) [11] is a variant of the R tree, where each non leaf entry augments an aggregate measure for some attribute value of all points in its sub tree. As an example, the tree shown in Fig. 2 can be upgraded to a MAX aR-tree over the point set, if entries e1,e2,e3 contain the maximum measure values of sets {p2,p3},{p1,p8,p7} ,{p4,p5,p6} respectively. Assume that the measure values of p4, p5, p6 are 0.2, 0.1, 0.4, respectively. In this case, the aggregate measure augmented in e3 would be max {0.2, 0.1,0.4} = 0.4. In this paper, we employ MAX aR-trees for indexing the feature data sets, in order to accelerate the processing of top-k spatial preference queries.

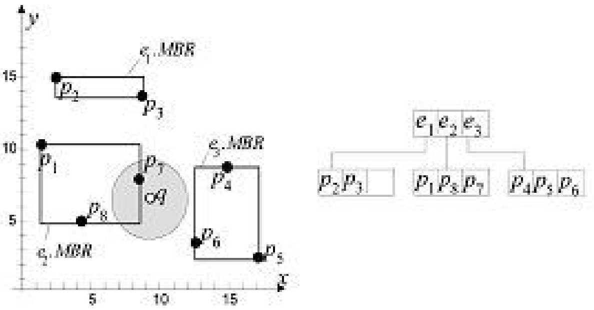


Fig. 2: Spatial queries on R-trees.

III. PRELIMINARIES:

In this section, we define the model used to represent the road networks, the set of objects of interest, and the problem statement.

Road networks. Here Road network model is treated as a graph G=(V,E,W)

Where V is the set of vertices, E is the set of edges, and W is the set of weights (network distances) that are associated with each edge. A vertex v ∊ V represents a road intersection or an end-point in the road network, an edge (v, v0) ∊ E represents a road segment, and the weight w ∊ W associated with each edge (v, v0) represents the length (network distance) |v, v0| of the road segment. For simplicity, we assume bidirectional traffic. However, unidirectional traffic is also support by our approach. In this case, (v, v0) 6≠ (v0, v) and the distance |v, v0| may be different from |v0, v|.

Objects of interest. We assume a set of spatio-textual objects p ∊ P on the edges E of the road network G. Each object p has a spatial location p.l and a textual description p.d. The size (cardinality) of the set of objects P is denoted as |P|. The network distance between an object p and the ends of the edge (vertices) in which it belongs is defined as |v, p| or |v!, p|, where v and v1 are the ends of the edge (v, v!) where p lies. The shortest path between two objects p and p1 in the road network graph G is defined as ||p, p1||. The set of objects in the edge (v, v1) and (v1, v) are the same, and the distance |v, p| is equal to the distance |v, v1| − |v1, p|. Therefore, knowing the distance between p and one vertex is sufficient to obtain the distance between p and the other vertex. We denote reference vertex, the vertex used to compute the distance to the objects on the edge. Note that if unidirectional traffic is considered, the set of objects lying on edge (v, v1) is different from the set of objects lying on the edge (v1, v), and the distance |v, p| may be different from |v, v1| − |v1, p|.

Problem statement. We define a top-k spatial keyword query on road network as $Q_N = \langle q.l, q.d, q.k \rangle$ where q.l is the query spatial location (latitude and longitude), q.d is the set of query keywords, and q.k is the number of results of interest. Without loss of generality, we assume that the query location q.l lies on an edge of the road network. This assumption can be bypassed finding the nearest edge of a given query location. Given a queryQ_N, a road network G, and a set of spatio-textual objects P; Q_N returns q.k spatio-textual objects in descending order of score τ. . The score $\tau(p)$ is defined as:

$$\tau(p) = \frac{\theta(p.d, q.d)}{1 + \alpha \cdot \delta(p.l, q.l)} \qquad (1)$$

where $\delta(p.l, q.l)$represents the network proximity between the query location $q.l$ and the object location *p.l*, and $\theta(q.d, p.d)$ represents the textual relevance of *p.d* according to the query keywords q.d. The query parameter α is a positive real number $(\alpha \in \mathbb{R}^+)$ and defines the importance of one measure over the other.

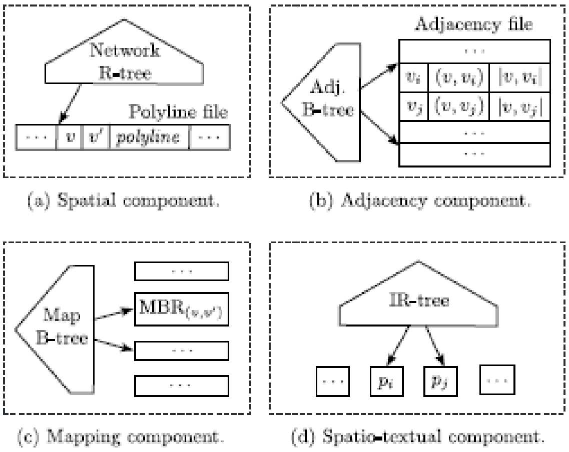
Network proximity (δ) gives the importance of the location of a spatio-textual object p.l to a query QN. Therefore, the network proximity is defined as:

$$\delta(p.l, q.l) = ||p.l, q.l|| \qquad (2)$$

Which is the shortest path between p and q.

Textual relevance (θ) gives the importance of the textual description of a spatio-textual object p.d to a query QN in terms of cosine similarity between p.d and q.d [22]. The textual relevance is defined as:

$$\theta(p.d, q.d) = \frac{\sum_{t \in q.d} w_{t,p.d} \cdot w_{t,q.d}}{\sqrt{\sum_{t \in p.d} (w_{t,p.d})^2 \cdot \sum_{t \in q.d} (w_{t,q.d})^2}} \qquad (3)$$



(a) Spatial component. (b) Adjacency component.

(c) Mapping component. (d) Spatio-textual component.

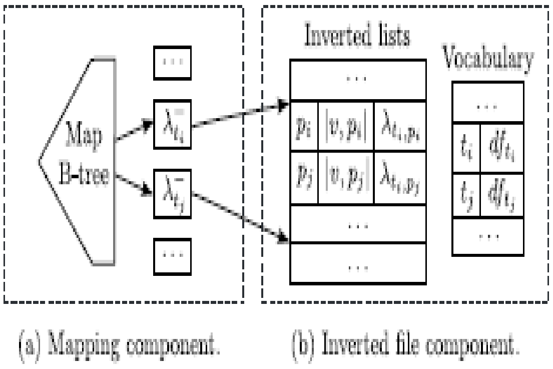
Fig.3 : Basic Indexing Architecture

IV. ENHANCED APPROACH

In this section we present an enhanced approach that indexes the objects lying on the edges of the road network for improving the query processing performance. First, we present the indexing architecture and query processing algorithm.

Enhanced Indexing Architecture:

Figure 5 presents the new components employed in the enhanced indexing architecture. The mapping component depicted in Figure 5(a) replaces the mapping component presented in Figure 3(c), and the inverted file component shown in Figure 5(b) replaces the spatio-textual component shown in Figure 3(d). Similarly to the basic indexing architecture, the new mapping component is the connection between the network and the objects through the edge of the network. The inverted lists and vocabulary compose the inverted file component. In the following, we describe the new components.



(a) Mapping component. (b) Inverted file component.

Fig. 5: Enhanced Indexing Architecture

Mapping Component:

The mapping component employs a B-tree named map B-tree that maps a key composed by the pair edge id and term id to the inverted list that contains the objects lying on the edge with term t in their description; see Figure 5(a).

The mapping component contains also the maximum impact λ_t^- of a given term t among the description of the objects lying on a given edge. The maximum impact λ_t^- is an upper-bound impact for any object on the edge that contains t. Therefore, the inverted list of a term t on an edge is accessed only if the upper bound score composed by minimum distance and λ_t^- may turn an object, present in the edge, inside the top-k objects found so far.

Inverted file component. The inverted file component (Figure 5(b) is composed by inverted lists and vocabulary. The inverted file contains inverted lists identified by a key composed by edge *id* and *term id.*

Each inverted list stores the objects lying on the edge *(v, v!)* that have a term t in their description. For each object, the inverted list stores:

1) The network distance between the object and the reference vertex of the edge

2) The impact of the term *ti* in the description of the object

In the enhanced indexing, the objects lying on a given edge are stored in inverted files. Therefore, it does not require performing a search on a spatio-textual index for finding the spatio-textual objects relevant for the query that lie on the index.

Furthermore, the enhanced indexing keeps an upper-bound score for each pair

Enhanced Query Processing:

Processing top-k spatial keyword queries on road network employing the enhanced indexing architecture can be performed using the basic algorithm (Algorithm 1). The only, but significant, change lies on the *find Candidates* procedure.

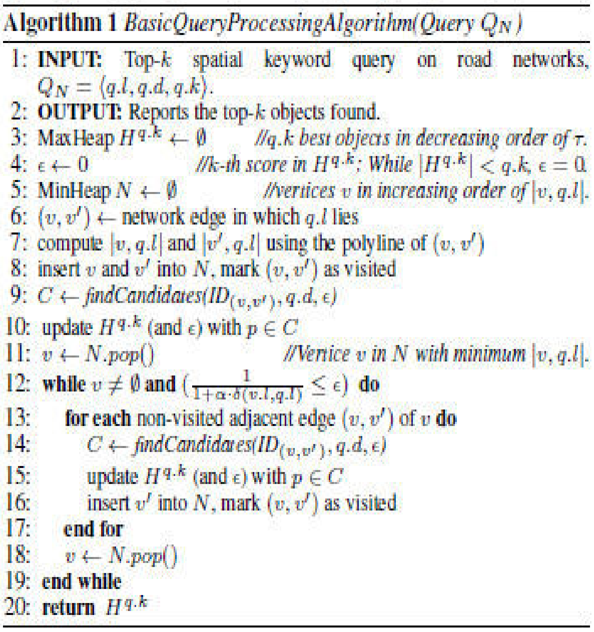
The new *findCandidates* procedure employed in the enhanced approach works as follows. **First**, the mapping component (Figure 5(a)) is accessed to compute an upper-bound score using the maximum impact λ_t^- of a given term $t \in q.d$ the minimum network distance between the edge and the query location. **Second**, if the upper-bound score is higher than ∈ the inverted lists (one list per query keyword) are accessed. The lists that contain objects are retrieved and the objects whose scores are higher than ∈ are returned.

The *findCandidates* procedure on the enhanced approach is more efficient. **First**, only the lists that can produce relevant objects are accessed. **Second**, retrieving the inverted lists is faster than processing a query location for retrieving all objects inside a given MBR. Only relevant objects are retrieved since the key of the mapping component incorporates the query keyword. **Third**, it does not require a filtering process to remove the objects that are inside the MBR of the edge, but not on the polyline of the edge. Finally, it does not require computing the network distance between the objects and the end vertices of the edges. The distances are computed and stored in the inverted lists during the index construction.

The enhanced query processing algorithm performs well when the network is populated, the query keywords are frequent, or the query preference parameter gives more weight to the network distance. In these cases, k objects with good scores are found rapidly, which permits the algorithm to terminate earlier. On the other hand, it can perform poorly if the k objects cannot be found rapidly, which can be common in top-k spatial keyword queries on road networks due to queries with non-popular terms, a large number of distinct terms in the datasets, or a sparse Network.

Algorithms used:

Exiting Algorithm

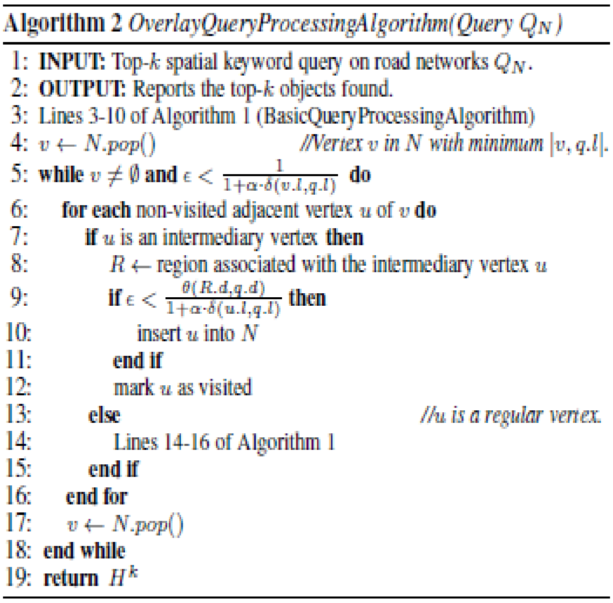


Algorithm 1 *BasicQueryProcessingAlgorithm(Query Q_N)*

```
1: INPUT: Top-$k$ spatial keyword query on road networks, $Q_N = \langle q.l, q.d, q.k \rangle$.
2: OUTPUT: Reports the top-$k$ objects found.
3: MaxHeap $H^{q.k} \leftarrow \emptyset$      //$q.k$ best objects in decreasing order of $\tau$.
4: $\epsilon \leftarrow 0$      //$k$-th score in $H^{q.k}$; While $|H^{q.k}| < q.k$, $\epsilon = 0$.
5: MinHeap $N \leftarrow \emptyset$      //vertices $v$ in increasing order of $|v, q.l|$.
6: $(v, v') \leftarrow$ network edge in which $q.l$ lies
7: compute $|v, q.l|$ and $|v', q.l|$ using the polyline of $(v, v')$
8: insert $v$ and $v'$ into $N$, mark $(v, v')$ as visited
9: $C \leftarrow$ findCandidates($ID_{(v,v')}, q.d, \epsilon$)
10: update $H^{q.k}$ (and $\epsilon$) with $p \in C$
11: $v \leftarrow N.pop()$      //Venice $v$ in $N$ with minimum $|v, q.l|$.
12: while $v \neq \emptyset$ and $(\frac{1}{1+\alpha\cdot\delta(v.l,q.l)} \leq \epsilon)$ do
13:     for each non-visited adjacent edge $(v, v')$ of $v$ do
14:         $C \leftarrow$ findCandidares($ID_{(v,v')}, q.d, \epsilon$)
15:         update $H^{q.k}$ (and $\epsilon$) with $p \in C$
16:         insert $v'$ into $N$, mark $(v, v')$ as visited
17:     end for
18:     $v \leftarrow N.pop()$
19: end while
20: return $H^{q.k}$
```

The basic query processing algorithm can be employed to process top-k spatial keyword queries on road networks. The main problem of this algorithm lies on the *find Candidates* procedure that is repeated for each adjacent edge. This procedure is expensive because it requires performing a search on a spatio-textual index, a filtering process for finding the relevant spatio-textual objects lying on a given polyline, and computing the network distance between the objects and the end vertices of the polyline.

Enhanced Overlay Query Processing Algorithm:



Algorithm 2 *OverlayQueryProcessingAlgorithm(Query Q_N)*

```
1: INPUT: Top-$k$ spatial keyword query on road networks $Q_N$.
2: OUTPUT: Reports the top-$k$ objects found.
3: Lines 3-10 of Algorithm 1 (BasicQueryProcessingAlgorithm)
4: $v \leftarrow N.pop()$      //Vertex $v$ in $N$ with minimum $|v, q.l|$.
5: while $v \neq \emptyset$ and $\epsilon < \frac{1}{1+\alpha\cdot\delta(v.l,q.l)}$ do
6:     for each non-visited adjacent vertex $u$ of $v$ do
7:         if $u$ is an intermediary vertex then
8:             $R \leftarrow$ region associated with the intermediary vertex $u$
9:             if $\epsilon < \frac{\theta(R.d,q.d)}{1+\alpha\cdot\delta(u.l,q.l)}$ then
10:                insert $u$ into $N$
11:            end if
12:            mark $u$ as visited
13:        else      //$u$ is a regular vertex.
14:            Lines 14-16 of Algorithm 1
15:        end if
16:    end for
17:    $v \leftarrow N.pop()$
18: end while
19: return $H^k$
```

The overlay query processing algorithm takes advantage of the overlay network to improve the performance of top-k spatial keyword queries on road networks. The idea is to avoid expanding the intermediary vertices that are not relevant for the query. Consequently, fewer vertices are expanded and the query processing terminates earlier.

V. PERFORMANCE EVALUATION:

Spatial preference (SP) query integrates two types of ranking that are Spatial Ranking and Non Spatial Ranking. Spatial Ranking refers to ranking objects based on distance from reference point. Non Spatial Ranking based on aggregated qualities of features in road network.

Performance metrics are measured based on the three concepts such as 1. Query size 2. Rank and neighbor range 3. Number of spatial objects.

Performance on Queries with Range Scores:

Fig. 6 plots the cost of the algorithms with respect to the number m of feature data sets. The costs of GP (**Group Probing**), BB (**Branch-and-Bound Algorithm**), and BB* rise linearly as m increases because the number of component score computations is at most linear to m. On the other Hand, the cost of FJ (**Feature Join Algorithm**) increases significantly with m, because the number of qualified combinations of entries is exponential to m.

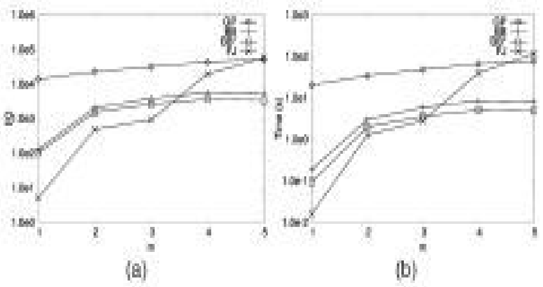


Fig. 6: Effect of m, range scores. (a) I/O. (b) Time.

Fig. 7 shows the cost of the algorithms as a function of the number k of requested results. GP, BB, and BB* compute the scores of objects in D in batches, so their performance is insensitive to k. As k increases, FJ has weaker pruning power and its cost increases slightly.

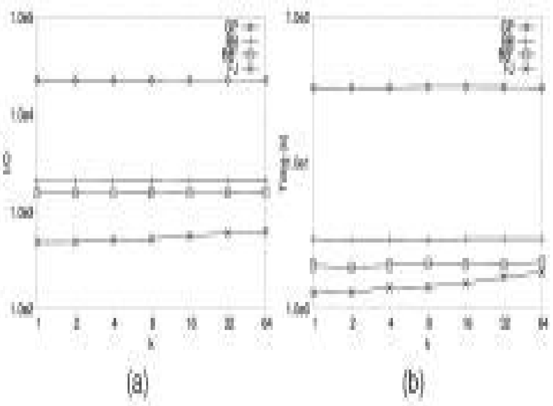


Fig. 7: Effect of k, range scores. (a) I/O. (b) Time.

Fig. 8 shows the cost of the algorithms, when varying the query range r. As r increases, all methods access more nodes in feature trees to compute the scores of the points. The difference in execution time between BB* and FJ shrinks as r increases.

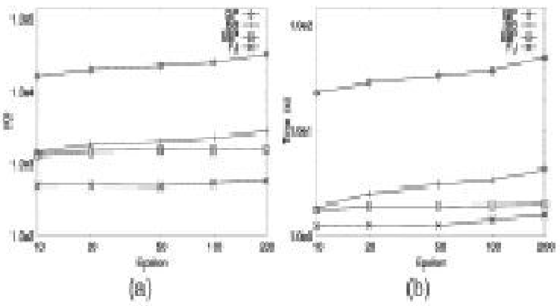


Fig. 8: Effect of r, range scores. (a) I/O. (b) Time.

VI. CONCLUSION:

In this paper, we introduced top-k spatial keyword queries on road networks. Given a spatial location and a set of query keywords; a top-k spatial keyword query on road networks returns the k best spatio-textual objects ranked in terms of both textual similarity to the query keywords and shortest path to the query location. We presented a straight-forward approach (basic approach) to process these queries combining state-of-the-art techniques. Then, we presented an enhanced approach that indexes the edges of the road network, and permits identifying and retrieving the objects relevant to the query efficiently. Finally we presented Performance Evaluation among Queries with Range Scores.

REFERENCES

[1] João B. Rocha-Junior_and Kjetil Nørvåg Top-k Spatial Keyword Queries on Road Networks in 2012

[2] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi. Collective spatial keyword querying. In SIGMOD, 2011.

[3]. Guttman, "R-Trees: A Dynamic Index Structure for Spatial Searching," Proc. ACM SIGMOD, 1984.

[4] G.R. Hjaltason and H. Samet, "Distance Browsing in Spatial Databases," ACM Trans. Database Systems, vol. 24, no. 2, pp. 265-318, 1999.

[5]. K.S. Beyer, J. Goldstein, R.Ramakrishnan, and U. Shaft, "When is 'Nearest Neighbor'Meaningful?" Proc. Seventh Int'l Conf. Database Theory (ICDT), 1999.

[6]. R. Fagin, A. Lotem, and M. Naor, "Optimal Aggregation Algorithms for Middleware,"Proc. Int'l Symp. Principles of Database Systems (PODS), 2001.

[7]. I.F. Ilyas, W.G. Aref, and A. Elmagarmid, "Supporting Top-k Join Queries inRelational Databases," Proc. 29th Int'l Conf. Very Large Data Bases (VLDB), 2003.

[8]. N. Mamoulis, M.L. Yiu, K.H. Cheng, and D.W. Cheung, "Efficient Top-k Aggregation of Ranked Inputs," ACM Trans. Database Systems, vol. 32, no. 3, p. 19, 2007.

◈◈◈

Automated SWIM (Single Window Infrastructure Management) for Cloud

Jyothi Ganig B S & Ramesh Namburi

Department of Information Science Engineering, MSRIT, Bangalore, India
E-mail : Jyothi.gng@gmail.com, ramram315.513@rediffmail.com

Abstract **-** Data centers manage complex server environments, including physical and virtual machines, across a wide variety of platforms, and often in geographically dispersed locations. Information Technology managers are responsible for ensuring that servers in these increasingly complex environments are properly configured and monitored throughout the IT life cycle. They also face challenges managing the physical and virtual environments and the fact that we must centralize, optimize and maintain both.

If the variation and complexity can be taken out of a process to make it more consistent, it can be automated. Through the use of virtual provisioning software, provisioning and re-purposing of infrastructure will become increasingly automatic. Staff will physically rack once, cable once, and thereafter (remotely) reconfigure repeatedly, effortlessly, as needed.

An automatic infrastructure will rapidly change which servers are running what software and how those servers are connected to network and storage. It will re-purpose machines according to the real-time demands of the business. It will enable capacity to be "dialed up" or "dialed down". And it will bring up a failed server on new hardware, with the same network and storage access and addressing, within minutes. All without needing to make physical machine, cable, LAN connection or SAN access changes.

Keywords **-** *Automation, cloud computing, virtualization.*

I. INTRODUCTION

IT is currently under a mandate to use resources more efficiently. In the past, IT organizations typically built an infrastructure that tightly coupled application workloads with physical assets, forcing IT administrators to continually perform repetitive manual tasks just to keep the operating environment up and running. In an effort to maintain some control, software is layered alongside the multiple instances of operating systems to enforce security, manage availability, and ensure performance. Then, to further compound the complexity, some applications require unique hardware, software, and skill sets to support specific business requirements. This entire approach is often referred to as IT being built in silos or IT operating on individual islands at best, but it is inefficient. In these scenarios, penetrating the walls between the various application workloads to help drive consolidation efforts can be a difficult task in terms of both management and security. Business owners have actually tolerated this approach for some time now, but as applications scale to new compliance and security mandates have been put in place and, in a less than perfect economy, companies are finding themselves tied at the ankles by an inefficient IT environment that is extremely costly to support and maintain. In many cases, it is prohibiting overall business growth and stability. So automation obviously becomes the critical components of the new model.

Automation means that many critical server and system resources can manage themselves, flawlessly completing tasks that once required a human catalyst. Virtualization gives the power to proactively deploy servers, easily adjusting and allocating resources when and where they are needed. Information Technology managers seek solutions to help them automate the consolidation of servers through virtualization, the provisioning of new servers, and processes to keep systems current with the latest software and security updates. At the same time they need to work diligently on monitoring the health of their infrastructure, industry compliance regulations, and reducing the total cost of managing that data center.

To create a cloud service, self-service and metering (feedback about the cost of the resources allocated) are offered in addition to automation. With self-service and metering, the computing model resembles a utility. The private cloud then is a technical strategy to turn computing power into utility computing, with the data and costs controlled and managed by the enterprise. Self service and metering are breakthrough capabilities for end users and business units, facilitating management and extension of the user experience. Now there is no intermediary between the consumer of a resource and the processes for the acquisition and allocation of resources for core businesses requirements and initiatives. Since the consumer initiates the service requests, now IT is an on-demand service rather than a gatekeeper. With the transition to an on-demand service, the cost structure is dramatically reduced, since the user uses and pays for only what is needed at a specific moment.

The business benefit of this change in cost structure is significant. While self-service and metering are breakthroughs private cloud capabilities for end users and business units, maintaining service delivery in a fully virtualized multi-tenancy environment and providing security, especially for information and services leaving the data center environment, are essential enterprise requirements for IT administrators.

With a private cloud utility model enabling these needs and requirements, enterprises can scale and expand by pooling IT resources in a single cloud operating system or management platform. They then can support anywhere from tens to thousands of applications and services and enable new architectures targeting very large-scale computing activities.

II RULES-BASED AUTOMATION

Automation is the next step in the evolutionary chain, making technology do the business's bidding and driving growth, innovation and profitability. This will drive incredible cultural change within organizations as constraints are removed and management can drive innovation and growth knowing that IT can respond to the ever-changing priorities of the business.

The automated infrastructure will, of course, need a set of rules, and the provision of its demands from the business will need to be delivered in a structured manner. This rulebook will be the service catalog, essentially a huge database of functions an organization undertakes and the parameters for each – such as a description of the service, timeframes, SLAs, costs and actions required. And a runbook will provide step-by-step procedures for governing workflow. In the automated infrastructure, these powerful workflow automation and management systems, with strict policy control, will:

- allocate resources to the applications and users that need them automatically in real-time
- continually monitor service levels to ensure business performance is on target
- provide a dynamic, on-demand environment, with support for the industry's leading virtualization, provisioning and re-purposing tools
- support major third-party servers, software and devices

III CHALLENGE

Labor Intensive, Time Consuming Virtual Machine (VM) Resource Provisioning

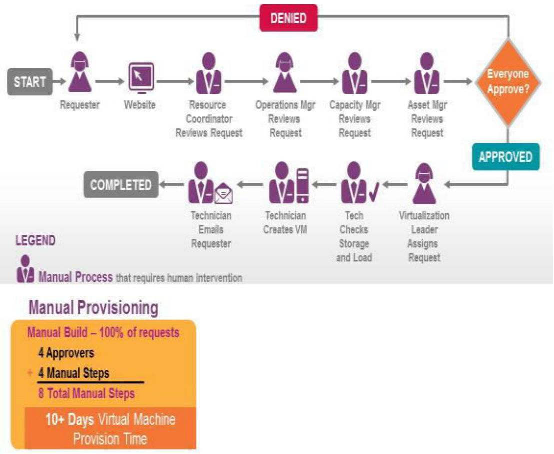

DENIED
START
Requester
Website
Resource Coordinator Reviews Request
Operations Mgr Reviews Request
Capacity Mgr Reviews Request
Asset Mgr Reviews Request
Everyone Approve?
APPROVED
COMPLETED
Technician Emails Requester
Technician Creates VM
Tech Checks Storage and Load
Virtualization Leader Assigns Request
LEGEND
Manual Process that requires human intervention
Manual Provisioning
Manual Build – 100% of requests
4 Approvers
+ 4 Manual Steps
8 Total Manual Steps
10+ Days Virtual Machine Provision Time

Fig. 1: Manual Provisioning

With the increased amount of requests for VMs, it was taking too long to provision virtual machines, provisioning those VMs was labor and time consuming. In addition, users were not releasing the VMs when they were done with them. They'd simply keep them running and not do anything with them. This caused stranded resources such as processor, memory and disk space that could be used for other VMs. This caused the need for the environment to continually grow to meet new demand for virtual machines. Virtual Machine request process went through as follows, where users would go to a web page, make a request for resources and approval process was reviewed by cloud administrator and the manual steps that were required to create a VM. We determined that 95% of our requests fell into a common configuration and that we could 100% automate provisioning that common configuration.

Solution

A) Automated VM Provisioning- Standard Build

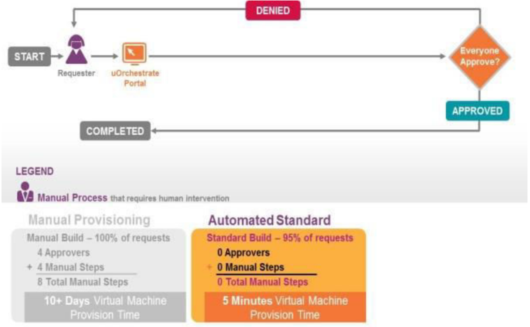


Fig. 2 :Automated Standard Provisioning

B) Automated VM Provisioning- Manual Build

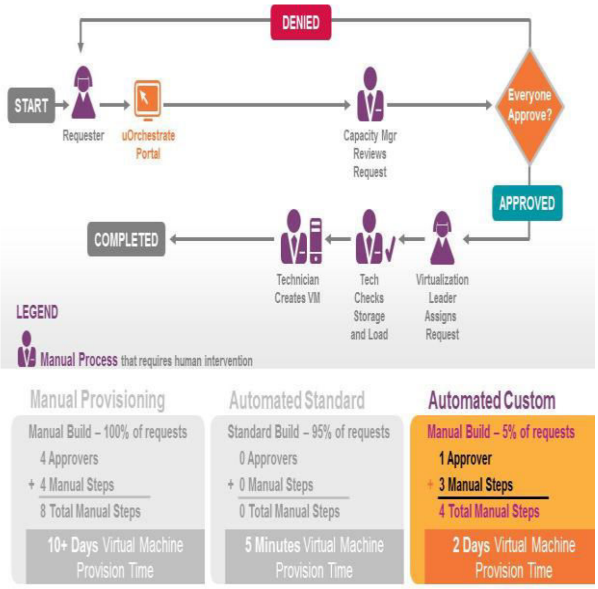

DENIED
START
Requester
uOrchestrate Portal
Capacity Mgr Reviews Request
Everyone Approve?
APPROVED
COMPLETED
Technician Creates VM
Tech Checks Storage and Load
Virtualization Leader Assigns Request
LEGEND
Manual Process that requires human intervention
Manual Provisioning
Manual Build – 100% of requests
4 Approvers
+ 4 Manual Steps
8 Total Manual Steps
10+ Days Virtual Machine Provision Time
Automated Standard
Standard Build – 95% of requests
0 Approvers
+ 0 Manual Steps
0 Total Manual Steps
5 Minutes Virtual Machine Provision Time
Automated Custom
Manual Build – 5% of requests
1 Approver
+ 3 Manual Steps
4 Total Manual Steps
2 Days Virtual Machine Provision Time

Fig. 3 : Automated Custom Provisioning

It is essential that the entire process be correctly understood before attempting to plan out what will be automated. It makes no sense to automate a bad process. Get all parties involved, to discover any hidden steps or misunderstandings in the flow. Once the process is understood, determine if anything needs to change, or be dropped, in order to streamline the automated process.

Also, make sure that the client requirements are clear. Have a comprehensive test and evaluation plan for the automation process.

IV IMPLEMENTATION

A) Automation ↑ 100% => Availability ↑ 100%

- Automate Application Management – When VMs are created, agents are automatically embedded for backup , performance and capacity management
- Automated Asset Tracking – When VMs are created, the Configuration Item is automatically updated in the CMDB
- Automate incident management – incidents are automatically sent during the provisioning process
- Automate performance and capacity management – insert agents and automate updates for the collection process
- Automate VM provisioning tasks – Standardize the VM templates and automatically allocate them based on pre-determined approval schemes
- Automate Security – Establish tiers and tenants so that they can be automatically isolated using vLAN technology or more sophisticated encryption technology
- Automate Identity and Access Management federation – Single sign-on and seamless management from datacenter management to cloud management

B) "Near Zero Touch" delivered through extreme automation

Secure Private Cloud includes a complete suite of tools to provision, operated, integrate and meter the use of cloud-based IT infrastructure. These tools allow automating and scaling to hundreds or even thousands of servers with little or no incremental administration and management cost.

Component	Value Delivered
uOrchestrate	The Virtual Orchestrator component (which includes LifeRay portal) is the basis of the PSO UI PSO UI is the interface by which the user and administrator interact with Secure Private Cloud "Run-book" automation – software that programmatically sequences a number of formerly manual steps into an automated sequence
uAdapt	Automatically provisions / de-provisions HW/SW platform from a pool of waiting servers Provisions physical servers Creates a "persona" which defines an instance of the HW attributes and SW components that are to be provisioned
VMware vCenter	Does most of the heavy lifting for provisioning of virtual servers Manages "templates" – defined set of SW that a VM will contain

Table 1 : Components used for automation

V RESULTS

Cost and Value Analysis

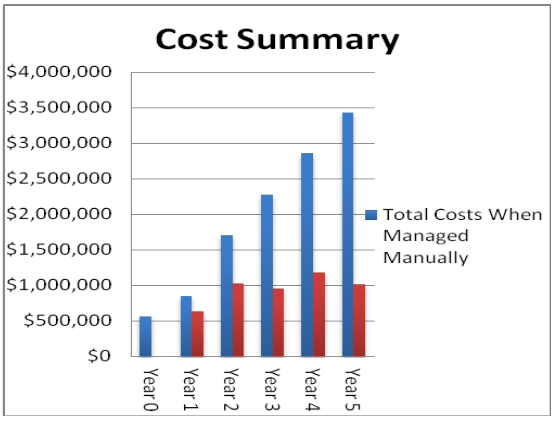


Fig. 5 Manual Provisioning cost versus Automated Provisioning cost

The economic conditions around the world have forced companies to be more cost conscious. With costs continuing to escalate, data centers are coming under unprecedented operational scrutiny. The Information Technology (IT) managers, faced with the challenge of squeezing more performance out of existing data center budgets, need to limit the capital and operational expense on additional equipment, simplify network operations, and ensure maximum efficiencies

Sophisticated automation can help to reduce operating expenses through:

- On-demand reallocation of computing resources
- Run-time response to capacity demands
- Trouble-ticket response automation (or elimination of trouble tickets for most automated response scenarios)
- I ntegrated system management and measurement

VI CONCLUSION

Enterprises need to move from managing underlying infrastructure to managing service levels based on what makes sense for the user of applications. For example, the customer may want to manage factors such as the minimum tolerable application latency or the availability level of an application. Enterprises also must implement automation for central IT and self-service for end users, thus extricating IT from the business of repetitive management procedures and enabling end users to get what they need quickly.

In this stage, virtualization optimizes IT resources and increases IT agility, thus speeding time-to-market for services. The IT infrastructure undergoes a transformation in which it becomes automated and critical IT processes are dynamic and controlled by trusted policies. Through automation, data centers systematically remove manual labor requirements for the run-time operation of the data center.

REFERENCES

[1] Unisys.com/secureprivatecloud

[2] http://www.unisys.com/unisys/ri/pub/bl/detail.jsp?id=1120000970018210153

[3] Enterprise Systems Journal http://esj.com/articles/2011/10 /10/private-clouds-just-vapor.aspx

[4] Software Architecture in Practice, 2nd Edition, Len Bass, Paul C. Clements, and Rick Kazman

[5] Software Engineering, Sixth Edition, Ian Somerville

[6] UML, http://www.uml.org/

Server Update Mechanism In MANETS

Nalini Suryadevara & P.Seshu babu

Computer Science and Engineering, Vignan University, Guntur(D.T), Andhra Pradesh, India
E-mail : nalini surya49@gmail.com, seshu.babu08@gmail.com

Abstract **-** A cache consistency scheme based on a previously proposed architecture for caching database data in MANETs. The original scheme for data caching stores the queries that are submitted by requesting nodes in special nodes, called query directories (QDs), and uses these queries to locate the data (responses) that are stored in the nodes that requested them, called caching nodes (CNs). The consistency scheme is server-based in which control mechanisms are implemented to adapt the process of caching a data item and updating it by the server to its popularity and its data update rate at the server. The system implements methods to handle disconnections of QD and CN nodes from the network and to control how the cache of each node is updated or discarded when it returns to the network. Estimates for the average response time of node requests and the average node bandwidth utilization are derived in order to determine the gains (or costs) of employing our scheme in the MANET.

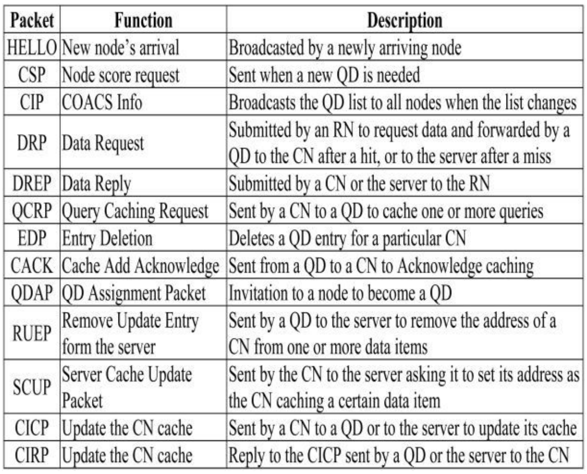
Keywords **-** *Data caching, cache consistency,QD,CN,server-based approach, MANET.*

I. INTRODUCTION

In a mobile ad hoc network (MANET), data caching is essential as it reduces contention in the network, increases the probability of nodes getting desired data, and improves system performance . The major issue that faces cache management is the aintenance of data consistency between the client cache and the server . In a MANET, all messages sent between the server and the cache are subject to network delays, thus, impeding consistency by download delays that are considerably noticeable and more severe in wireless mobile devices. All cache consistency algorithms are developed with the same goal in mind: to increase the probability of serving data items from the cache that are identical to those on the server. A large number of such algorithms have beenproposed in the literature, and they fall into three groups: server invalidation, client polling, and time to live (TTL). With server invalidation, the server sends a report upon each update to the client. Two examples are the Piggyback server invalidation and the Invalidation report mechanisms. In client polling, like the Piggyback cache validation of , a validation request is initiated according to a schedule. If the copy is up to date, the server informs the client that the data have not been modified; else the update is sent to the client. Finally, with TTL algorithms, a server-assigned TTL value (e.g., T) is stored alongside each data item d in the cache. The data d are considered valid . until T time units pass since the cache update. Usually, the first request for d submitted by a client after the TTL expiration will be treated as a miss and will cause a trip to the server to fetch a fresh copy of d. Many algorithms were proposed to determine TTL values, including the fixed TTL approach, adaptive TTL , and Squid's LM-factor. TTL-based consistency algorithms are popular due to their simplicity, sufficiently good performance, and flexibility to assign TTL values for individual data items. However, TTL-based algorithms, like client polling algorithms, are weakly consistent, in contrast to server invalidation schemes that are generally strongly consistent. According to , with strong consistency algorithms, users are served strictly fresh data items, while with weak algorithms, there is a possibility that users may get inconsistent (stale) copies of the data. This work describes a server-based scheme implemented on top of the COACS caching architecture we proposed in . In COACS, elected query directory (QD) nodes cache submitted queries and use them as indexes to data stored in the nodes that initially requested them (CN nodes). Since COACS did not implement a consistency strategy, the system described in this paper fills that void and adds several improvements: 1) enabling the server to be aware of the cache distribution in the MANET, 2) making the cached data items consistent with their version at the server, and 3) adapting the cache update process to the data update rate at the server relative to the request rate by the clients. With these changes, the overall design provides a complete caching system in which the server sends to the clients selective updates that adapt to their needs and reduces the average query response time.

II. PROBLEM DEFINITION

Traditional server-based schemeThey are not usually aware of what data items are currently cached, as they might have been replaced or deleted from the network due to node disconnections. If the server data update rate is high relative to the nodes request rate, unnecessary network traffic would be generated, which could increase packet dropout rate and cause longer delays in answering node queries.



Packet	Function	Description
HELLO	New node's arrival	Broadcasted by a newly arriving node
CSP	Node score request	Sent when a new QD is needed
CIP	COACS Info	Broadcasts the QD list to all nodes when the list changes
DRP	Data Request	Submitted by an RN to request data and forwarded by a QD to the CN after a hit, or to the server after a miss
DREP	Data Reply	Submitted by a CN or the server to the RN
QCRP	Query Caching Request	Sent by a CN to a QD to cache one or more queries
EDP	Entry Deletion	Deletes a QD entry for a particular CN
CACK	Cache Add Acknowledge	Sent from a QD to a CN to Acknowledge caching
QDAP	QD Assignment Packet	Invitation to a node to become a QD
RUEP	Remove Update Entry form the server	Sent by a QD to the server to remove the address of a CN from one or more data items
SCUP	Server Cache Update Packet	Sent by the CN to the server asking it to set its address as the CN caching a certain data item
CICP	Update the CN cache	Sent by a CN to a QD or to the server to update its cache
CIRP	Update the CN cache	Reply to the CICP sent by a QD or the server to the CN

2.1 Dealing with Query Replacements and Node Disconnections

A potential issue concerns the server sending the CN updates for data that have been deleted (replaced), or sending the data out to a CN that has gone offline. To avoid this and reduce network traffic, cache updates can be stopped by sending the server Remove Update Entry Packets (RUEPs). This could occur in several scenarios. For example, if a CN leaves the network, the QD, which first tries to forward it a request and fails, will set the addresses of all queries whose items are cached by this unreachableCN in its cache to -1, and sends an RUEP to the server containing the IDs of these queries. The server, in turn,changes the address of that CN in its cache to -1 and stops sending updates for these items. Later, if another node A requests and then caches one of these items, the server, upon receiving an SCUP from A, will associate A with this data item. Also, if a CN runs out of space when trying to cache a new item i_n, it applies a replacement mechanism to replace $_{\text{Id}}$ with I_n and instructs the QD that caches the query associated with I_d to delete its entry. This causes the QD to send an RUEP to the server to stop sending updates for i_d in the future.If a caching node CN_d returns to the MANET after disconnecting, it sends a Cache Invalidation Check Packet (CICP) to each QD that caches queries associated with items held by this CN. A QD that receives a CICP checks for each item to see if it is cached by another node and then sends a Cache Invalidation Reply Packet (CIRP) to CN_d containing all items not cached by other nodes. CN_d then deletes from its cache those items whose $_{\text{IDs are}}$ not in the CIRP but were in the CICP. After receiving a CIRP from all QD_s to which it sent a CICP and deleting nonessential data items from its cache, CN sends a CICP containing the ID_s of all queries with data remaining in its cache to the server along with their versions. In the meanwhile, if CN_d receives a request from a QD for an item in its cache, it adds the request to a waiting list. The server then creates a CIRP and includes in it fresh copies of the outdated items and sends it to CN , which, in turn, updates its cache and answers all pending requests. Finally, and as described in , QD disconnections and reconnections do not alter the cache of the CNs, and hence, the pointers that the server holds to the CN_s remain valid.

2.2 Adapting to the Ratio of Update Rate and Request Rate

SSUM suspends server updates when it deems that they are unnecessary. The mechanism requires the server to monitor the rate of local updates, R_u , and the rate of RN requests, R_r, for each data item d_i . Each CN also monitors these values for each data item that it caches. Whenever a CN receives an update from the server, it calculates R and compares it to a threshold Γ. If this ratio is greater than or equal to Γ, the CN will delete d and the associated information from its cache and will send an Entry Deletion Packet (EDP) to the QD (say, QD di) that caches query q . The CN includes in the header of EDP a value for R_u , which tells QD_d that d_i is being removed due to its high update-torequest ratio. Normally, when a QD gets an EDP, it removes the cached query from its cache, but here, the nonzero value of $_{\text{Ru in}}$ the EDP causes QD to keep the query cached, but with no reference to a CN. Next, QD d will ask the server to stop sending updates for di . Afterward, when QD receives a request from an RN node that includes q , it forwards it to the server along with a DONT_CACHE flag in the header to be later passed in the reply, which includes the results, to the RN. Under normal circumstances in COACS, when an RN receives a data item from the server in response to a query it had submitted, it assumes the role of a CN for this item and will ask the nearest QD to cache the query. The DONT_CACHE flag instructs the RN to treat the result as if it were coming from the cache and not become a CN for it. Now, at the server, each time an update for qi occurs and a new R_u=Rri is computed, if this ratio falls below a second threshold, _ (_<_), the server will reply to the RN with a DREP that includes the CACHE_NEW flag in the header. Upon receiving the DREP, the RN sends a QCRP with the CACHE_NEW flag to its nearest QD. If this QD caches the query of this item (with _1 as its CN address), it sets its address to its new CN, else it forwards the request to its own nearest QD. If the QCRP traverses all QDs without being processed (implying that the QD caching this item has gone offline), the last QD at which the QCRP arrives will cache the query with the CN address. By appropriately selecting the values of Γ and , the system can reduce unnecessary network traffic. The processing time of qi will suffer though when R_u=R is above _after it had passed _ since QD d_r will be sending q to the server each time it receives it. However, the two thresholds allow for favoring bandwidth consumption over response time, or vice versa. This makes SSUM suitable for a variety of mobile computing applications: a large _ may be used when disconnections are frequent and data availability is important, while a low _ could be used in congested environments where requests for data are infrequent or getting fresh data is not critical. summarizes the interactions among the entities of the system.

2.3 Accounting for Latency in Receiving Server Updates

Given the different processes running at the server and since it sends the updates to the CNs via unicasts, there may be a time gap between when an update occurs and when the CN actually receives the updated data item d. Hence, if the CN gets a request for d during this time, it will deliver a stale copy of d to the RN. Our design uses the time stamp that the server sends with each update in an attempt to mitigate this issue. To explain this, suppose that the time stamp sent with d is ts and the time of receiving d by the CN is t. Upon getting an update, the CN checks if it had served any RN a copy of d from its cache in the past t_s-t milliseconds. If it is the case, the CN sends a new DREP to the RN, but now it includes the fresh copy of d. The above solution assumes that the clocks of the nodes in the MANET and that of the server are synchronized. This assumption is realistic given that node clock synchronization is part of the MAC layer protocol, as specified by the IEEE 802.11 standards. In particular, IEEE 802.11 specifies a Timing Synchronization Function (TSF) through which nodes synchronize their clocks by broadcasting their timing information using periodic beacons. Since the Access Point (AP) is considered a node in the MANET and it can synchronize its clock with that of the server asynchronously with respect to the MANET through the wired network, it will be possible to synchronize the clocks of the mobile nodes with that of the server at almost a zero cost to them(no protocol besides the MAC layer's TSF is needed). The suitability of TSF for SSUM depends on its effectiveness. It was shown in that using TSF, the maximum clock offset in the case of 500 nodes is 700 s. Several approaches were proposed to reduce this error. The Automatic Self-time Correcting Procedure (ASP) reportedly reduced the maximum offset to 300s, while the Multihop Adaptive TSF (MATSF) method cut it down to 50 _s, but at the expense of adding 8 bits to the IEEE 802.11 frame. we show that SSUM is best characterized by the delta consistency model , where the upper bound for the delta between the time of the server's update and the time the RN gets a fresh copy is in tens of milliseconds. It follows that TSF will virtually not increase this delta, especially in small to moderately sized networks.

2.4 Overhead Cost

When a node joins the network, the server will know about it when it first gets a query from it in a DRP that is forwarded by one of the QD nodes. Each data item at the server that is cached in the network is associated with a query id, a request rate, an update rate, and the address of the CN caching it. This additional information could cost the server about 16 bytes of extra storage per record. Hence,from a storage perspective, this cost may be deemed insignificant when considering the capabilities of modern servers. In terms of communication cost, the server communicates with the CNs information about the rates using header information in the exchanged packets, and uses control packets (RUEP, SCUP, CICP, and CIRP) to manage the updating process. Here, it suffices to state that the simulation results indicate that the overall overhead traffic (including other packets that do not concern the server) is a small portion of the data traffic.Finally, from a processing load point of view, the server is only required to manipulate the update rate when appropriate. In conclusion, the server will not incur any notably additional load due to its role in this architecture. Hence, the system should be able to scale to a large number of cached items.A similar argument can be made for the CNs, although the main concern here is the impact on the cache space and replacement frequency. Using the same value of 5 KB for the average data item size (as in the simulations of , with caching capacity of 200 KB, a CN can cache about 40 data items. The additional overhead required for storing the request and update rates of one single data item is 8 bytes, and therefore, the overhead for storing the request and update rates at the CN is less than 0.16 percent of the available space. It follows that the space for caching at the CNs is minimally impacted by this additional information. Also, the frequency of cache replacements will not increase in a major way because of this.

2.5 Consistency Model.

- H_C is the average number of hops between the corner of the topology and a random node in the MANET. It applies when a packet is sent between the server and the random node.

- H_R is the expected number of hops between any two randomly selected nodes.
- H_r is the expected number of hops to traverse all the QDs in the system, which usually occurs in the case of a cache miss.
- H_A is the expected number of hops to reach the QDwhich holds the reference to the requested data, in the case of a hit.

III. ANALYSIS

We evaluate our scheme in terms of bandwidth and query response time gains. These are the differences between the corresponding measures when no cache updating is in place and when SSUM is employed. Requests for data in the ad hoc network and data updates at the server are assumed to be random processes and may be represented by exponential random variables, as was suggested in and We use _R to denote the rate of requests and _ the rate of updates. The probability density functions of requests and updates are thus given:$_{Pr(t)=\lambda r}$e-λ_rt ,p$_{u(t)}$= $^{\lambda ue\text{-}\lambda ut}$$_{Both}$ the andwidth gain Gb and response time gain Gt are influenced by the number of data requests issued by requesting nodes relative to the number of data updates that occur at the server. In the remainder of the paper, we refer to no cache updating as NCU._ 1(C2). C2, in turn, comprises two scenarios: after _Wenowconsider two cases: $\lambda_{u/\lambda r}$< 1 (C1) and $\lambda_{u/\lambda r}$>1 (c2)(c1)in trun comprise two scenario :after $\lambda_{u/\lambda r}$>Γ and then while $\lambda_{u/\lambda r}$> Γ(C2S2), where updates are suspended by the server, and the remaining scenario (C2S1), where updates are sent by the server.Finally, the average update and request rates can be related as follows: In the first case (i.e., C1),λ_r=N$\lambda_{u,}$N >1, while in the first scenario of C2 (i.e., C2S1), λ_u=MΓ$_{r1<}$M<T . Starting with C2S2, SSUM suspends the server updates to the RN, thus, acting like NCU and resulting in a zero bandwidth gain. The cache entry would have been outdated upon getting the RN's next request, thus causing a cache miss. This causes a traversal of all QDs, then a transmission o and from the server, a transmission back to the requesting node, and then a packet sent to a QD asking it to cache the query. In C2S1, where $\lambda_{u/\lambda r}$< 1 and the updates are sent to the RNs, the requests are served from the cache. To compute the gain, we take a time period that includes K update periods such that K>M. The cache entry will be updated, and hence, an update packet is sent to the CN from the server, plus the request sent from the RN to the QD caching the query and then forwarded to the CN holding the data, and finally, the reply from the CN.

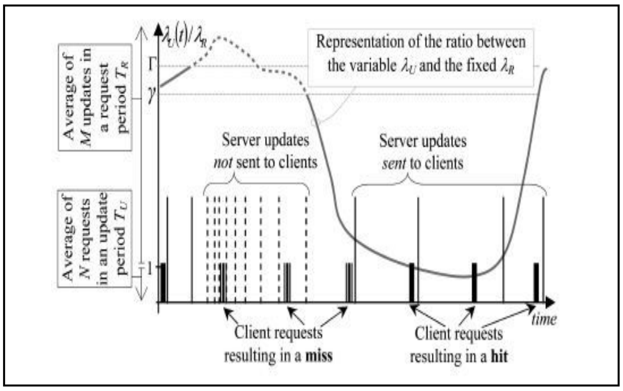

Average of M updates in a request period T_R
Average of N requests in an update period T_U
$\lambda_U(t)/\lambda_R$
Γ
γ
1
Representation of the ratio between the variable λ_U and the fixed λ_R
Server updates not sent to clients
Server updates sent to clients
time
Client requests resulting in a miss
Client requests resulting in a hit

Fig. 2: Behavior of SSUM as $\lambda_{u/\lambda r}$

REFERENCES

[1] S. Acharya, R. Alonso, M. Franklin, and S. Zdonik, "Broadcast Disks: Data Management for Asymmetric Environments," Proc. ACM SIGMOD, pp. 199-210, May

[2] H. Artail, H. Safa, K. Mershad, Z. Abou-Atme, and N. Sulieman, COACS: A Cooperative and Adaptive Caching System for MANETS," IEEE Trans. Mobile Computing, vol. 7, no. 8, pp. 961- 977, Aug. 2008.

[3] B. Krishnamurthy and C.E. Wills, "Piggyback Server Invalidationfor Proxy Cache Coherency," Proc. Seventh World Wide Web(WWW) Conf., Apr. 1998.

[4] B. Krishnamurthy and C.E. Wills, "Study of Piggyback Cache Validation for Proxy Caches in the World Wide Web," Proc USENIX Symp. Internet Technologies and Systems, Dec. 1997.

[5] D. Li, P. Cao, and M. Dahlin, "WCIP: Web Cache InvalidationProtocol," IETF Internet Draft, http://tools.ietf.org/html/draftdanli-wrec-wcip-01, Mar. 2001.

[6] W. Li, E. Chan, Y. Wang, and D. Chen, "Cache InvalidationStrategies for Mobile Ad Hoc Networks," Proc. Int'l Conf. Paralle Processing, Sept. 2007.[7] S. Lim, W.-C. Lee, G. Cao, and C.R. Das, "Performance Comparison of Cache Invalidation Strategies for Internet-Based Mobile-Ad Hoc Networks," Proc. IEEE Int'l Conf. Mobile Ad-Hoc and Sensor Systems, pp. 104-113, Oct. 2004.

[7] M.N. Lima, A.L. dos Santos, and G. Pujolle, "A Survey oSurvivability in Mobile Ad Hoc Networks," IEEE Comm. Survey and Tutorials, vol. 11, no. 1, pp. 66-77, First Quarter 2009.

Removal of High Density Salt-And-Pepper Noise Through Improved Adaptive Median Filter

Anisha Bhatia & Ramesh K Kulkarni

EXTC, Vivekanand Education Society's Institute Of Technology**,** Mumbai, India
E-mail : anisha_bhatiya@yahoo.com, rk1_2002@yahoo.com

***Abstract* -** In this paper, a new image-denoising filter that is based on the standard median (SM) filter is proposed. In our method, the adaptive standard median is used to detect noise and change the original pixel value to a newer that is closer to or the same as the standard median. With our experimental results, we have made a comparison among our method, the standard median (SM) filter, the median filter with mask of 3-by-3, 5-by-5, 7-by-7, the center weighted median (CWM) filter, the adaptive center weighted median (ACWM) filter, the progressive switching median (PSM) filter, the decision based median (DBM) filter, and the untrimmed (UT) median filter, in which our method proves to be superior.

***Keywords* -** *Salt-and-pepper noise, the standard median (SM) filter, the median filter with mask of 3-by-3, 5-by-5, 7-by-7, the center weighted median (CWM) filter, the adaptive center weighted median (ACWM) filter , the progressive switching median (PSM) filter, the decision based median (DBM) filter, the untrimmed (UT) median filter.*

I. INTRODUCTION

Digital images get corrupted by impulse noise when acquired by a defective sensor or when transmitted through a faulty channel. Impulse noise removal image processing is an important pre-processing step which involves the removal of salt and pepper noise from digital images so that the restored images can be applied to subsequent phases of segmentation [1]. Impulsive noises can be commonly found in the sensor or transmission channel during the acquisition and transfer procedure for the digital signals images. Salt-and-pepper noise is a typical kind of impulsive noise. The nonlinear filter algorithms are often adopted for the salt-and-pepper noise removal [2].

It is well known that linear filtering techniques fail when the noise is non-additive and are not effective in removing impulse noise. This has led the researchers to the use of nonlinear signal processing techniques. Classes of widely used nonlinear digital filters are median filters. Median filters are known for their capability to remove impulse noise as well as preserve the edges. The main drawback of a standard median filter (SMF) is that it is effective only for low noise densities. At high noise densities, SMFs often exhibit blurring for large window sizes and insufficient noise suppression for small window sizes [3].

When the noise level is over 50% the edge details of the original image will not be preserved by standard median filter, hence this poses a drawback for noise to be removed at higher density.

In the low-density noise cases, SM filter has better performance for noise removal and detail preservation. Since SM filter is implemented uniformly across the image, thus it modifies both noisy and noise-free pixels, so the denoising performance greatly degrades in the high noise density cases.

Hence to solve the above problem the adaptive median filter method has been proposed which eradicates the noise even at very high noise density levels.

The results obtained are then compared with the standard median (SM) filter, the median filter with mask of 3-by-3, 5-by-5, 7-by-7, the center weighted median (CWM) filter, the adaptive center weighted median (ACWM) filter, the progressive switching median (PSM) filter, the decision based median (DBM) filter, and the untrimmed median filter, in which our method proves to be superior.

Our method gives better Peak Signal-to-Noise Ratio (PSNR) and Mean square error (MSE), Correlation ratio (COR), universal quality index (UQI), Structural similarity index mean (SSIM) values than the existing algorithm.

II. METHODOLOGY

REVIEW OF FILTERS

The weighted median (WM) filter is an extension of the median filter, which gives more weight to some values within the window. This WM filter allows a degree of control of the smoothing behavior through the weights that can be set, and therefore, it is a promising image enhancement technique. In this paper, we focus our attention on a special case of WM filters called the center weighted median (CWM) filter. This filter gives more weight only to the central value of a window, and thus it is easier to design and implement than general WM filters. We shall analyze the properties of CWM filters and observe that CWM filters preserve more details at the expense of less noise suppression like the other non-adaptive detail preserving filters. In an attempt to improve CWM filters further, an adaptive CWM (ACWM) filter having a variable central weight has been proposed [4].

Next a median-based filter, progressive switching median (PSM) filter, is proposed to restore images corrupted by salt–pepper impulse noise.

The algorithm is developed by switching scheme—an impulse detection algorithm is used before filtering, thus only a proportion of all the pixels will be filtered and progressive methods—both the impulse detection and the noise filtering procedures are progressively applied through several iterations [5].

THE PROPOSED METHOD

The adaptive median filter is based on the following steps:

1. It checks for pixels that are noisy in the image, i.e. pixels with values 0 or 255 are considered.
2. For each such pixel P, a window of size 3×3 around the pixel P is taken.
3. Find the absolute differences between the pixel P and the surrounding pixels.
4. The arithmetic mean (AM) of the differences for a given pixel p is computed.
5. The AM is then compared with the “threshold” to detect whether the pixel p is informative or corruptive.

a) If AM is greater than or equal to the threshold the pixel is considered noisy.

b) Otherwise the pixel is considered as information.

Median filters produce the best result for a mask of size 3×3 at low noise density levels though the image is considerably blurred. The filter fails to perform well at higher noise densities and hence an alternative that works well at low as well as higher noise densities is required. This lies in the fact that when noise density is high it is highly unlikely that there might be more informative pixels than corruptive pixels.

The proposed method overcomes the shortcomings faced by the normal median filter at high noise densities by considering only those pixels that are informative in the neighbourhood.

The algorithm for the improved adaptive median filtering is as follows

1. Noise is detected by the noise detection algorithm as mentioned above.
2. Filtering is done only at those pixels that were detected as noisy.
3. Once a given pixel p is found to be noisy the following steps are followed.

a) A 3×3 mask is centred at the pixel p and finds if there exists at least one informative pixel around the pixel P.

b) If found so, the pixel p is replaced by the median of the informative pixels found in the 3×3 neighbourhood of P.

4. The above steps are repeated if noise still persists in the output image for betterment.

Peak Signal-to-Noise Ratio (PSNR) and Mean square error (MSE), universal quality index (UQI), Structural similarity index mean (SSIM) of the output image are computed to analyse the performance of the proposed filter as a denoising technique.

III. EXPERIMENTAL RESULTS

The performance of the proposed improved adaptive median filter, a comparison among our method, the standard median (SM) filter, the median filter with mask of 3-by-3, 5-by-5, 7-by-7, the center weighted median (CWM) filter, the adaptive center weighted median (ACWM) filter , the progressive switching median (PSM) filter, the decision based median (DBM) filter, and the untrimmed median filter, were analyzed for high noise density (ND) of salt-and-pepper noise added to gray level Lena image shown in Fig.1

Fig. 1 : Original Lena image

RESULT FOR PROPOSED FILTER

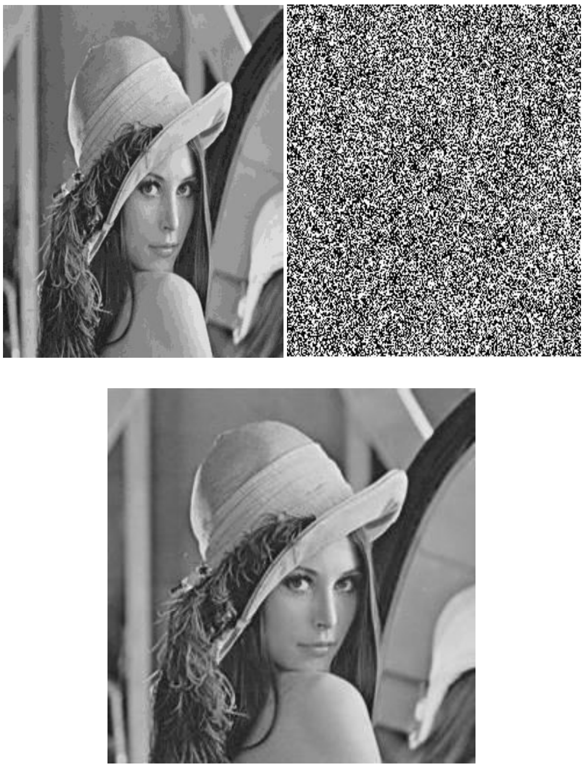


Fig. 2 : Original image, image corrupted with 90% noise, output of proposed filter

The parameters used to define the performance are:

Peak Signal-to-Noise Ratio (PSNR):

$PSNR = 20 \log_{10}(255/RMSE)$

where Root Mean square error (MSE):

$$RMSE = \surd 1/MN \sum_{i,j} (Y_{ij}-X_{ij})^{2}$$

Correlation ratio (COR):

$$COR = \sum_{i,j} (Y_{ij}-\mu_y)(X_{ij}-\mu_x) / \surd \sum_{i,j} (Y_{ij}-\mu_y)^{2} \sum (X_{ij}-\mu_x)^{2}$$

Structural similarity index mean (SSIM):

SSIM is designed to improve on traditional methods like peak signal-to-noise ratio (PSNR) and mean squared error (MSE), which have proved to be inconsistent with human eye perception [7].

The Structural Similarity (SSIM) index is a method for measuring the similarity between two images. The SSIM index can be viewed as a quality measure of one of the images being compared provided the other image is regarded as of perfect quality. It is an improved version of the universal image quality index [6].

The structural similarity index correlates with human visual system. Thus SSIM is used as a perceptual image quality evaluation metric. The SSIM is defined as function of luminance, contrast and structural components(s) [7].

Universal image quality index (UQI):

The universal quality index is given by

$Q = 1/M \sum Q_j$

Where M is the steps and Q_j is the local universal quality index [8].

COMPARISON TABLE OF VARIOUS FILTERS WITH PROPOSED FILTER (PF)

(A) PSNR TABLE

MF 3X3	MF 5X5	MF 7X7	CW M	ACW M	DBMF	PSMF	UT	PF
6.42	6.46	6.18	3.35	25.005	8.82	2.913	8.8	∞

(B) MSE TABLE

MF 3X3	MF 5X5	MF 7X7	CWM	ACWM	DBMF	PSMF	UT	PF
4.24	4.4	5.3	1.4	13.9	0.1	1.7	15.8	0

(C) COR TABLE

MF 3X3	MF 5X5	MF 7X7	CWM	ACWM	DBMF	PSMF	UT	PF
-.03	-.04	-.05	0.10	0.99	0.95	0.83	0.99	1

(D) SSIM TABLE

MF 3X3	MF 5X5	MF 7X7	CWM	ACWM	DBMF	PSMF	UT	PF
0.23	0.33	0.34	0.67	0.98	0.9964	0.2236	0.99	1

(E) UQI TABLE

MF 3X3	MF 5X5	MF 7X7	CWM	ACWM	DBMF	PSMF	UT	PF
-.03	- .042	- .047	0.016	0.99	0.995	0.77	0.9	1

IV. CONCLUSION

In this paper, our algorithm is proposed which gives better performance in comparison with the median filter with mask of 3-by-3, 5-by-5, 7-by-7, the center weighted median (CWM) filter, the adaptive center weighted median (ACWM) filter, the progressive switching median (PSM) filter, the decision based median (DBM) filter, and the untrimmed (UT) median filter.At high noise density levels this algorithm gives better results in comparison with other existing algorithms. Also the quantitive parameters like Peak Signal-to-Noise Ratio (PSNR) and Mean square error (MSE), Correlation ratio (COR), universal quality index (UQI), Structural similarity index mean (SSIM) prove to be superior. Hence the proposed algorithm is effective for salt and pepper noise removal in images at high noise densities.

REFERENCES

[1] KrishnanNallaperumal,JustinVarghese,S.Saudia, K.Krishnaveni,Sri.S.Ramasamy,Santhosh.P.Mathew,P.Kumar “An efficient Switching Median Filter for Salt & Pepper Impulse Noise Reduction”, IEEE 2006, pp. 160-166

[2] Changhong Wang, Taoyi Chen, and Zhenshen Qu, “A novel improved median filter for salt-and-pepper noise from highly corrupted images”, IEEE 2010, pp. 718-722

[3] K. S. Srinivasan and D. Ebenezer, “A New Fast and Efficient Decision-Based Algorithm for Removal of High-Density Impulse Noises”, IEEE SIGNAL PROCESSING LETTERS, Vol. 14, No. 3, March 2007, pp. 189-192

[4] Sung-Jea KO, “Center Weighted Median Filters and Their Applications to Image Enhancement”, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, Vol. 38, No. 9, September 1991, pp. 984-993

[5] Zhou Wang and David Zhang, “Progressive Switching Median Filter for the Removal of Impulse Noise from Highly Corrupted Images”, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: ANALOG AND DIGITAL SIGNAL PROCESSING, Vol. 46, No. 1, January 1999, pp. 78-80

[6] 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, pp. 600-612.

[7] Megha.P.Arakeri, G.Ram Mohana Reddy “Comparative Performance Evaluation of Independent Component Analysis in Medical Image Denoising”, IEEE-International Conference on Recent Trends in Information Technology June 3-5, 2011, pp. 770-774.

[8] Zhou wang, Alan . C bovik, “A universal image quality index”, IEEE Signal processing letters, Vol. XX, No Y, March 2002, pp. 1-4

◈◈◈

Design and Implementation of Adaptive Filter for Better Throughput

Shivaraj S B & Bhagya R

Dept. of Telecom, RV College of Engineering, Bangalore, India

***Abstract* -** This paper proposes design and implementation of high throughput adaptive digital. The Filter structure is based on Fast Block LMS and Distributed Arithmetic (DA). With DA we can able to calculate inner product by shifting and accumulating of partial products and storing in Look up table, also the desired filter will be multiplier less. Thus DA based implementation of adaptive filter is highly computational, power and area efficient. DA based architecture map well to the today's Field Programmable Gate Arrays (FPGA) architecture. FPGA results conforms that proposed DA based filter requires 45% less area and 50% less power than that of FBLMS.

I. INTRODUCTION

Adaptive digital filters are widely used in the area of Signal processing such as echo cancelation, noise cancelation and channel equalization for communications and networking systems [1], [2]. The necessity of hardware implementation requires various of performances such as high speed, low power dissipation and good convergence characteristics. Fast block least mean square (FBLMS) algorithm proposed by Clark et al. [3] is one of the fastest and computationally efficient adaptive algorithm since here the process of filtering and adaption is done in frequency domain by using FFT algorithms. But there is still possibility to further enhanced the throughput of FBLMS algorithm based adaptive filters by applying the concept of Distributed arithmetic (DA) proposed by A. Peled and B. Liu [4]. Using bit level rearrangement of a multiply accumulate terms, DA can hide the complex hardware multipliers and therefore, the desired system becomes multiplier-less. DA is a powerful technique for reducing the size of a parallel hardware multiply-accumulate that is well suited for FPGA designs. Recently there has been a trend to implement DSP functions using FPGAs. While application specific integrated circuits (ASICs) are the traditional solution to high performance applications due to high development costs and time-to-market factor. The main reason behind the popularity of the FPGA is due to balance that FPGAs provide the designer in terms of flexibility, cost, and time-to-market. The concept of DA has already applied to LMS based adaptive filters [5], [6], [7], [8] but, not to FBLMS algorithm based adaptive filters. This paper proposes a new hardware efficient implementation of FBLMS algorithm based adaptive filter using DA.

II. THE EXISTING FAST BLOCK LMS (FBLMS) ADAPTIVE FILTER

Consider a BLMS based adaptive filter, that takes an input sequence x(n), which is portioned into non overlapping blocks of length P each by means of a serial to parallel converter, and the blocks of data so produced are applied to an FIR filter of length L, one block at a time. The tap weights of the filter are updated after the collection of each block of data samples. With the j-th block, (j ∈ Z) consisting of $x(jP+r)$, r ∈ Z= {0, 1..... P -1}, the filter coefficients are updated from block to block as,

$$w(j+1) = w(j) + \mu\sum_{r=0}^{P-1} x(jP+r)e(jP+r) \quad (1)$$

Where, $w(j) = [w_0(j)\; w_1(j)\; w_2(j) \ldots\ldots\ldots. w_{L-1}(j)]^t$ is the tap weight vector corresponding to j-th block $\boldsymbol{x(jP+r) = [x(jP+r)\; x(jP+r-1) \ldots\ldots x(jP+r-L+1)]^t}$**.**

And error at $n = (jP+r)$ is given by,

$$e(jP+r) = d(jP+r) - y(jP+r) \quad (2)$$

$d(jP+r)$ is the so called desired response available during initial training period.

Filter output $y(jP+r)$ at $n = (jP+r)$ is given by,

$$y(jP+r) = w^t(j)x(jP+r) \quad (3)$$

The parameter μ, popularly called the step size parameter is to be chosen as $0 \le \mu \le 2/P_{tr}$ for convergence of the algorithm. For the l -th sub-block within the i -th block, $0 \le l \le K-1$ i.e., for $n = (jP+r)$, $r = 0,1,2 \ldots\ldots\ldots . P-1, j = iK + l, filter\ output\ y(n) = w^t(j)x(n)$ is obtained by convolving the input data sequence $x(n)$ with the filter coefficient vector $w^t(j)$ and thus can be realized efficiently by the overlap-save method via M = L+ P -1 point FFT, where the first L -1 points come from the previous sub-block, for which the output is to be discarded. Similarly, the weight update term in (1) above, $\sum_{r=0}^{P-1} x(jP+r)e(jP+r)$ can be obtained by the usual circular correlation technique, by employing M point FFT and Setting the last P -1 output terms as zero.

III. DISTRIBUTED ARITHMETIC BACKGROUND

Consider the following inner product of two L dimensional vectors c and x, where c is a constant vector, x is the input sample vector, and y is the result.

$$y = \sum_{k=0}^{L-1} c_k x_k \qquad (4)$$

Using B-bit 2's complement binary representation scaled such that $|x_k| \le 1$ produces,

$$x_k = -b_{k0} + \sum_{n=1}^{B-1} b_{kn} 2^{-n} \qquad (5)$$

Where b_{kn} s are the bits (0 or 1) of x_k , b_{k0} is the most significant bit. Substituting (5) into (4) yields,

$$\mathbf{y} = \sum_{k=1}^{L-1} c_k \left[-b_{k0} + \sum_{n=0}^{B-1} b_{kn} 2^{-n} \right] \qquad (6)$$

$$y = -\sum_{k=1}^{L-1} c_k b_{k0} + \sum_{n=1}^{B-1} \left[\sum_{k=1}^{L-1} c_k b_{kn} \right] 2^{-n} \qquad (7)$$

The computation in distributed arithmetic is represented by (6). The values of b_{kn} s are either 0 or 1, resulting in bracketed term in (7) having only 2B possible values. Since c is a constant vector, the bracketed term can be recomputed and stored in memory using either lookup table (LUT) or ROM. The lookup table is then addressed using the individual bits of input samples, x_k with the final result y computed after B cycles, regardless of lengths of vectors c and x. A comprehensive tutorial review of DA linear filters is given in [9].

IV. PROPOSED IMPLEMENTATION

The throughput of FBLMS based adaptive filter is limited by computational complexity lies in FFT (and IFFT) block. It is possible to enhance the throughput of system by implantation of that FFT (and IFFT) block with reduced hardware complexity. DA is one of the efficient techniques, in which, by means of a bit level rearrangement of a multiply accumulate terms FFT can be implemented without multiplier. Since the main hardware complexity of the system is due to hardware multipliers and introduction of DA eliminates the need of that multipliers and resulting system will have high throughput and also have low power dissipation. There are many fast Fourier transform (FFT) algorithm like radix-2, Cooley-Tukey, Winograd, Good-Thomas, Rader etc. But using DA, FFT can be efficiently calculated by jointly employing the Good-Thomas and Rader algorithms. Good-Thomas algorithm re-expresses the discrete Fourier transform (DFT) of a size $N = N_1$ N_2 as a two-dimensional N_1 N_2 DFT, but only for the case where N_1 and N_2 are relatively prime. It is easily seen from the definition of the DFT that the transform of a length N real sequence x(n) has conjugate symmetry, i.e. X(N -K) = X*(K). This property facilitates to compute only half of the transform, as the remaining half is redundant and need not be calculated. Rader algorithm provides straightforward way to compute only half of the conjugate symmetric outputs without calculating the others, which is not possible with other algorithms like radix-2, Cooley-Tukey and Winograd. Algorithm presented here first decomposes the one dimensional DFT into a multidimensional DFT using the index map proposed by Good [10]. Next, a method which is based on the index permutation proposed by Rader [11] is used to convert the short DFTs into convolution. This method changes a prime length N DFT of real data into two convolutions of length (N - 1)/2. One convolution is cyclic and the other is cyclic or skew-cyclic. The index mapping suggest by Good and Thomas for n is,

$$n = N_2 n_1 + N_1 n_2 \bmod N \left\{ \begin{matrix} 0 \le n_1 \le N_1 - 1 \\ 0 \le n_2 \le N_2 - 1 \end{matrix} \right\} \qquad (8)$$

And as index mapping for k results,

$k = N_2\langle N_2^{-1}\rangle_{N_1} k_1 + N_1\langle N_1^{-1}\rangle_{N_2} k_2 \; modN$ (9)

$0 \le k_1 \le N_1 - 1,\ 0 \le k_2 \le N_2 - 1$

If we substitute the Good-Thomas index map in the equation for DFT matrix it follows

$$X[k_1, k_2] = \sum_{n_2=0}^{N_2-1} W_{N_2}^{n_2 k_2} \left(\sum_{n_1=0}^{N_1-1} x[n_1, n_2] W_{N_1}^{n_1 k_1} \right) \qquad (10)$$

$$= \sum_{n_2=0}^{N_2-1} x[n_2, k_1] W_{N_2}^{n_2 k_2} \qquad (11)$$

Steps for Good-Thomas FFT Algorithm,

An N = $N_1 N_2$ point DFT can be computed according to following steps:

(1) Index transform of input sequence, according to (8).

2) Computation $N_2$of DFTs of length N_1 using Rader algorithm

3) Computation of N_1 DFTs of length N_2 using Rader algorithm.

4) Index transform of input sequence, according to (9).

Consider the length N = 14, suppose we have N_1 = 7 and N_2= 2 then mapping for the input index according to (n = $2n_1 + 7n_2$mod 14) and ($k = 4k_1 + 7k_2$mod 14) for output index results and using these index transforms we can construct the signal flow graph as shown in Fig. 1

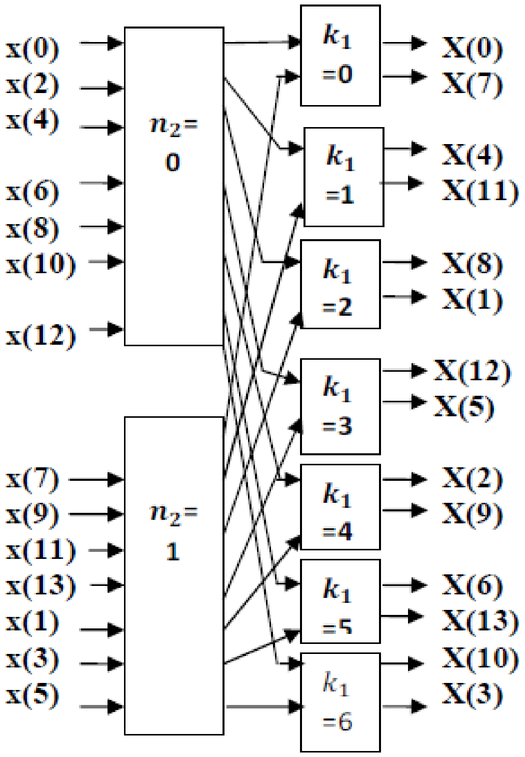


Fig.1 Mapping using Good-Thomas algorithm.

From figure.1 we realize that first stage has 2 DFTs each having 7 -points and second stage has 7 DFTs each having of length 2.

One of the interesting thing here is multiplication with twiddle factors between the stages is not required. Now consider N_1 = 7 if the data are real we need to calculate only half of the transform. Also, as Rader showed the zero frequency term must be calculated separately.

In matrix t form, we write,

$$\begin{bmatrix} X(1) \\ X(2) \\ X(3) \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 6 & 1 & 3 & 5 \\ 3 & 6 & 2 & 5 & 1 & 4 \end{bmatrix} \begin{bmatrix} x(1) \\ x(2) \\ x(3) \\ x(4) \\ x(5) \\ x(6) \end{bmatrix} + \begin{bmatrix} x(0) \\ x(0) \\ x(0) \end{bmatrix} \qquad (12)$$

Replacing W^k by $W^{(n-k)\star}$

$$\begin{bmatrix} X(1)\\ X(2)\\ X(3)\end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 3^* & 2^* & 1^*\\ 2 & 3^* & 1^* & 1 & 3 & 2^*\\ 3 & 1^* & 2 & 2^* & 1 & 3^*\end{bmatrix} \begin{bmatrix} x(1)\\ x(2)\\ x(3)\\ x(4)\\ x(5)\\ x(6)\end{bmatrix} + \begin{bmatrix} x(0)\\ x(0)\\ x(0)\end{bmatrix} \qquad (13)$$

If real and imaginary parts of W matrix in (13) are separated, a simplification is possible. Consider first the real part using notation in matrix that k stands for cos ($2\pi k/7$). The real part of (13) becomes,

$$\begin{bmatrix} X_R(1)\\ X_R(2)\\ X_R(3)\end{bmatrix} = \begin{bmatrix} 1 & 2 & 3\\ 2 & 3 & 1\\ 3 & 1 & 2\end{bmatrix} \begin{bmatrix} x(1)+x(6)\\ x(2)+x(5)\\ x(3)+x(4)\end{bmatrix} + \begin{bmatrix} x(0)\\ x(0)\\ x(0)\end{bmatrix} \qquad (14)$$

Using the notation of k for sin (2πk/7) gives, for imaginary part of (13).

$$\begin{bmatrix} X_I(1)\\ X_I(2)\\ X_I(3)\end{bmatrix} = \begin{bmatrix} -1 & -2 & 3\\ -2 & 3 & -1\\ 3 & -1 & -2\end{bmatrix} \begin{bmatrix} x(1)-x(6)\\ x(2)-x(5)\\ x(3)-x(4)\end{bmatrix} \qquad (15)$$

The (14) and (15) are cyclic convolution relation. Since in our problem we always convolve with the same coefficients (In case of DFT it is twiddle factor matrix), arithmetic efficiency can be improved by pre calculating some of the intermediate results. These are stored in table in memory and simply addressed as needed. Using distributed arithmetic this can be implemented efficiently and detail can be found in [12]. Here it will present only the structure Fig. 2 best suited to the DFT calculated by cyclic convolution. Initially R_1 to R_7 are cleared to zero and the x_i' s are loaded into registers R_1 to R_3 after addition. Then all R_1 to R_3 are shifted by one bit, the last bit of each register is in R_4. The ROM output will be added to R_5. Circular shift of R_4 produces at ROM output are added to R_1 to R_7. when the first cycle is completed content of all R_1 to R_7 except R_4 are right shifted by one bit and the second cycle starts. At B-th cycle function at ADD _SUB changed from adder to subtraction and after this B-th cycle the content of R5 to R7 gives final FFT coefficient and zero frequency component can be calculated separately applying accumulate and addition of x_i's as shown in Fig. 2 and the FFT block so obtained has no multipliers at the expanse of increased adder requirement and memory requirement in order to store some pre calculated values.

AS ADD_SUB

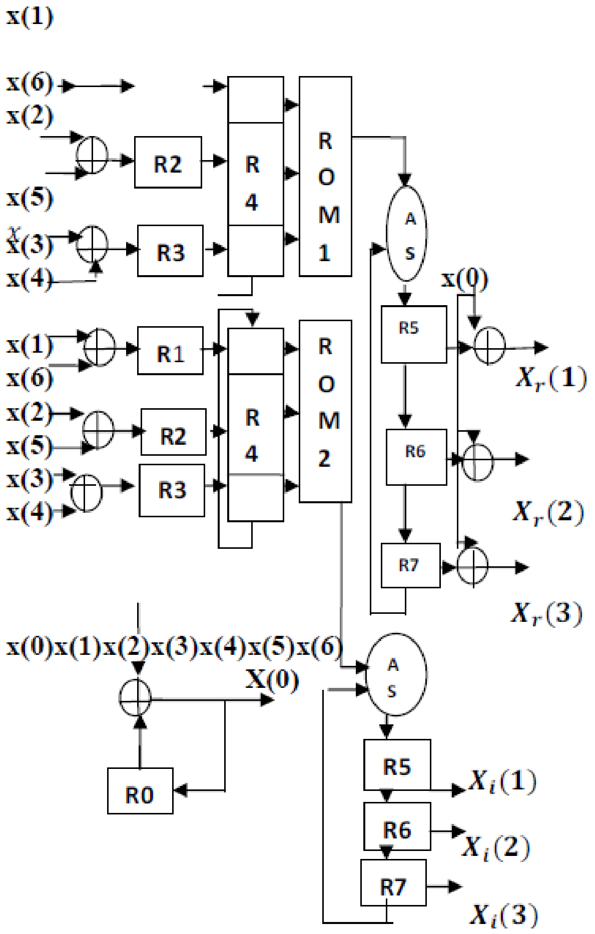


Fig. 2 Architecture for FFT using DA.

In the proposed architecture, using the DA based FFT (and IFFT) block that is described above, we recast the existing FBLMS based adaptive filter. Since, the DA based FFT block provides only half of the conjugate symmetric outputs without calculating others (reaming can calculated by conjugating them) and in DA based IFFf block we require to feed only half of the conjugate symmetric coefficients not all, which is not possible in existing FBLMS based adaptive filters since here radix-2 based FFT (and IFFT) blocks are employed and in radix-2 there is no such facility to calculate only half of the conjugate symmetric outputs hence, in this case half of the processed data are redundant. Since in our proposed FBLMS algorithm based adaptive filter, we require to calculate only half of the conjugate symmetric coefficients, under that condition the hardware requirements for our proposed system is approximately half of that of existing one. In our proposed architecture shown in Fig. 3. the computation of frequency domain outputs requires 8(N/2+ 1) or 8(M + 1) number of multiplication and required number of addition is 16N + 2.5(N1 + N2) + 2, here total number of multiplications are much less than that of required number of multiplications in the existing FBLMS algorithm based adaptive filter (can be compared from Table I and Table II) at the expanse of increased memory and adder requirement, which drastically reduces the hardware complexity for higher order filters as shown in Fig. 4 and it results with a adaptive filter which has high throughput and low power dissipation with reduced area requirement.

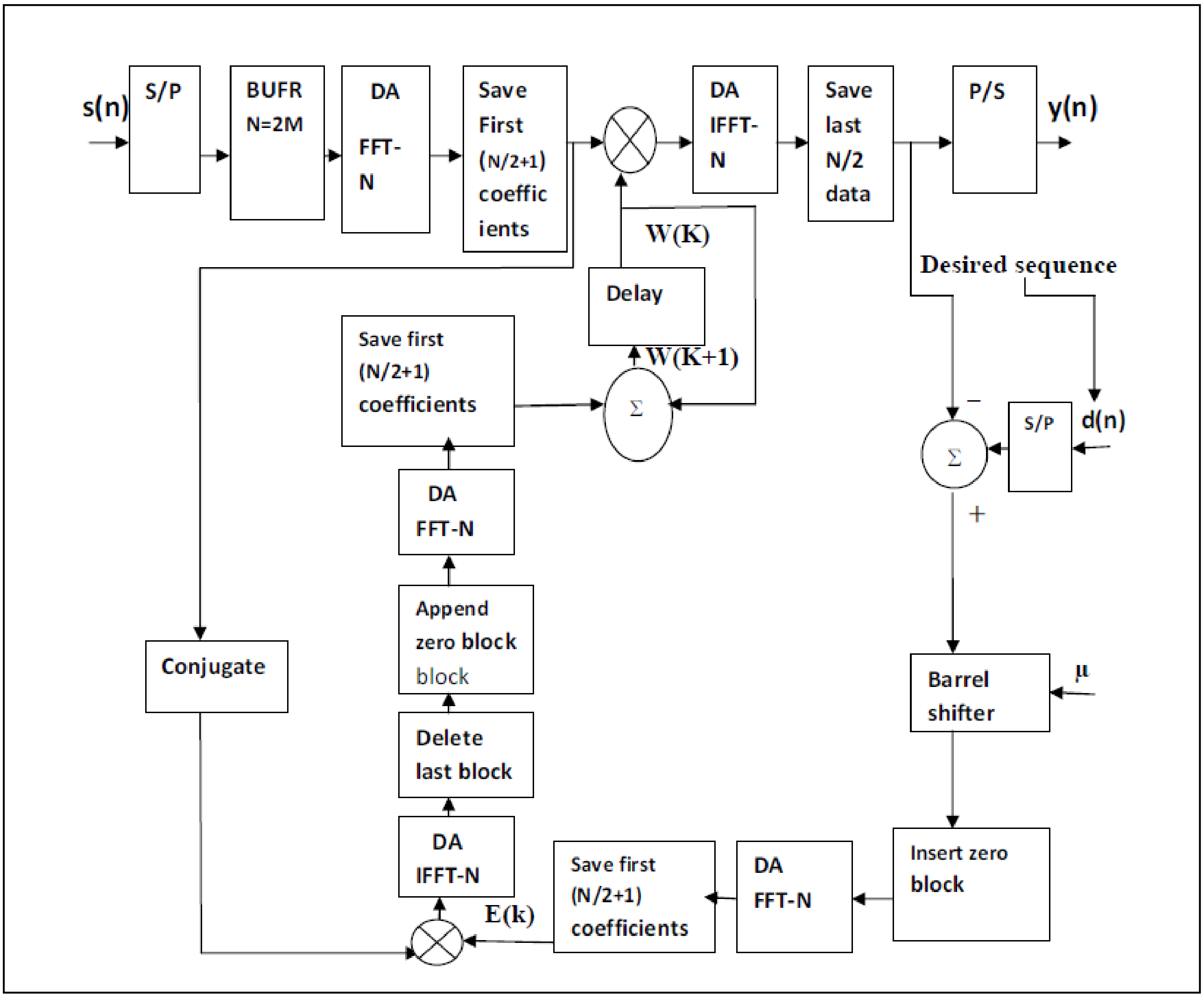


Fig.3 Proposed DA based FBLMS after optimization

Table I COMPUTATIONAL COMPLEXITY IN EXISTING FBLMS ALGORITHM.

Multipliers	10M $log_2M + 26M$

Adders	2N+5N$log_2N + 1$

Table II COMPUTATIONAL COMPLEXITY IN PROPOSED FBLMS ALGORITHM.

Multipliers	$8(N/2 + 1)1$
Adders	16N+2.5$(N_1 + N_2)$

V EXPERIMENTAL RESULTS

Proposed and existed architectures are implemented for filter length-7 and length-8 respectively. Verilog codes are written for both of these designs and synthesized using Xilnx 10.1 version. Family of device was Virtex II Pro and target device was 2vpl 00-ff1696-6. Fig. 5 shows the logic utilization of both the architectures. Although the filter length used in case of existed system is 1 more then that of our proposed one, but the FPGA resource utilization and power utilization in case of existed system was approximately 45% greater than that of proposed one which can be clearly seen from Fig. 5 and 6. From Table III it is clear that our proposed architecture based adaptive filter is 30.2% faster than that of existed one.

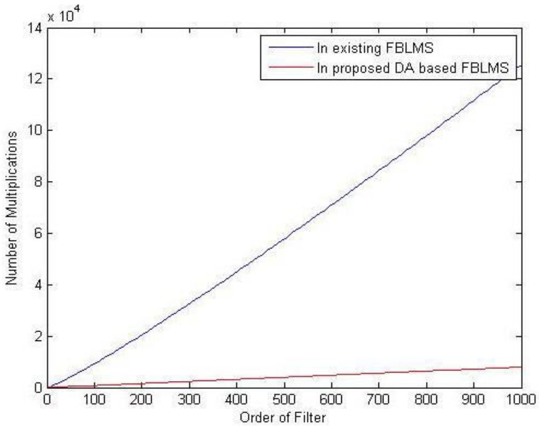


Fig.4 Comparison of hardware complexity

Table III : Timing comparison of Presented Architectures

Design	Delay (ns)
Existing Architecture	9.61
Proposed Architecture	6.709

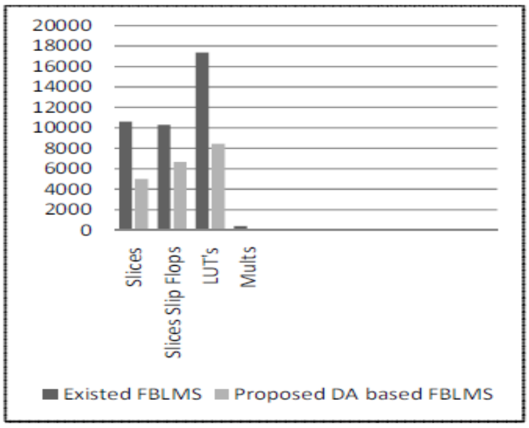
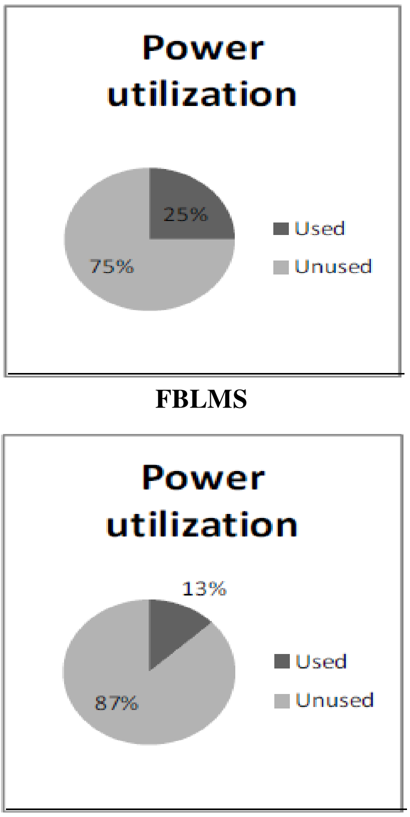


Fig.5 Comparison of resources utilization in FPGA

VI. CONCLUSION

In this paper it is proposed a new hardware-efficient adaptive filter structure for very high throughput FBLMS adaptive filters and its implementation details were presented. The concept of DA involves for implementation of FFT block without any hardware multiplier using LUT and adders. Due to reduced hardware complexity the proposed DA based FBLMS adaptive filter is best suitable for implementation of higher order filters in FPGA efficiently with minimum area requirement, low power dissipation and high throughput.



FBLMS

DA based FBLMS

Fig. 6 Comparison of Power utilization

REFERENCES

[I] S. Haykin, Adaptive Filter Theory, 4th ed., T. Kailath, Ed. Pearson Education, 2008.

[2] B. Farhang-Boroujeny, Adaptive Filters: Theory and Applications, Chichester, Ed. Wiley, 1998.

[3] S. K. M. Gregory A. Clark and S. R. Parker, "Block implementation of adaptive digital filters," IEEE Transactions on Circuits and Systems,vol. 28, pp. 584 - 592, 1981.

[4] A. Peled and B. Liu, "A new hardware realization of digital filters," IEEE Transactions On Acoustics, Speech, And Signal Processing, vol. 22, pp. 45□6, December 1974.

[5] c. H. Wei and 1. 1. Lou, "Multi memory block structure for implementing a digital adaptive filter using distributed arithmetic," lEE Proceedings, Electronic Circuits and Systems, vol. 133, February 1986.

[6] DJ Allred, W. Huang, Y.Krishnan, H. Yoo, D.V Anderson, "An FPGA implementation for a high throughput adaptive filter using distributed arithmetic." 12th Annual IEEE Symposium on Field-Programmable Custom Computing Machines, pp. 324 - 325, 2004.

[7] Daniel 1. Allred, Walter Huang, Venkatesh Krishnan, Heejong Yoo, and David Y. Anderson, "LMS adaptive filters using distributed arithmetic for high throughput," IEEE Transactions on Circuits and Systems, vol. 52, pp. 1327 - 1337, July 2005.

[8] N. J. Sorawat Chivapreecha, Aungkana Jaruvarakul and K. Dejhan, "Adaptive equalization architecture using distributed arithmetic for partial response channels," IEEE Tenth International Symposium on Consumer Electronics, 2006.

[9] S. A. White, "Applications of distributed arithmetic to digital signal processing: A tutorial review," IEEE ASSP Magazine, July 1989.

[10] C. S. Burrus, "Index mappings for multidimensional formulation of the DFT and convolution," IEEE Transactions On Acoustics, Speech, And Signal Processing, vol. 25, pp. 239-242, June 1977.

[11] C. M. Rader, "Discrete fourier transforms when the number of data samples is prime," IEEE Proceedings, vol. 56, pp. 1107-1108, June 1968.

[12] S. Chu and C. S. Burrus, "A prime factor FFT algorithm using distributed arithmetic," IEEE Transactions On Acoustics, Speech, And Signal Processing, vol. 30, April 1982.

◈◈◈

Backbone Node Election In Presence of Misbehaving Node Using Mechanism Design Approach In MANET

Manasa K M, Shanthi M B & Jitendranath Mungara

Dept of Computer science & Engg. CMRIT, Bangalore, India
Visvesvaraya Technological University, Bangalore, India
E-Mail: kmmanasa19@gmail.com, gowri.pc@gmail.com

Abstract - The proposed system highlight about a robust scheme of selection of cluster leader in mobile adhoc network. The proposed scheme has security consideration where the evaluation is carried out in presence of selfish node for intrusion detection system. The issue of unwanted resource consumption is addressed for enhancing the network lifetime in mobile adhoc network. There are two challenges considered in the design e.g. first without any incentives for serving others, a mobile node may behave selfishly by furnishing false information about the residual resources in order to avoid participation in cluster leader election process. Secondly, a restricted overhead in network performance may occur while choosing an efficient cluster leader. The proposed system has considered reverse game theory for mitigating issues with selfish node. Simulation results show that after adopting the proposed system, the above discussed issues has fairer chances of minimization.

Keywords*-component; Selfish Node, Intrusion detection system, mobile adhoc network.*

I. INTRODUCTION

A mobile ad hoc network (MANET) is a self-configuring network that is formed automatically by a collection of mobile nodes without the help of a fixed infrastructure or centralized management. Each node is equipped with a wireless transmitter and receiver, which allow it to communicate with other nodes in its radio communication range. There are both passive and active attacks in MANETs. For passive attacks, packets containing secret information might be eavesdropped, which violates confidentiality. Active attacks, including injecting packets to invalid destinations into the network, deleting packets, modifying the contents of packets, and impersonating other nodes violate availability, integrity, authentication, and non-repudiation. Proactive approaches such as cryptography and authentication [1][2][3][4] were first brought into consideration, and many techniques have been proposed and implemented. However, these applications are not sufficient. If we have the ability to detect the attack once it comes into the network, we can stop it from doing any damage to the system or any data. Here is where the intrusion detection system comes in. The various architecture of intrusion detection system in mobile adhoc network is Stand-alone Intrusion Detection Systems, Distributed and Cooperative Intrusion Detection Systems [5], Hierarchical Intrusion Detection Systems and Mobile Agent for Intrusion Detection Systems [6].

Authentication of entities and messages can be realized in different ways using either symmetric (3DES, AES) or asymmetric (ElGamal, RSA) cryptographic algorithms (see e.g. [7] for details). In order to protect the secret from attackers that move around and compromise multiple share holders over a long period of time, a proactive secret sharing (PSS) scheme should be used in ad hoc networks. In PSS schemes, secret shares are changed periodically without changing the secret itself, so an attacker cannot use a secret's whole lifetime to compromise k participants. All information an attacker collected about the secret becomes worthless after refreshing the shares [8]. Threshold shared secret schemes can be transformed into PSS schemes using discrete logarithms [8]. Proactive digital signatures, which are used in our work, are an implementation of PSS schemes [9], [10].

IDS solutions for fixed wired networks are often hierarchical and deploy network-based sensors at key traffic concentration points, such as switches, routers, and firewalls. These IDS sensors are physically secured, and use the signature-based detection technique to detect attacks. Alerts generated by these distributed IDS sensors are sent to centralized security servers for analysis and correlation. The centralized security server distributes attack signature updates to the network-based IDS sensors. The effectiveness of IDS solutions that were designed for fixed wired networks are limited for wireless ad-hoc networks as described below:

- Wireless ad-hoc networks lack key concentration points where network traffic can be monitored. This limits the effectiveness of a network-based IDS sensor, since only the traffic generated within radio transmission range may be monitored.
- In a dynamically changing ad-hoc network, it may be difficult to rely on the existence of a centralized server to perform analysis and correlation.
- The secure distribution of signatures may be difficult, due to the properties of wireless communication and mobile nodes that operate in disconnect mode.
- It may be difficult to physically secure a mobile host that could be captured, compromised, and later rejoin the network as a Byzantine node.

The proposed system highlights a solution for balancing the resource consumption of IDSs among all nodes while preventing nodes from behaving selfishly. To address the selfish behavior, we design incentives in the form of reputation to encourage nodes to honestly participate in the election scheme by revealing their cost of analysis. The cost of analysis is designed to protect nodes' sensitive information (resources level) and ensure the contribution of every node on the election process (fairness). In section 2 we give an overview of related work which identifies all the major research work being done in this area. Section 3 highlights proposed system. Implementation is discussed in Section 4 followed by results and performance analysis in Section 5. Section 6 makes some concluding remarks.

II. RELATED WORK

Zachary K. Baker and Viktor K. Prasanna [11] present a tool for automatic synthesis of highly efficient intrusion detection systems using a high-level, graph-based partitioning methodology, and tree-based look ahead architectures.

Amritha Sampath et.al [12] presents an effective algorithm for selecting cluster heads in mobile ad hoc networks using ant colony optimization.

ZHANG Jian, DING Yong, and GONG Jian [13] applies fuzzy default theory to transform reasoning and response engine of IDS, based on the proving of IDS as non-monotonic, and set up an intelligent IDSFDL-IDS.

Ashish Bagwari and Raman Jee [14] proposed The Criteria Require for Cluster Head Gateway Selection in Integrated Mobile Ad hoc Network. They also provide the criteria for Cluster Head Gateway (CHG) selection in Mobile Ad hoc network.

Caleb C. Noble and Diane J. Cook [15] introduce two techniques for graph-based anomaly detection. In addition, they introduce methods for calculating the regularity of a graph, with applications to anomaly detection.

T. Shivaprakash et.al [16] has proved that Passive Clustering becomes practically possible by implementing the intelligent gateway selection heuristic and on-demand timeout mechanism.

Thomas Guyet et.al [17] presents a self-adaptive intrusion detection system which relies on a set of local model-based diagnosers. The redundancy of diagnoses is exploited, online, by a meta-diagnoser to check the consistency of computed partial diagnoses, and to trigger the adaptation of defective diagnoser models (or signatures) in case of inconsistency.

NGAI Cheuk Han [18] presents a public key authentication service to protect security in the network in the presence of malicious nodes. They develop a novel authentication service based on trust and clustering models.

James Cannady et.al [19] presents an analysis of the progress being made in the development of effective intrusion detection systems for computer systems and distributed computer networks.

Jens Tölle and Oliver Niggemann [20] presents a description of a system supporting the detection of intrusions and network anomalies by analyzing and visualizing traffic flows in computer networks

Harley Kozushko [21] believes that combined network-based and host-based intrusion detection systems effectively prevent attacks from insider as well as outsider sources.

S. Staniford-Chen et.al [22] presents the design of GrIDS (Graph-Based Intrusion Detection System). GrIDS collect data about activity on computers and network traffic between them.

Scott Fazackerley et.al [23] presents the LEACH algorithm for selecting cluster heads is a probabilistic method which produces clusters with a large variation of link distances and uneven energy consumption during the data transmission phase. To address this issue, a RF signal strength algorithm based on link quality is presented.

Brian Tung [24] describes the construction of operators that combine graphs from two or more systems into one graph.

Dang Nguyen et.al [25] investigate the problems of cluster head selection for large and dense MANETs Two variants of the cluster head selection are examined: (1) the distance-constrained selection where every node in the network must be located within a certain distance to the nearest cluster head; and (2) the size-constrained selection where each cluster is only allowed to have a limited number of members.

Saira Beg et.al [26] surveyed the effectiveness and upcoming challenges of security needs in a network environment, especially a larger one. They considered different tools of security, their types and detection schemes along with summarizing basic details of Firewall, IDS, IPS and IDPS.

III. PROPOSED SYSTEM

The main aim of the project work is to develop an architectural framework to elect leader in the presence of selfish nodes for intrusion detection in mobile ad hoc networks (MANETs). To balance the resource consumption among all nodes and prolong the lifetime of an MANET, nodes with the most remaining resources should be elected as the leaders. In our proposed project, we propose a solution for balancing the resource consumption of IDSs among all nodes while preventing nodes from behaving selfishly. To address the selfish behavior, we design incentives in the form of reputation to encourage nodes to honestly participate in the election scheme by revealing their cost of analysis. The cost of analysis is designed to protect nodes' sensitive information (resources level) and ensure the contribution of every node on the election process (fairness). To motivate nodes in behaving normally in every election round, we relate the amount of detection service that each node is entitled to the nodes' reputation value. Besides, this reputation value can also be used to give routing priority and build a trust environment. The design of incentives is based on a classical mechanism design model, namely, Vickrey, Clarke, and Groves (VCG).

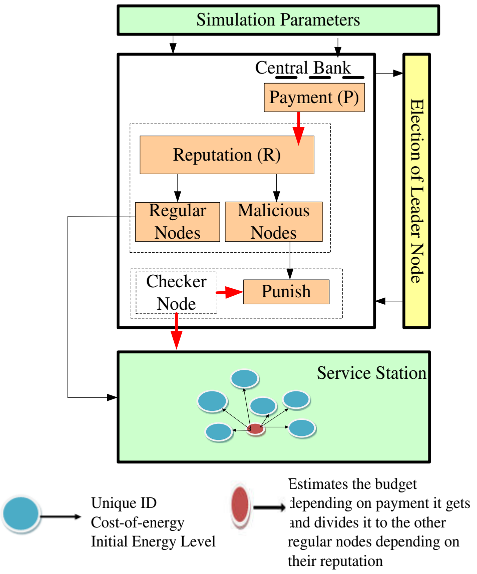


Figure 1. Proposed Architecture

The model guarantees that truth-telling is always the dominant strategy for every node during each election phase. On the other hand, to find the globally optimal cost-efficient leaders, a leader election algorithm is devised to handle the election process, taking into consideration the possibility of cheating and security flaws, such as replay attack. The algorithm decreases the percentage of leaders, single-node clusters, and maximum cluster size, and increases average cluster size. Last but not least, we address these issues in two possible settings, namely, Cluster-Independent Leader Election (CILE) and Cluster-Dependent Leader Election (CDLE). In the former, the leaders are elected according to the received votes from the neighbor nodes. The latter scheme elects leaders after the network is formulated into multiple clusters. In both schemes, the leaders are elected in an optimal way in the sense that the resource consumption for serving as IDSs will be balanced among all nodes overtime. Finally, we justify the correctness of proposed methods through analysis and simulation. Empirical results indicate that our scheme can effectively improve the overall lifetime of an MANET. The main contribution of this paper is a unified model that is able to: 1) balance the IDS resource consumptions among all nodes by electing the most cost-efficient leaders and 2) motivate selfish nodes to reveal their truthful resources level. To design the leader election algorithm, the following requirements are needed:

- To protect all the nodes in a network, every node should be monitored by a leader and
- To balance the resource consumption of IDS service, the overall cost of analysis for protecting the whole network is minimized. In other words, every node has to be affiliated with the most cost-efficient leader among its neighbors.

The main goal of using mechanism design is to address this problem by: 1) designing incentives for players (nodes) to provide truthful information about their preferences over different outcomes and 2) computing the optimal system-wide solution, which is defined according to (1). A malicious node can disrupt our election algorithm by claiming a fake low cost in order to be elected as a leader. Once elected, the node does not provide IDS services, which eases the job of intruders. On the other hand, if node i still wins, then its utility remains the same since the payment does not depend on the value it reports. Second, suppose the real valuation function ci of node i is not the lowest, then reporting a higher value will never help the node to win. Last but not least, the checkers are able to catch and punish the misbehaving leaders by mirroring a portion of its computation from time to time. A caught misbehaving leader will be punished by receiving a negative payment. Thus, it discourages any elected node from not carrying out its responsibility. We can thus conclude that our mechanism is truthful and it guarantees a fair election of the most cost-efficient leader.

To execute the election mechanism, a leader election algorithm is proposed which assists to elect the most cost-efficient leaders with less performance overhead compared to the network flooding model. We devise all the needed messages to establish the election mechanism taking into consideration cheating and presence of malicious nodes. Moreover, we consider the addition and removal of nodes to/from the network due to mobility reasons. Finally, the performance overhead is considered during the design of the given algorithm where computation, communication, and storage overhead are derived.

IV. IMPLEMENTATION

The proposed system is designed on 32 bit Linux (Fedora 8) with 1.84 GHz processor and 2 GB of RAM. The proposed system is simulated using Network Simulator 2 (NS2). To implement the proposed system, the energy model is used to evaluate the influence of running IDS. Initially, we randomly assign 60 to 100 joules to each node. The proposed system assumes that the energy required for running the IDS for one time slot as 10 joules. The energy required to live and transmit packets to capture the silent aspect of the problem is ignored. The transmission radius of each node to 200 meters is configured. Two nodes are assumed as neighbor nodes if their Euclidean distance is less than or equal to 200 meters. Besides, we deploy different number of nodes, which varies from 20 to 50 in an area of 500 x 500 square meters. It helps us to measure the performance of the nodes from sparse networks to dense networks.

V. RESULTS & PERFORMANCE ANALYSIS

The main theme of this project work is to elect leader in the presence of selfish nodes for intrusion detection in mobile ad hoc networks, which is achieved by viewing the different node activities inside the network and the specified way has to be mentioned to monitor those activities. Performance analysis is done to find out whether to balance the resource consumption among all nodes and prolong the lifetime of an MANET, nodes with the most remaining resources should be elected as the leaders or not. It is essential that the process of performance analysis and definition must be conducted in parallel. This section will discuss about the results accomplished as well as comparative analysis of the results accomplished.

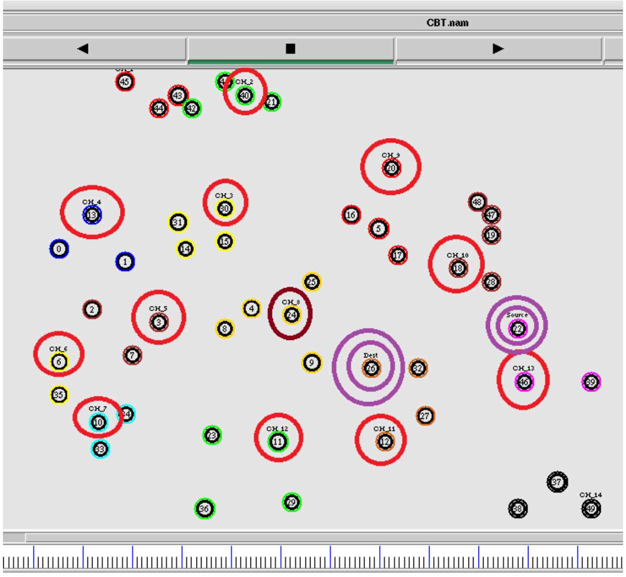


Figure 2. NS2 interface showing mobile of the nodes

The above figure 2 shows the mobile nodes, where it is shown basically three types of nodes e.g. feasible channels for communication, source, and destination node.

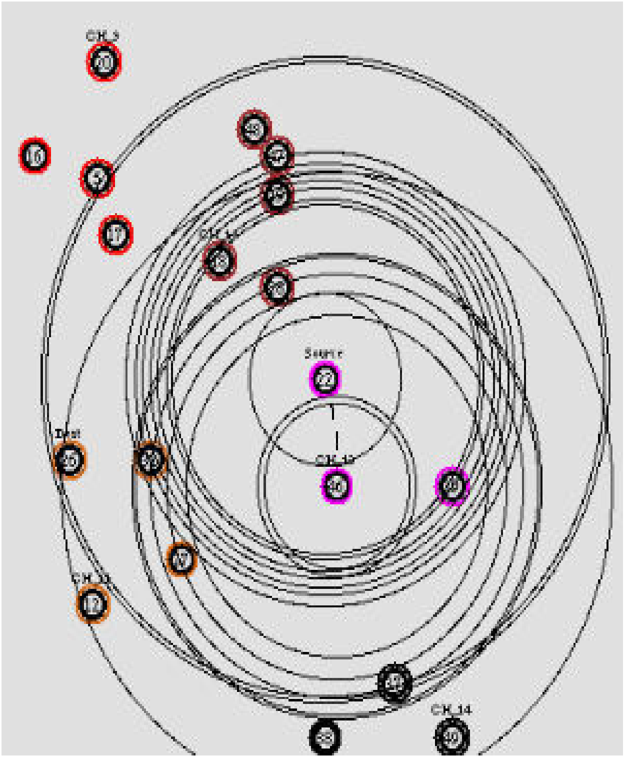


Figure 3 Communication from source node to nearest cluster leader node.

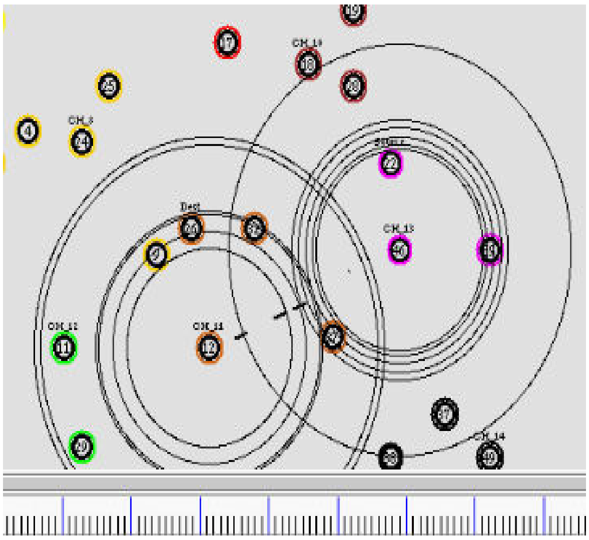


Figure 4. Communication from previous cluster leader node to next cluster leader node.

The main motive of the experiment is to analyze the effect of node selection for intrusion detection system on the life of all nodes. In order to show the negative impact of selfish node, dual experiments has been conducted: Duration taken for the first node to die and proportion of packet analysis. The considered metrics used are Percentage of alive nodes, power level of nodes, fractions of leader node, mean cluster size, upper limit cluster size, and quantity of single-node clusters. The mobile nodes can behave selfishly before and after the election. A node shows selfishness before election by refusing to be a leader. On the other hand, selfishness after election is considered when nodes misbehave by not carrying out the detection service after being a leader. Both kinds of selfishness have a serious impact on the normal nodes.

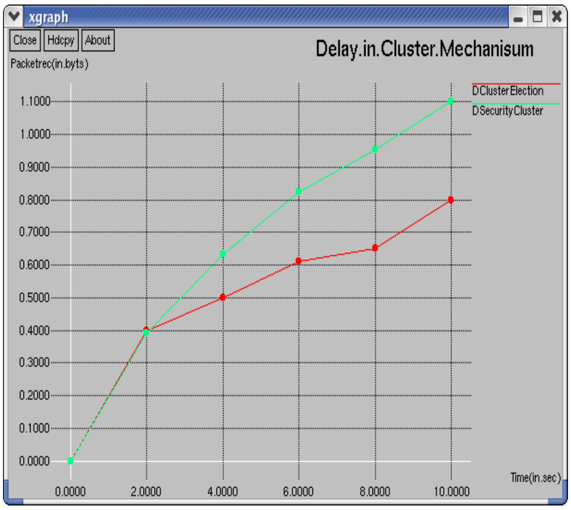


Figure 5. Delay in Clustering mechanism

Figure 5 shows the simulation results for delay in clustering mechanism for both traditional clustering and clustering using proposed security techniques. Due to maximum size of cluster, the security of the proposed technique can be assured.

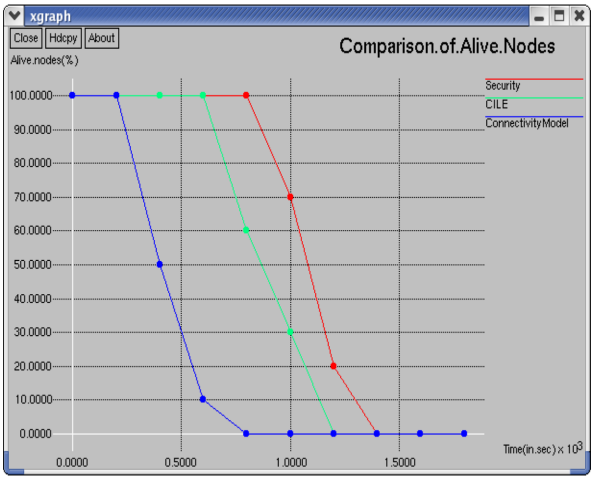


Figure 6 Comparative analysis of Alive nodes

Figure 6 highlights comparative analysis of alive nodes for cluster leader independent election, connectivity model, as well as security aspect considered. The proposed framework is compared with the connectivity model since the expected performance of the random model can be expected to be close to the one given with low mobility. The above graphical representation highlights that more nodes are alive in the proposed framework as compared to the connectivity one. As the quantity of mobile nodes maximizes, the life of nodes also increases since there are more nodes to act as leaders. Thus, the detection service is distributed among the nodes which prolongs the live time of the nodes in mobile adhoc network.

Fig. 7 compares the average cluster size of both the models for different number of nodes. The proposed model has a higher average cluster size than the other one, which proves that the proposed framework is able to uniformly distribute the load of the leaders. The graphical representation also explains the size of the maximum cluster. The upper limit of cluster size for both models is increasing with the number of nodes. For our model, the maximum cluster size is less, and thus, avoids many problems, such as message collisions, transmission delays, etc. This could also improve the detection probability since more number of packets are analyzed per node compared to the other model. Moreover, the proposed model is able to reduce the number of single-node clusters as the density of nodes is increasing.

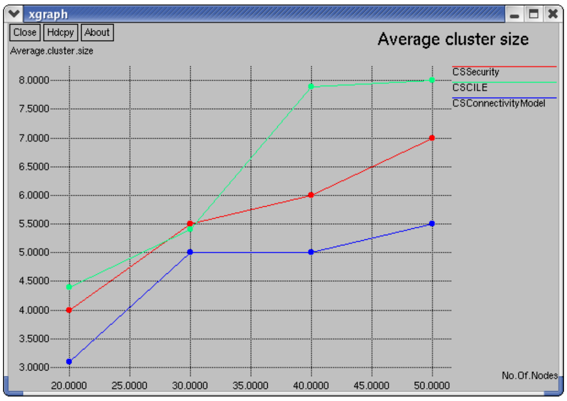


Figure 7 Average Cluster Size

Fig. 8 shows the impact of selfishness after election on security. We consider the presence of 20 percent of selfish nodes out of 10 nodes. As selfish nodes do not exhaust energy to run the IDS service, it will live longer than the normal nodes.

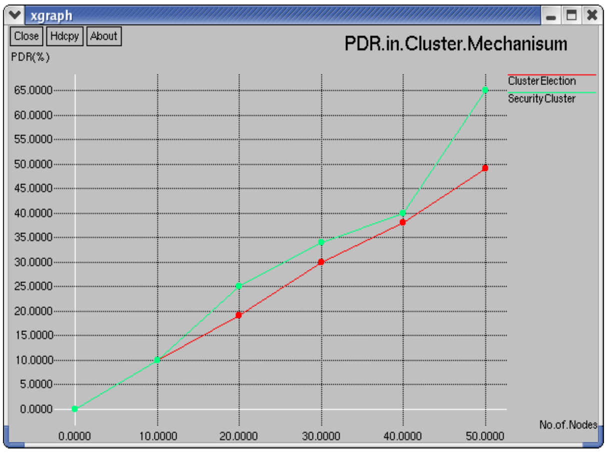


Figure 8 Comparative analysis for Packet Delivery Ratio

Thus, the more the time goes, the more the chances that the selfish node will be the leader node. Hence, the percentage of packet analysis decreases with time, which is shown in above figures. This is a severe security concern since fewer packets are analyzed.

CONCLUSION

The unbalanced resource utilization of intrusion detection system in mobile adhoc network and the presence of selfish nodes have motivated us to propose an integrated solution for prolonging the lifetime of mobile nodes and for preventing the emergence of selfish nodes. The solution motivated nodes to truthfully elect the most cost-efficient nodes that handle the detection duty on behalf of others. Moreover, the sum of the elected leaders is globally optimal. To achieve this goal, incentives are given in the form of reputations to motivate nodes in revealing truthfully their costs of analysis. To implement our mechanism, we devised an election algorithm with reasonable performance overheads. We also provided the algorithmic correctness and security properties of our algorithm. Simulation results showed that the proposed model is able to prolong the lifetime and balance the overall resource consumptions among all the nodes in the network. Moreover, we are able to decrease the percentage of leaders, single-node clusters, and maximum cluster size, and increase the average cluster size. These properties allow us to improve the detection service through distributing the sampling budget over less number of nodes and reduce single nodes to launch their intrusion detection system.

REFERENCES

[1] M. G. Zapata, \Secure Ad Hoc On-Demand Distance Vector (SAODV) Routing," ACM Mobile Computing and Communication Review (MC2R), Vol. 6, No. 3, pp. 106-107, July 2002.

[2] Y. Hu, D. B. Johnson, and A. Perrig, \SEAD: Secure E±cient Distance Vector Routing for Mobile Wireless Ad Hoc Networks," Proceedings of the 4th IEEE Workshop on Mobile Computing Systems and Applications (WMCSA'02), pp. 3-13, June 2002.

[3] Y. Hu, A. Perrig, and D. B. Johnson, \Ariadne: A secure On-Demand Routing Protocol for Ad hoc Networks," Proceedings of the 8th Annual International Conference on Mobile Computing and Networking (MobiCom'02), pp. 12-23, September 2002.

[4] A. Perrig, R. Canetti, D. Tygar and D. Song, \The TESLA Broadcast Authentication Protocol," RSA CryptoBytes, 5 (summer), 2002.

[5] Y. Zhang, W. Lee, and Y. Huang, \Intrusion Detection Techniques for Mobile Wireless Networks," ACM/Kluwer Wireless Networks Journal (ACM WINET), Vol. 9, No. 5, September 2003.

[6] A. Mishra, K. Nadkarni, and A. Patcha, \Intrusion Detection in Wireless Ad Hoc Networks," IEEE Wireless Communications, Vol. 11, Issue 1, pp. 48-60, February 2004.

[7] B. Schneier, “Applied Cryptography” John Wiley, 1996.

[8] A. Herzberg, M. Jakobsson, S. Jarecki, H. Krawczyk, and M. Yung, “Proactive public key and signature systems,” in ACM Conf. on Computer and Comm. Security, Zürich, 1997.

[9] A. Herzberg, S. Jarecki, H. Krawczyk, and M. Yung, “Proactive secret sharing, or: How to cope with perpetual leakage,” in Advances in Cryptology, Proc. CRYPTO’95, ser. LNCS, vol. 936. Santa Barbara, California: Springer-Verlag, Aug. 1995, pp. 339–352.

[10] K. Takaragi, K. Miyazaki, and M. Takahashi, “A threshold digital signature issuing scheme without secret communication,” IEEE P1363 Study, Nov. 2000.

[11] Zachary K. Baker and Viktor K. Prasanna, “Automatic Synthesis of Efficient Intrusion Detection Systems on FPGAs” proceeding of the 14th annual international conference on field-programmable logic and application (FPL-‘04).

[12] Amritha Sampath, Tripti. C, Sabu M. Thampi, “An ACO Algorithm for Effective Cluster Head Selection” Department of Computer Science and Engineering Rajagiri School of Engineering and Technology, Kochi, India

[13] Zhang Jian, Ding Yong, and Gong Jian, “Intrusion Detection System based on Fuzzy Default Logic” 0-7803-7810-5/03/$17.00 ©2003 IEEE, The IEEE International Conference on Fuzzy Systems.

[14] Ashish Bagwari, Raman Jee, “The Criteria Require for Cluster Head Gateway Selection in Integrated Mobile Ad hoc Network” Ashish Bagwari et al. / International Journal of Engineering Science and Technology (IJEST), ISSN : 0975-5462, Vol. 3 No. 7 July 2011.

[15] Caleb C. Noble, Diane J. Cook “Graph-Based Anomaly Detection” ISBN:1-58113-737-0 Proceeding KDD '03 Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM New York, NY, USA ©200.

[16] T. Shivaprakash, C. Aravind, A.P. Deepak, S. Kamal, H.L. Mahantesh, K.R. Venugopal, and L.M. Patnaik, “Efficient Passive Clustering and Gateway Selection in MANETs”, A. Pal et al. (Eds.): IWDC 2005, LNCS 3741, pp. 548–553, 2005. Springer-Verlag Berlin Heidelberg 2005

[17] Thomas Guyet, René Quiniou, Wei Wang, Marie-Odile Cordier, “Self-adaptive web intrusion detection system” INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE, ISSN 0249-6399 ISRN INRIA/RR--6989--FR+ENG, June 2009.

[18] NGAI Cheuk Han, “Trust- and Clustering-based Authentication Service in MANET”, at The Chinese University of Hong Kong in June 2004.

[19] James Cannady and Jay Harrell, “A Comparative Analysis of Current Intrusion Detection Technologies” Proceedings of Technology in Information Security Conference (TISC) '96, 212-218. 12.

[20] Jens Tölle, Oliver Niggemann, “Supporting Intrusion Detection by Graph Clustering and Graph Drawing” Publisher: Springer, Pages: 197-210, ISBN: 9783540415541

[21] Harley Kozushko “Intrusion Detection: Host-Based and Network-Based Intrusion Detection Systems” Thursday, September 11, 2003Independent Study

[22] S. Staniford-Chen, S. Cheung, R. Crawford, M. Dilger, J. Frank, J. Hoagland, K. Levitt, C. Wee, R. Yip, D. Zerkle “GrIDS -A graph based intrusion detection system for large networks” The work reported here is supported by DARPA under contract DOD/DABT 63-93-C-0045.

[23] Scott Fazackerley, Alan Paeth, Ramon Lawrence, “Cluster Head Selection Using Rf Signal Strength” The authors are funded by NSERC CGS, a NSERC Discovery grant, and the Irving K. Barber Endowment Fund.

[24] Brian Tung, “A Graph Theory Approach to Combining Intrusion Diagnoses” November 12, 2004, ISI Technical Report ISI-TR-2004-587. This work funded by the National Science Foundation (NSF) under Award 0209046.

[25] Dang Nguyen, Pascale Minet, Thomas Kunz and Louise Lamont, “On the Selection of Cluster Heads in MANETs” IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011 ISSN (Online): 1694-0814 www.IJCSI.org.

[26] Saira Beg1, Umair Naru1, Mahmood Ashraf, Sajjad Mohsin1 “Feasibility of Intrusion Detection System with High Performance Computing: A Survey” ISSN - 2218-6638, International Journal for Advances in Computer Science, Volume 1, Issue 1, December 2010.

◈◈◈

Categorization of SMS in Android Smart Phones

Mohammed Mateenuddin Siddiqui & K. V. Bhosle

GS Mandal's MIT College of Engineering Aurangabad ,India
E-mail : smateen81@gmail.com, kvbcse@gmail.com

Abstract - This paper involved concept of categorization of SMS in different format particularly the festival based categorization of SMS in android smartphones. The propose ideas of this paper is to make flexible categorization of SMS from the bulk of SMS in smartphones.The ideas of project is another additional feature can be implemented in android smartphones.This paper introduces the architecture and component models of android and analyzes the anatomy of an android application including the functions of Activity,Intent receiver,Serice,Content Provider and etc.The categorization of SMS can be implemented with the help of different Natural Lnaguage Processing Algorithms.

Keywords: *Android, SMS, Kernel, NLP*

I. INTRODUCTION

A mobile operating system or a mobile OS is an OS for handheld devices or mobiles. The operating system controls a mobile device—just like Mac OS, Linux or Windows.There are already many mobile platforms on the market today, includingSymbian, iPhone, Windows Mobile, BlackBerry, Java Mobile Edition,Linux Mobile (LiMo), and more. When I tell people about Android,their first question is often, Why do we need another mobile standard?

Where's the "wow"?

Although some of its features have appeared before, Android is the first environment that combines the following.'

1.⌋ A truly open, free development platform based on linux and open source

2∟ A component-based architecture inspired by Internet mashups.

3∟ Tons of built-in services out of the box.

4∟ Automatic management of the application life cycle.

5. ∟High-quality graphics and sound.

6. ∟Portability across a wide range of current and future Hardware.

Smartphones in today's age are found to be based on a number of different Operating Nokia's Symbian OS, Apple's IOS, RIM's BlackBerry OS, Microsoft's Windows Phone OS, Linux, Palm WebOS, Google's Android, Samsung's Bada (operating system) and Nokia's Maemo. Android, Bada, WebOS and Maemo are in turn built on top of Linux, and the iPhone OS is derived from the BSD and NeXTSTEP operating systems, which all are related to UNIX.

TABLE I. MARKET SHARE OF DIFFERENT SMARTPHONE

OPERATING SYSTEMS.

Source	**IDC (International Data Corp.)**
Date	*2011*
Android	33%
BlackBerry OS	29 %
iOS (Apple)	25 %
Microsoft	8 %
Palm	3%

Table 1, shows the market share of different Smartphone Operating Systems. As we see, Android's market has caught up very fast, as compared to other OS which have been in the market for more time than Android.

II. ANDROID

Android is a Linux-based operating system for mobile devices such as smartphones and tablet computers. It is developed by the Open Handset Alliance, led by Google, and other companies.

Google purchased the initial developer of the software, Android Inc., in 2005. The unveiling of the Android distribution in 2007 was announced with the founding of the Open Handset Alliance, a consortium of 86 hardware, software, and telecommunication companies devoted to advancing open standards for mobile devices.Google releases the Android code as open-source, under the Apache License. The Android

Open Source Project (AOSP) is tasked with the maintenance and further development of Android.

Android has a large community of developers writing applications ("apps") that extend the functionality of the devices. Developers write primarily in a customized version of Java Apps can be downloaded from third-party sites or through online stores such as Google Play (formerly *Android Market*), the app store run by Google. As of February 2012 there were more than 450,000 apps available for Android, and the estimated number of applications downloaded from the Android Market as of December 2011 exceeded 10 billion.

Fig 1. Android Logo

A. *Android Releases*

Google has updated Android from time-to-time, and has released the following versions:

- ∟ ┘1.5 (Cupcake) - released April 2009
- ∟ ┘∟1.6 (Donut) – released September 2009
- ∟ ┘∟2.0 / 2.1 (Eclair) - released October 20
- ∟2.2 (Froyo) –released May 20
- 2.3(Gingerbread) - released Dec. 2010
- ∟3.0(Honeycomb) -released Feb 2011
- 3.0(Honeycomb) -released Feb 2011
- ∟ ┘∟4.0(Ice Cream Sandwich) -released Oct 2011

Android 4.0 (Ice Cream Sandwich) is the latest version of the Android platform for phones, tablets, and more. It builds on the things people love most about Android easy multitasking, rich notifications, customizable home screens, resizable widgets, and deep interactivity and adds powerful new ways of communicating and sharing.

With the release of the latest version, Android 4.0 has come up with the following enhancements on the previous versions:

- Android 4.0 Ice Cream Sandwich is the brand new font system, which is much more appealing than previous versions. It is known as Roboto which is specially designed for high resolution screens and brings a magazine-like feel to the whole interface.
- Your Android smartphone will allow you to take screenshots without rooting the smartphone or installing any third-party apps. Just hold down the volume down key and the power button to capture the screen.
- While this is not as compelling as Siri on iPhone, unlike other versions of Android which support voice commands, Android 4.0 will not take very long to transcribe your words. Just speak it in and it instantly transforms it into text.
- This features is based on NFC (Near Field Communication) and it allows two Android smartphones to securely exchange Web pages, contacts, media or even applications. It is based on NDEF Push technology .
- Android ICS has some great new camera features, a brand new camera UI, options to edit your images using multiple effects right after you click them and it captures images at a ridiculously fast speeds, probably the fastest amongst smartphones. Video recording modes have also improved along with support for a time lapse mode.

B. *Android Availability*

Android costs nothing and the source code is freely available. Its license terms are commercial-friendly, basically, one can do whatever he feels like with it, without the intention of blaming Google, if anything goes wrong. The one exception to this rule is the Linux kernel, which is licensed under the GNU Public License. Because of this, manufacturers must release their device's Linux kernel source code after product shipment.

The Android Software Development Kit (SDK) consists of a debugger, libraries, a handset emulator, documentation, sample code, and tutorials.

C. *Linux Kernel*

As we have seen, Android is based on a Linux kernel. Android works on Linux kernel 2.6.x of the Linux kernel tree.

Google has come up with various versions of Android, each of which is based on different Linux versions such Android 1.5 (Cupcake) based on Linux

Kernel 2.6.27, 1.6 (Donut) based on Linux Kernel 2.6.29, 2.0 / 2.1 (Eclair) based on Linux Kernel 2.6.29, 2.2 (Froyo) based on Linux Kernel 2.6.32 and so on.

Android 4.0 (Ice Cream Sandwich) is based on on Linux 3.0 Kernel and later updated to 3.0.1.

III. ANDROID ARCHITECTURE

The Android architecture and it's main components are shown in following figure

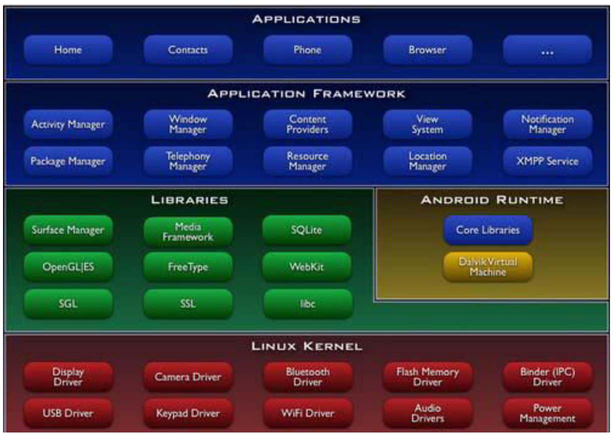


Fig 2. Android Architecture

A. *Application*

A set core application are top level of the framework include basic application,emailclient,SMS program calender,maps, browsers , phones and others .All application are written in java programming language.

B. *Application framework*

Developers have full access to the same framework APIs used by the core applications. The application architecture is designed to simplify the reusing of all components. This mechanism allows every component to be replaced by the user. Underlying all applications is a set of services and systems including a rich and extensible set of ActivitiesViews that can be used to build an application, including grids, lists, textViews editIntroductionTexts, Spinners, Buttons, an embeddable web browser and even an MapView which can be put into every app within very few lines of code; Content Providers that enable applications to access data from other applications (such as Contacts), or to share their own data; a automatic Resource Manager, making non-code resources accessible from code; a Notification Manager that enabling all applications to show custom alerts in the upper status bar.

An Activity Manager managing the life of each applications and providing a useful navigation backtrack.

C. *Libraries*

Android includes a set of C/C++ libraries used by various components of the Android system. These capabilities are exposed to developers through the Android application framework. Some of the core libraries are listed in Fig.1.

D. *Android Runtime*

Android includes a set of core libraries that provides most of the functionality available in the core libraries of the Java programming language. Every Android application runs in its own process given by the OS, and owns its own instance of the Dalvik virtual machine. Dalvik has been written so that a device can run multiple VMs efficiently. The Dalvik VM is executing files in the .dex (Dalvik Executable) format which was optimized for minimal cpu-and-memory-usage. The Virtual Machine is register-based, and runs classes compiledby a Java language compiler that have been transformed at compile-time into the .dex format using the "dx" tool, that are shipped with the SDK. The Linux Kernel can run multiple instances of the Dalvik VM, also providing underlying functionality such as threads and lowest-level memory management.

IV. ANATOMY OF AN ANDROID APPLICATION

Application components are the essential building blocks of an Android application. Each component is a different point through which the system can enter your application. Not all components are actual entry points for the user and some depend on each other, but each one exists as its own entity and plays a specific role—each one is a unique building block that helps define your application's overall behavior.

There are four different types of application components. Each type serves a distinct purpose and has a distinct lifecycle that defines how the component is created and destroyed.

Here are the four types of application components:

A. *Activity*

An Activity is an application component that provides a screen with which users can interact in order to do something, such as dial the phone, take a photo, send an email, or view a map. Each activity is given a window in which to draw its user interface. The window typically fills the screen, but may be smaller than the screen and float on top of other windows.

An application usually consists of multiple activities that are loosely bound to each other. Typically, one activity in an application is specified as the "main" activity, which is presented to the user when launching the application for the first time. Each activity can then start another activity in order to perform different actions. Each time a new activity starts, the previous activity is stopped, but the system preserves the activity in a stack (the "back stack"). When a new activity starts, it is pushed onto the back stack and takes user focus. The back stack abides to the basic "last in, first out" stack mechanism, so, when the user is done with the current activity and presses the **Back** button, it is popped from the stack (and destroyed) and the previous activity resumes.

When an activity is stopped because a new activity starts, it is notified of this change in state through the activity's lifecycle callback methods. There are several callback methods that an activity might receive, due to a change in its state—whether the system is creating it, stopping it, resuming it, or destroying it—and each callback provides you the opportunity to perform specific work that's appropriate to that state change. For instance, when stopped, your activity should release any large objects, such as network or database connections. When the activity resumes, you can reacquire the necessary resources and resume actions that were interrupted. These state transitions are all part of the activity lifecycle.

B. Service

A *service* is a component that runs in the background to perform long-running operations or to perform work for remote processes. A service does not provide a user interface. For example, a service might play music in the background while the user is in a different application, or it might fetch data over the network without blocking user interaction with an activity. Another component, such as an activity, can start the service and let it run or bind to it in order to interact with it. A Service is an application component representing either an application's desire to perform a longer-running operation while not interacting with the user or to supply functionality for other applications to use. Each service class must have a corresponding <service> declaration in its package's AndroidManifest.xml. Services can be started with Context.startService() and Context.bindService().

C. Content Providers

A content provider manages a shared set of application data. You can store the data in the file system, an SQLite database, on the web, or any other persistent storage location your application can access. Through the content provider, other applications can query or even modify the data (if the content provider allows it). For example, the Android system provides a content provider that manages the user's contact information. As such, any application with the proper permissions can query part of the content provider (such as ContactsContract.Data) to read and write information about a particular person.

Content providers manage access to a structured set of data. They encapsulate the data, and provide mechanisms for defining data security. Content providers are the standard interface that connects data in one process with code running in another process.

When you want to access data in a content provider, you use the ContentResolver object in your application's Context to communicate with the provider as a client. The ContentResolver object communicates with the provider object, an instance of a class that implements ContentProvider. The provider object receives data requests from clients, performs the requested action, and returns the results.

You don't need to develop your own provider if you don't intend to share your data with other applications. However, you do need your own provider to provide custom search suggestions in your own application.

You also need your own provider if you want to copy and paste complex data or files from your application to other applications.

Android itself includes content providers that manage data such as audio, video, images, and personal contact information. You can see some of them listed in the reference documentation for the android.provider package. With some restrictions, these providers are accessible to any Android application.

D. Broadcast Receiver

A broadcast receiver is a component that responds to system-wide broadcast announcements. Many broadcasts originate fromthe system for example, a broadcast announcing that the screen has turned off, the battery is low, or a picture was captured. Applications can also initiate broadcasts—for example, to let other applications know that some data has been downloaded to the device and is available for them to use. Although broadcast receivers don't display a user interface, they may create a status bar notification to alert the user when a broadcast event occurs. More commonly, though, a broadcast receiver is just a "gateway" to other components and is intended to do a very minimal amount of work. For instance, it might initiate a service to perform some work based on the event.

A broadcast receiver is implemented as a subclass of BroadcastReceiver and each broadcast is delivered as an Intent object. For more information, see the BroadcastReceiver class.

V. NATURAL LANGUAGE PROCESSING

A. Introduction

Natural Language Processing (NLP) is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies. And, being a very active area of research and development, there is not a single agreed-upon definition that would satisfy everyone, but there are some aspects, which would be part of any knowledgeable person's definition. The definition I offer is:

"Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications."

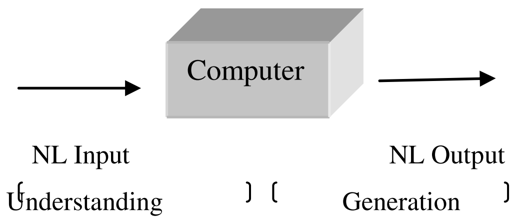


Fig 3. Natural Language Processing

B. Levels in Natural Language Processing

The most explanatory method for presenting what actually happens within a Natural Language Processing system is by means of the 'levels of language' approach. This is also referred to as the synchronic model of language and is distinguished from the earlier sequential model, which hypothesizes that the levels of human language processing follow one another in a strictly sequential manner? Psycholinguistic research suggests that language processing is much more dynamic, as the levels can interact in a variety of orders. Introspection reveals that we frequently use information we gain from what is typically thought of as a higher level of processing to assist in a lower level of analysis. For example, the pragmatic knowledge that the document you are reading is about biology will be used when a particular word that has several possible senses (or meanings) is encountered, and the word will be interpreted as having the biology sense.

Phonology.

This level deals with the interpretation of speech sounds within and across words. There are, in fact, three types of rules used in phonological analysis: 1) phonetic rules – for sounds within words; 2) phonemic rules – for variations of pronunciation when words are spoken together, and; 3) prosodic rules – for fluctuation in stress and intonation across a sentence. In an NLP system that accepts spoken input, the sound waves are analyzed and encoded into a digitized signal for interpretation by various rules or by comparison to the particular language model being utilized.

Morphology

This level deals with the componential nature of words, which are composed of morphemes – the smallest units of meaning. For example, the word preregistration can be morphologically analyzed into three separate morphemes: the prefix pre, the root registra, and the suffix tion. Since the meaning of each morpheme remains the same across words, humans can break down an unknown word into its constituent morphemes in order to understand its meaning. Similarly, an NLP system can recognize the meaning conveyed by each morpheme in order to gain and represent meaning. For example, adding the suffix –ed to a verb, conveys that the action of the verb took place in the past. This is a key piece of meaning, and in fact, is frequently only evidenced in a text by the use of the -ed morpheme.

Lexical

At this level, humans, as well as NLP systems, interpret the meaning of individual words. Several types of processing contribute to word-level understanding – the first of these being assignment of a single part-of-speech tag to each word. In this processing, words that can function as more than one part-of-speech are assigned the most probable part-of-speech tag based on the context in which they occur.

Additionally at the lexical level, those words that have only one possible sense or meaning can be replaced by a semantic representation of that meaning. The nature of the representation varies according to the semantic theory utilized in the NLP system. The following representation of the meaning of the word launch is in the form of logical predicates. As can be observed, a single lexical unit is decomposed into its more basic properties. Given that there is a set of semantic primitives used across all words, these simplified lexical representations make it possible to unify meaning across words and to produce complex interpretations, much the same as humans do.

launch (a large boat used for carrying people on rivers, lakes harbors, etc.) ((CLASS BOAT) (PROPERTIES (LARGE)(PURPOSE (PREDICATION (CLASS CARRY) (OBJECT PEOPLE))))

The lexical level may require a lexicon, and the particular approach taken by an NLP system will determine whether a lexicon will be utilized, as well as the nature and extent of information that is encoded in the lexicon. Lexicons may be quite simple, with only the words and their part(s)-of-speech, or may be increasingly complex and contain information on the semantic class of the word, what arguments it takes, and the semantic limitations on these arguments, definitions of the sense(s) in the semantic representation utilized in the particular system, and even the semantic field in which each sense of a polysemous word is used.

VI. CATEGORIZATION OF SMS IN EMULATOR.

Android provides a QEMU based emulator with the SDK, which can be used to test applications before they are loaded onto the phone.

This emulator can also be used for testing of the kernel and the compiled Android images.

This feature has been tested by us for Linux kernel 2.6.29 Goldfish (as Goldfish is the kernel which is required for emulator) and Android 2.3.

The Linux kernel was first obtained using git from http://android.git.kernel.org. The kernel was cross compiled using the inbuilt toolchain provided with Android source code- arm-eabi.

Having cross-compiled the Linux kernel, it was used with the obtained Android images on the emulator. The environment variable 'ANDROID_PRODUCT_OUT' stores the location of the Android images to be used. The kernel can be used by setting the '-kernel' option with the emulator.

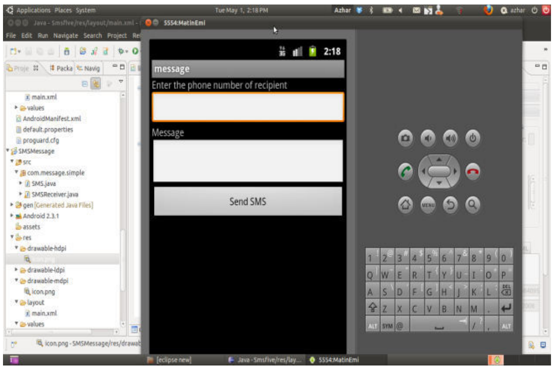


Fig 4:Android Emulator

VII. CONCLUSION.

The Android operating system for the mobile smart phones is making new stepping stone in every new aspects.Android has contributed in many of successful projects, the recentmost being the categorization of SMS in different fromats based on particularly Natural Language Programming approach.

Based on the minimum hardware requirement and less time consumining approach this catergorization is added another feature in android smart phones.

We have added the another feature in this project the categorization of SMS based on festivals so it will be easy for user to generalized this smart phone in their messaging prespective.

REFERENCES

[1] www.developer.android.com

[2] Android Source Code Repository. http://android.git. kernel.org/

[3] Nicolas Gramlich. Android Programming. (2nd edition). [Online]. Available: http:// andbook.anddev.org .

[4] A Spectrum White Paper: Thoughts on Google Android

[5] *android-developers.blogspot.com.*

[6] Winograd, T., (1971). Procedures as a Representation for Data in a Computer Programfor Understanding Natural Language. MIT-AI-TR- 235

[7] Woods, W. A. (1970). Transition Network Grammars for Natural Language Analysis.Communications of the ACM 13:10.

[8] Stock, O. (2000). Natural language processing and intelligent interfaces. Annals of Mathematics and Artificial Intelligence.

[9] Sparck Jones, K. (1999). What is the role for NLP in text retrieval. In T. Strzalkowski (Ed.). Natural language information retrieval. Kluwer, pp. 1—25.

[10] Roux, M.& Ledoray, V. (2000) Understanding of medico-technical reports. Artificial Intelligence in Medicine, 18, 149-72

◈◈◈

Real Time Implementation of Adaptive Image Enhancement Method on Android Platform

Jharna Majumdar[1], Shiva Sumanth Reddy[2] & Manoj Kumar M[3]

Nitte Meenakshi Institute of technology, Yelahanka, Bangalore – 560 064
Email: [1]jhama.majumdar@gmail.com, [2]saishivasumanth.reddy43@gmail.com, [3]dbamanoj@gmail.com

Abstract - Android is an open-source platform developed by Google and Open handset alliance. Image Processing on Android based mobile devices is an emerging field in today's World. Android also comes with a vast library of useful functions, including functions for user interfaces, image/bitmap manipulation, and camera controls. Histogram equalization is widely used for contrast enhancement in a variety of applications due to its simplicity and effectiveness. Examples include medical image processing and radar signal processing. One drawback of the histogram equalization can be found on the fact that the brightness of an image can be changed after the histogram equalization, which is mainly due to the flattening property of the histogram equalization. This paper presents the implementation of Brightness Preserving Histogram Equalization in Android Platform. The drawback of histogram equalization can be overcome by using the above mentioned method. This paper also shows the output on Android platform.

Keywords: *Histogram Equalization, Image Processing, Image Acquisition, Android, BBHE.*

I. INTRODUCTION

Histogram equalization is the one of the well-known methods for enhancing the contrast of given images in accordance with the sample distribution of an image. Useful applications of the histogram equalization scheme include medical image processing and radar image processing**.** In general, histogram equalization flats the density distribution of the resultant image and enhances the contrast of the image as a consequence, since histogram equalization has an effect of stretching dynamic range.

The BBHE firstly decomposes an input image into two sub-images based on the mean of the input image. One of the sub-images is the set of samples less than or equal to the mean whereas the other one is the set of samples greater than the mean. Then the BBHE equalizes the sub-images independently based on their respective histograms with the constraint that the samples.

Then the BBHE equalizes the sub-images independently based on their respective histograms. In other words, one of the sub-images is equalized over the range up to the mean and the other sub-image is equalized over the range from the mean based on the respective histograms. Thus, the resulting equalized sub-images are bounded by each other around the input mean, which has an effect of preserving mean brightness.

The architecture of the image processing in the Android mobile is shown in below.

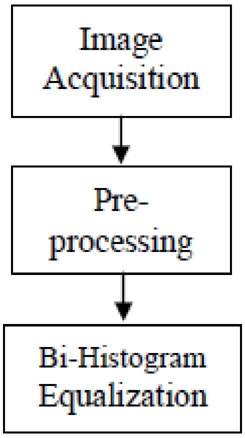


Figure 1: Architecture of the Image processing

Image Acquisition refers to the capturing of image data by a particular sensor or data repository. Once the image data is acquired, Pre-Processing often includes removing of noise from the acquired image data. Histogram Equalization is the process of enhancing the image.

II. ANDROID

A. Motivation

Image processing on mobile phones is a new and exciting field with many challenges due to limited hardware and connectivity. Phones with cameras, powerful CPUs, and memory storage devices are becoming increasingly common. The need for benchmarking basic image processing routines such as: addition, convolution, thresholding and edge detection is important for comparison of systems. With this information developers and researchers can design complex computer vision and image processing applications while being aware of the current state of the art limitations and bottlenecks on mobile phones.

B. Goals

The goal of this paper is to focus on Image Acquisition, Pre-Processing through implementing image noise removal algorithms; implementing Histogram Equalization and Histogram Equalization based Enhancement method on Android based mobiles such as HTC G1, Samsung Galaxy Y etc. using the available Android Software Development Kit (SDK).

C. Approach and Challenges

The Android operating system is preferable for benchmarking due to its recent growth and popularity. Few of the hardware manufacturers are e.g. HTC, Motorola, LG and Samsung. The Android operating system is supported and a part of the Open Handset Alliance. This alliance positions key manufacturers, cellular providers and the Android operating system in a collaborative environment which has caused large growth since October 2008 when the first Android mobile phone was released.

Few challenges when implementing Android mobile devices with Android OS includes architecting software and optimizing code for

a. Memory limitations

b. CPU limitations

c. Image Quality limitations

III. HISTOGRAM EQUALIZATION

ALGORITHM:

Input: An Image file, Row M, Column N

Output: Enhanced Image after histogram Equalization

Steps:

1. Let X = { X (i ,j) } denote a given image composed of L discrete gray levels denoted as $\{X_O, X_1, .., . . , X_{L-1}\}$, where X (i , j) represents an intensity of the image at the spatial location (i , j) and X (i , j) $\in$ $\{X_O, X_1, .., . . , X_{L-1}\}$. For a given image X, the probability density function p (X k) is defined as

$$p(x_k) = \frac{n^k}{n}$$

for **k** = 0 , 1 , . . . , L - 1, where n^k represents the number of times that the level X_k appears in the input image X and n is the total number of samples in the input image.

2. Based on the probability density function, we define the cumulative density function as

$$C(x) = \sum_{j=0}^{k} p(x_j)$$

Where, $\mathbf{X_k}$ = x, for k = 0, l , . . . , L - 1. Note that $C(X_{L-l})$ = 1 by definition. Histogram equalization is a scheme that maps the input image into the entire dynamic range, (X_o, X_{L-1}), by using the cumulative density function as a transform function.

3. Define a transform function f(x) based on the cumulative density function as

$$f(x) = X_0 + (X_{L-1} - X_0)c(x)$$

4. Then the output image of the histogram equalization, Y = {Y (i,j) } , can be expressed as

$$\mathrm{Y} = f(X)$$
$$= \{\ \{f(X(i,j))|\in X\}$$

IV. BBHE

ALGORITHM:

Input: An Image file, Row M, Column N

Output: Enhanced Image after histogram Equalization

Steps:

1. Denote by X_m the mean of the image X and assume that

$X_m \quad \{X_0, X_1, X_2, \ldots\ldots X_{L-1}\}$, ,Based on the mean, the input image is decomposed into two sub-images X_L and X_U as

$$X = X_L \cup X_U$$

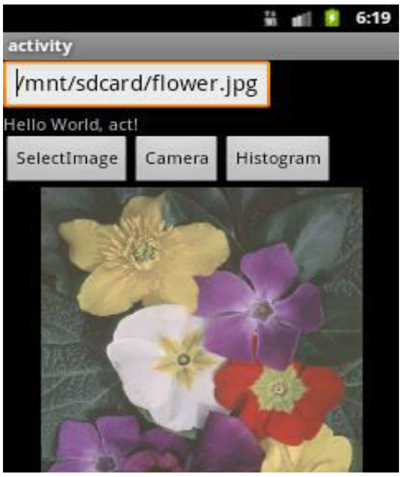
2. The probability density functions of the sub-image X_L and X_U as

$$p_L(X_k) = \frac{n_L^k}{n_L}, \quad \text{where } k = 0, 1, \cdots, m,$$

and

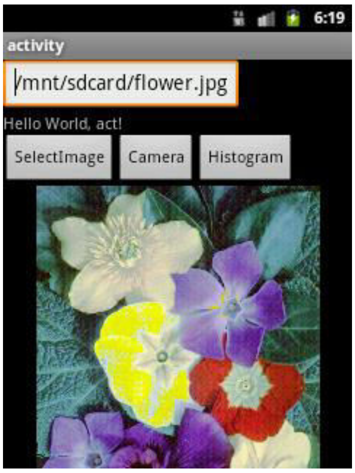
$$p_U(X_k) = \frac{n_U^k}{n_U}, \quad \text{where } k = m+1, m+2, \cdots, L-1,$$

V. RESULTS


/mnt/sdcard/flower.jpg

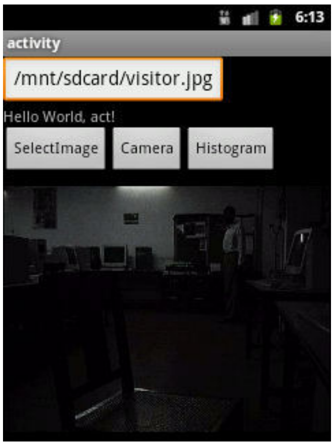
Hello World, act!

i. Input Image of Flowers


/mnt/sdcard/flower.jpg

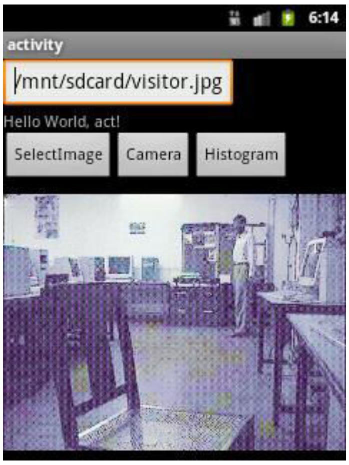
Hello World, act!

ii. Output Image of Flowers


/mnt/sdcard/visitor.jpg

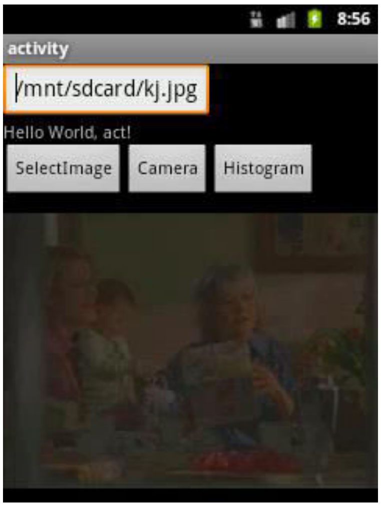
Hello World, act!

iii. Input Image of Furniture


/mnt/sdcard/visitor.jpg

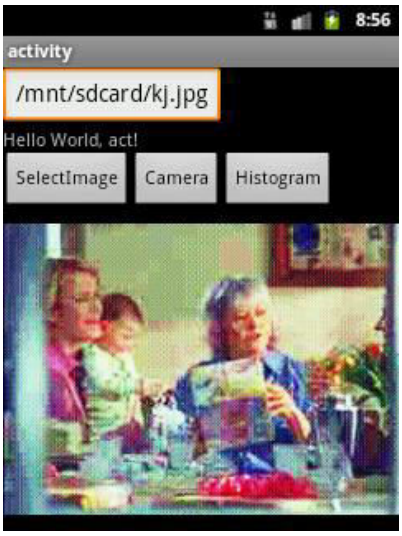
Hello World, act!

iv. Output Image of Furniture


/mnt/sdcard/kj.jpg

Hello World, act!

v. Input Image of persons


/mnt/sdcard/kj.jpg

Hello World, act!

vi. Output Image of persons

VI. CONCLUSION

In this paper we have used two different image enhancement methods that are popularly used in image understanding studies. Our results show that the images after enhancement have better visibility than the original images. This paper shows that the processing can be applied even for the real time image that is the image captured from mobile. This paper also avoids the flattening property of Histogram Equalization by using the Bi-Histogram Equalization.

REFERENCES

[1]. J.S Lim, Two-Dimensional Signal and Image processing, Prentice Hall, Englewood Cliffs, New Jersey 1990.

[2]. R.C Gonzalez and P. Wints, Digital Image Processing, 3rd Edition, Addison-Wesley Publishing Co., Reading, Massachusetts.

[3]. Y. Li, Wang, and D. Y. Yu, "Application of adaptive histogram equalization to x-ray chest image," Proc. Of the SPIE, pp. 513-514, Vol. 2321 1994.

[4]. Y.T. Kim, "Contrast Enhancement using Brightness Preserving Bi–Histogram Equalization". IEEE Transactions on Consumer Electronics, Volume 43, No.1, 1997.

[5]. Tutorial on Using Android for Image Processing, EE368, spring 2010, Linux Version

◈◈◈

Improved Near Duplicate Matching Scheme for E-mail Spam Detection

[1] **M. Siva Kumar Reddy** & [2] **B. Krishna Sagar**

Department of CSE, Madanapalli Institute of Technology and Science, Madanapalli, Andhra, Pradesh, India.

***Abstract* -** Today the major problem that the people are facing is spam mails or e-mail spam. In recent years there are so many schemes are developed to detect the spam emails. Here the primary idea of the similarity matching scheme for spam detection is to maintain a known spam database, formed by user's feedback, to block the subsequent near-duplicate spam's. We propose a novel e-mail abstraction scheme, which considers e-mail layout structure to represent e-mails. We present a procedure to generate the e-mail abstraction using HTML content in e-mail, and this newly devised abstraction can more effectively capture the near-duplicate phenomenon of spams. Moreover, we design a complete spam detection system Cosdes (standing for Collaborative Spam Detection System), which possesses an efficient near-duplicate matching scheme and a progressive update scheme. To detect fastly near duplicates and duplicate spam mails in Cosdes, we propose a new approach SimHash.

***Keywords* -** Spam mails, Emails, Near Duplicate SimHash, Spam Trees.

I. INTRODUCTION

Internet is the most widely used area. In internet most widely used are E-mails. E-mails play a major role for the communication between the people .The people who are using emails cannot verify the duplicate and near duplicate web documents creating the more problems on the web search engines. These documents will increase the space required to store the index, slow down the searching results and the annoy users. According to the data availability on the internet, the huge data are shorts texts such that mobile phone short messages, instant messages, chat log, BBS titles etc.

The statistical information is given by the Information Industry Ministry of china that more than 1.56 billion mobile phone short messages are sent each day in Mainland China. You already know how much of email is spam, but here are a bunch of other factoids as per [9] you may not be aware of:

- **90%** of spam is in English. A year ago it was 96%, so spam is getting more "international."
- **88%** of all spam is sent from botnets (networks of compromised PCs).
- **91%** of spam contains some form of link.
- Unsolicited newsletters are increasing and are now the second most common type of spam.
- Spam from webmail services like Gmail and Hotmail isn't as common as you might think. Only **0.7%** of spam is sent from webmail accounts.
- **1 in 284** emails contain malware.
- **1 in 445** emails are phishing emails.
- As many as **95 billion** phishing emails were in circulation in 2010.
- Unfortunately, the status of duplicate and near duplicate messages is very complex. Among these especially near duplicates and spam mails.

These differences may result from several causes: 1) same contents appearing on different sites are all crawled, processed and indexed; 2) mistake introduced while parsing these loosely structured and noisy text (HTML page may contain ads., and it is known as shorting of semantics useful for parsing); 3) manual typos (all information on Internet are created by people originally) and manual revising while being referred and reused; 4) explicit modification to make the short message suitable for difference usage.

Checking may be applicable manually when the scale of repository is small. E.g. hundreds or hundreds or thousands of instances. When the amount of instances increases to millions and more, obviously, it becomes impossible for human beings to check them one by one, which is tedious, costly and prone to error. Resorting to computers for such kind of repeatable job is desired, of which the core is an algorithm to measure the difference between any pair of short messages, including duplicated and near duplicated ones.

In Section 2, we define near duplicate and the construction of SP Tree and in section 3 we describe how SimHash works, in section 4 SimHash advantages and disadvantages A brief review of conventional work is presented in Section 4, followed by conclusion in Section 5.

2. Preliminaries

2.1 Near Duplicate

Near-duplicate spam detection is to exploit reported spams and to subsequently block one which have similar content. The definition of similarity between two e-mails are diverse for different forms of email. representing e-mails based mainly on content text, we represent e-mail using an HTML tag sequence, which depicts the layout Structure of e-mail, and look forward to more effectively capturing the near-duplicate phenomenon of spams.

Near-Duplicate:

Let I ={t1, t2, ...tn} be the set of valid HTML Tags with two types of newly created HTML tags <mytext/> and <anchor/>. An e-mail abstraction derived as <e1, e2, . . . , ei; . . . , em>, which is an ordered list of tags, where ei Є I.

The definition of near duplicate is: "Two e-mail abstractions A=<a1, a2, . . . ,ai, . . . , an> and B=<b1, b2, . . . , bi, . . . , bm> are viewed as

near-duplicate if for all ai= bi and n = m.

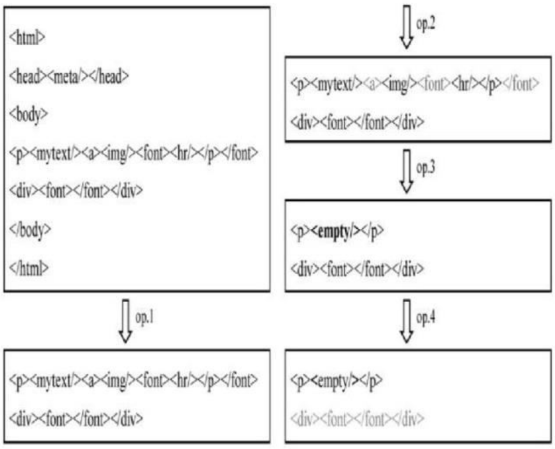
2.2 Related Works

Since the e-mail spam problem is increasingly serious various techniques have been explored to solve the problem. They can be categorized into the categories: 1) content-based methods,2) non content-based methods, and 3) others. Researchers analyze e-mail content text and model this problem as a binary text classification task. The solutions of this category are Naive Bayes, and Support Vector Machines (SVMs) methods. Naive Bayes methods train a probability model using classified e-mails, and each word in e-mails will be given a probability of being a suspicious spam keyword. As for SVMs, it is a supervised learning method, which possesses outstanding performance on text classification tasks. Markov random field model, neural network and logic regression, and certain specific features, such as URLs and images have also been taken into account for spam detection. The other group attempts to exploit noncontent information such as e-mail header, e-mail social network, and e-mail traffic to filter spams. Collecting notorious and innocent sender addresses (or IP addresses) from e-mail header to create blocked list and allowable mail list.

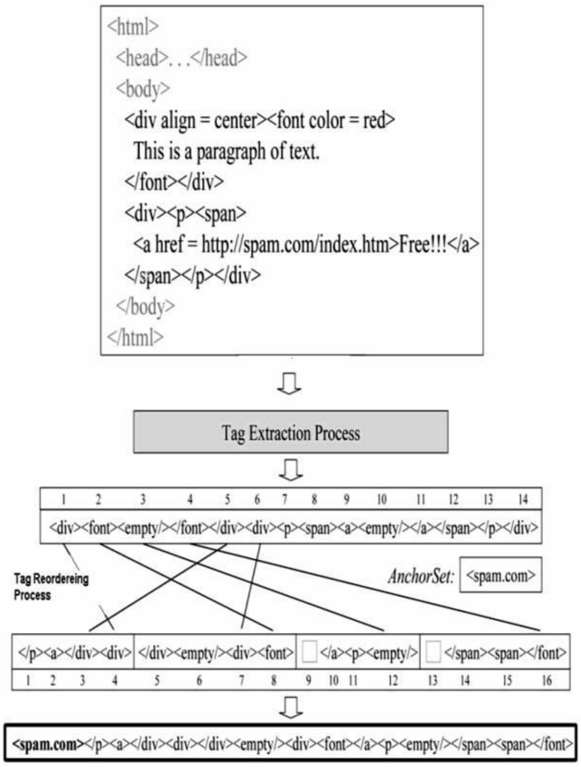
2.3 Structure Abstraction Generation

We propose the COSDES as a specific procedure SAG to generate the e-mail abstraction using HTML content in e-mail. Procedure SAG is composed of three major phases, Tag Extraction Phase, Tag Reordering Phase, and <anchor> Appending Phase. In Tag Extraction Phase, the name of each HTML tag is extracted, and tag attributes and attribute values are eliminated. In addition, each paragraph of text without any tag embedded is transformed to <mytext/>.

An example of the preprocessing step in Tag Extraction Phase of SAG.



Procedure flow of Structure Abstraction Generation



<html>
<head>. . .</head>
<body>
<div align = center><font color = red>
This is a paragraph of text.
</font></div>
<div><p><span>
<a href = http://spam.com/index.htm>Free!!!</a>
</span></p></div>
</body>
</html>
Tag Extraction Process
1 2 3 4 5 6 7 8 9 10 11 12 13 14
<div><font><empty/></font></div><div><p><span><a><empty/></a></span></p></div>
Tag Reordereing Process
AnchorSet: <spam.com>
</p><a></div><div> </div><empty/><div><font> </a><p><empty/> </span><span></font>
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
<spam.com></p><a></div><div></div><empty/><div><font></a><p><empty/></span><span></font>

2.4 Design of Spam Tree

SP tree is a data structure to facilitate the process of near-duplicate matching. SpTable and SpTrees (Sp stands for spam) are proposed to store large amounts of the e-mail abstractions of reported spams. Several SpTrees are the kernel of the database, and the e-mail abstractions of collected spams are maintained in the corresponding SpTrees. According to near duplicate definition, two e-mail abstractions are possible to be near-duplicate only when the numbers of their tags are identical.

For efficient matching Sp Trees are designed to be binary trees. The branch direction of each SpTree is determined by a binary hash function. If the first tag of a subsequence is a start tag (e.g.,<div>), this subsequence will be placed into the left child node. A subsequence whose first tag is an end tag (e.g.,</div>) will be placed into the right child node. Since most HTML tags are in pairs and the proposed e-mail abstraction is reordered in SAG, subsequences are expected to be uniformly distributed. Moreover, on level i of each SpTree (with the root on level 0), each node stores subsequences whose tag lengths are equal to 2i. For instance, as shown in Fig, the subsequence <spam:com> is placed into level 0, the subsequence </p><a> (whose tag length is 21) is placed into level 1, and so forth.

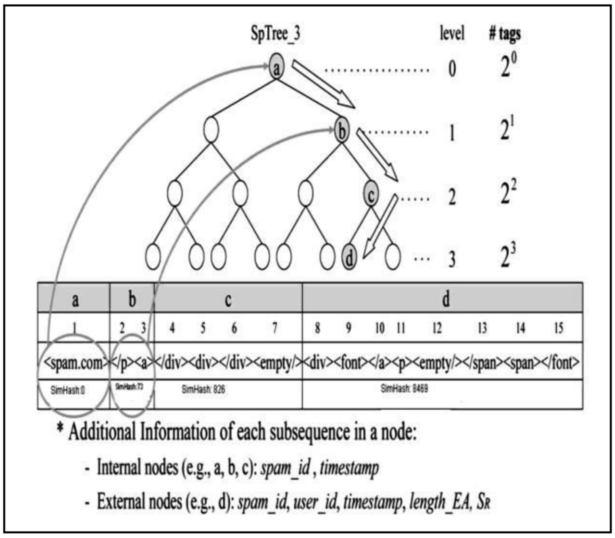

SpTree_3
level
# tags
a
b
c
d
0
1
2
3
2^0
2^1
2^2
2^3
a
b
c
d
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
<spam.com></p><a></div><div></div><empty/><div><font></a><p><empty/></span><span></font>
SimHash:0
SimHash:73
SimHash: 826
SimHash: 8469
* Additional Information of each subsequence in a node:
- Internal nodes (e.g., a, b, c): spam_id, timestamp
- External nodes (e.g., d): spam_id, user_id, timestamp, length_EA, Sr

Figure: Illustration of SP Tree with an example

3. Near-Duplicate Detection by SimHash

Charikar's SimHash [4], actually, is a fingerprinting technique that produces a compact representation of the objects may be documents or images. So, it allows for various processing, once applied to original data sets, to be done on the compact sketches, a much smaller and well formatted (fixed length) space. With documents, SimHash works as follows: a Web document is converted into a set of features, each feature tagged with its weight. Then, we transform such a high dimensional vector into an *f bit* - fingerprint where *f* is quite small compared with the original dimensionality.

The calculation of the hash is performed in the following way:

1. Document is splitted into tokens (words for example) or super-tokens (word tuples)
2. Each token is represented by its hash value; a traditional hash function is used
3. Weights are associated with tokens
4. A vector V of integers is initialized to 0, length of the vector corresponds to the desired hash size in bits
5. In a cycle for all token's hash values (h), vector V is updated: ith element is decreased by token's weight if the ith bit of the hash h is 0, otherwise ith element is increased by token's weight if the ith bit of the hash h is 1
6. Finally, signs of elements of V correspond to the bits of the final fingerprint

Sample program to show how SimHash works:

```
public class HtmlSimhash {
    private static final Logger LOG =
            Logger.getLogger(HtmlSimhash.class );
    public static void main(String[] args) {
Tap inputTap = new Hfs( new TextDelimited(new
Fields("docid","body"), " " ),args[0] );
        Tap outputTap = new StdoutTap();
// create the flow
Flow simhashFlow = Simhash.simhash(inputTap,
outputTap, 1, HtmlText.tokenizer(3));
        simhashFlow.complete(); // or add to your
                                    Cascade, etc
}
}
```

In this paper, we show that SimHash is indeed effective and efficient in detecting both duplicate (with *k* = 0) and near-duplicate (with *k* > 0) (see the two typical examples in TABLE II.) among large short message repository. However, we also notice that due to the born feature of short messages, *k* = 3 may not be an Ideal parameter for. For example, as shown in TABLE III. , *k* = 2 is enough to detect the one-character difference, but *k* has to be 5 to detect the same pair of messages with two-character difference. Besides, with the same one-character difference, short messages require larger *k* for effective detection. This may be explained by an observation, that the same difference, e.g. having one different character on the same position of two spam messages, would be more influential to short text than to long text.

This is a paper focusing on discussing Solution for real application. Firstly, we demonstrate a series of practical values of SimHash-based approach by experiments and our experience.

Secondly, we point out that *k* = 3 may be suitable for near-duplicated spam mail detection, but obviously not suitable for short messages.

Thirdly, we propose one empirical choice, *k* = 5 , as applied on our Online short message search. TABLE1. TYPICAL NEAR-DUPLICATES OF SPAM MAILS WITH DIFFERENCES HIGHLIGHTED IN GREY

1.International Monetary Fund congratulate you as our Ten(10) Star Prize Winner in our 2011 End of Year IAP held in London.This makes you a cash prize of £750,000.00 GBP 2. IMF congratulate you as our Ten(10) Star Prize Winner in our 2011 End of Year IAP held in London.This makes you a cash prize of £750,000.00 GBP

TABLE II. EXAMPLE: DETECT DUPLICATE WITH 0 *k* = AND NEAR-DUPLICATE WITH 0 *k* > (WITH DIFFERENCES HIGHLIGHTED IN GRAY)

K=0	1.Great Opportunity -- IT Professionals only IIPM LOOKING FOR INDIAN PROFILES 2.Great Opportunity -- IT Professionals only IIPM LOOKING FOR INDIAN PROFILES
K>0	1. Your e-mail has won you, (£750,000.00.Pounds) from COCA COLA NATIONAL LOTTERY On our 2011 charity bonanza 2. Your e-mail has won you, ($750,000.00.Dollors) from COCA COLA NATIONAL LOTTERY On our 2011 charity bonanza

TABLE III. EXAMPLE: DETECT SAME LONG TEXT BUT MORE DIFFRENCE REQUIRES LARGER *k* (WITH DIFFERENCES HIGHLIGHTED IN GRAY)

K=2	1.We are Pleased to inform you that you have won a prize money of GBP750,000.00 2.We are Pleased to inform you that you have won a prize money of INR750,000.00
K=5	1) Your e-mail address attached to Winning number 20-12jan-2010-02MSW, serial number S/N-00168, drew the lucky numbers 887-13-866-37-10-83 (2) Your e-mail address attached to Winningnumber20-12DEC-2010-02MSW, serial number S/N-00168, drew the lucky numbers 887-13-865-37-10-83

TABLE IV. EXAMPLE: DETECT SAME DIFFRENCE BUT SHORTER TEXT REQUIRES LARGER k (WITH DIFFERENCES HIGHLIGHTED IN GRAY)

K=2	1. Your e-mail has won you, (£750,000.00.Pounds) from COCA COLA NATIONAL LOTTERY On our 2011 charity bonanza 2. Your e-mail has won you, ($750,000.00.Dollors) from COCA COLA NATIONAL LOTTERY On our 2011 charity bonanza
K=5	1.Great Opportunity --- IT Professionals only IIPM seeing FOR INDIAN PROFILES ! 2.Great Opportunity -- IT Professionals only IIPM LOOKING FOR INDIAN PROFILES *

4. Advantages and Disadvantages of SimHash

SimHash has several advantages for application based on our experience:

1. Transforming into a standard fingerprint makes it applicable for different media content, no matter text, video or audio;
2. Fingerprinting provides compact representation, Which not only reduces the storage space greatly? but allows for quicker comparison and search.
3. Similar content has similar SimHash code, which permits easier distance function to be? determined for application.
4. It is applicable for both duplicate and near duplicate Detection, with *k* = 0 and *k* > 0 respectively.
5. Similar processing time for different setting of *k* if via the proposed divide-and-search mentioned above, and this is valuable for practice since we are able to detect more near duplicates with no extra cost.
6. The search procedure of similar encoded objects is easily to be implemented in distributed environment based on our implementation experience.

7. From the point of software engineering view, this procedure may be implemented into standard module and be re-used on similar applications, except that the applicants may determine the related parameters themselves.

5. Challenges to Detect Spam E-Mails

In this day and age, spammers are becoming more and more sophisticated. They are finding ways to trick people into thinking their unsolicited junk messages are worth the time you spend reading them. While many users are savvy enough to figure out what's real and what's bogus among their electronic correspondence, there are many out there who take what they receive at face value and open it.

This is alright though because sometimes the electronic junk mail swindlers are clever enough to pull the wool over our eyes. It's in the best interests of your computer's health and your sanity to research how to tell if an email is spam or genuine. We researched this topic extensively and generated a list of the top five ways to tell if an email is spam. These rules can help you when spam slips through the protection of your Spam filter.

Here are the some of the following list:

If it ends up in Spam Folder:

You might be reading this entry and thinking "Duh!" But you would be surprised how many people go rummaging through their spam folder like there's something they need in there. Unless you accidentally categorized legitimate emails as spam, you can be pretty sure that all the emails you need will appear in your inbox. Sometimes emails from certain websites end up in the spam folder. You must deal with those on a case-by-case basis to determine whether or not they're legitimate of pushing garbage into your inbox.

Look at the Email Address:

Legitimate companies send emails through a server based out of their company website (for example, support@microsoft.com). If you see a long string of numbers in front of the @ sign or the name of a free email service before the .com (or any other domain), you need to question the legitimacy of the email in question.

Look at the Content:

Keep an eye out for emails that say you need to do something right at that second or within a certain number of hours. Also, be wary of any emails that include links. Most companies tell you what to do, but they never direct you to where to do it with a link. Finally, rampant grammatical and spelling errors within the body of an email are good signs that it's spam. Spammers don't care enough about the actual messages they're sending to take the time to make them make sense.

If it asks for personnel Information:

Most institutions you deal with come right out and say they're never going to ask for personal information in an email. They don't need to ask you for your personal information anyway because they usually have it on hand. So, if you get an email that asks you for any personal information, no matter how legitimate it might seem, delete it right away. Personal information is only meant to be entered in secure, encrypted forms, not emails where anyone and everyone can get their hands on your information.

Look at the Greeting:

When you receive a genuine email, the sender addresses you directly, using either your first or last name. If you receive an email where they refer to you as a "Valued Customer" or as a member of some company, its spam. Senders of your genuine emails want to get your attention, so they always address you directly. We don't know about you, but when we read "Dear Valued Customer," our eyes begin to glaze over and our mouse cursor can't drag it to the trash fast enough.

VI. CONCLUSION AND FUTURE WORK

Uses an innovative tree structure, SpTrees, to store large amounts of the e-mail abstractions of reported spams. To achieve efficient matching with balanced tree structure, SpTrees are designed to be binary trees. The branch direction of each SpTree is determined by a binary hash function.

The improvement is limited since we map each subsequence in a node of an SpTree to a hash value. Therefore, the subsequences that have some prefix tags in common still can be differentiated with one comparison. In this paper, Instead of mapping each subsequence in a node of an SpTree to a hash value using a binary hash function we propose to replace it with a special hash function, namely Simhash.

The advantage of this over other hash functions is that it sets a minimum on the number of members that the two sets must share in order to match. This mitigates the effect of extremely common set members on data clusters.

SimHash based approach is Fast, Flexible, Customizable (HtmlSimhash), Scalable and is patented.

ACKNOWLEDGEMENT

We are greatly delighted to place my most profound appreciation to all my friends and faculty for encouragement and kindness in carrying out the paper.

Their pleasure nature, directions, concerns towards us and their readiness to share ideas rejuvenated our efforts towards our goal. We also thank the anonymous references of this paper for their valuable comments.

REFERENCES

[1] Chi-Yao Tseng, Pin-Chieh Sung, and Ming-Syan Chen "Cosdes: A Collaborative Spam Detection System with a Novel E-Mail Abstraction Scheme" IEEE transactions on knowledge and data engineering, vol. 23, no. 5, may 2011

[2] E. Blanzieri and A. Bryl, "Evaluation of the Highest Probability SVM Nearest Neighbor Classifier with Variable Relative Error Cost," Proc. Fourth Conf. Email and Anti-Spam (CEAS), 2007.

[3] M.-T. Chang, W.-T. Yih, and C. Meek, "Partitioned Logistic Regression for Spam Filtering," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data mining (KDD), pp. 97-105, 2008.

[4] S. Chhabra, W.S. Yerazunis, and C. Siefkes, "Spam Filtering Using a Markov Random Field Model with Variable Weighting Schemas," Proc. Fourth IEEE Int'l Conf. Data Mining (ICDM), pp. 347-350, 2004.

[5] P.-A. Chirita, J. Diederich, and W. Nejdl, "Mailrank: Using Ranking for Spam Detection," Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM), pp. 373-380, 2005.

[6] R. Clayton, "Email Traffic: A Quantitative Snapshot," Proc. of the Fourth Conf. Email and Anti-Spam (CEAS), 2007.

[7] A.C. Cosoi, "A False Positive Safe Neural Network; The Followers of the Anatrim Waves," Proc. MIT Spam Conf., 2008.

[8] E. Damiani, S.D.C. di Vimercati, S. Paraboschi, and P. Samarati, "An Open Digest-Based Technique for Spam Detection," Proc. Int'l Workshop Security in Parallel and Distributed Systems, pp. 559-564,2004.

[9] http://royal.pingdom.com/2011/01/19/ email-spam-statistics/

◈◈◈

Continuous Integration Software Build Enhancement System Based On Agile Model

Harish Reddy Bappa & Poornima M

Dept. of Information Science and Engineering, SJB Institute of Technology, Bangalore, India
E-mail :: harishbidar@gmail.com, poornima_urs2004@yahoo.co.in

Abstract - The term software build refers either to the process of converting source code files into standalone software artifact(s) that will be ready for deployment, or the result of doing so. Agile methodology is used in case of large projects where time critical factor and the situations where deliverables are submitted in stages. Agile methodology uses CI(continuous integration) tools in software build process, CI at any time picks up all latest code for package creation. For a build system where any error/non-working code cannot be afforded, mechanism is needed to pick a code which is the most functional state(locally tested by the developer before committing in the VCS) for the package(load/build) creation. Mechanism to view the old build details should be there. Customer-released loads will require immediate delta fixes which requires creation of branches for required components and privileged users should be allowed to start the load creation on safe decision. Our build enhancement system allows to enter the revision(version) numbers directly into the build environment making sure that only working code will be taken into the main build, Our system will provide UI(user interface) to view old build revision numbers based on the load name [Convention]or year and month of load creation. We have also provided auto-suggest feature for load selection. We have provided automated branch creation based on the load name and the component name without waiting for scm(software configuration management) engineer , so that the developer can start working on the patch faster, In our build enhancement system we have provided UI for triggering build.

Keywords*-continuous integration; agile model ; svn; software build.*

I. INTRODUCTION

Some people use the term “build” to mean compile, and this is not correct. On the teams that use no automated tools, the compile might be the only step in their build process, but in case of larger projects that has many subsystems (components) “build” is a process of taking the source of a software system and making it ready for deployment[9].

The Wikipedia defines it as follows,

The term software build refers either to the process of converting source code files into standalone software artifact(s) that can be run on a computer, or the result of doing so[7]. One of the most important steps of a software build is the compilation process where source code files are converted into executable code.

Agile model is best model for larger projects where entire project has to be divided into different subsystems(components). In agile module code will be reviewed, compiled and tested at regular intervals, Agile model can be used in case of urgent projects that are critical to the organization. which is also excels when requirements are unknown or changing[4]. Continuous Integration is integral part of agile model for improving software quality and reducing project risks. where members of a team integrate their work frequently, usually each subsystems code integrates at least daily - leading to multiple integrations per day[2]. Each integration is verified by an automated build (including test) to detect integration errors as quickly as possible. Many teams find that this approach leads to significantly reduced integration problems and allows a team to develop cohesive software more rapidly.

Here are some tasks that need to be done when creating a software build system,

- Getting Source Code from source control.
- Update the build numbers in code and documentation.
- Tag the code in source control.
- Build the code.
- Run automated unit tests.
- Build the documentation[7].

Manual effort is needed at different levels of the complete s/w build process. This is error prone and sometimes will lead to incorrect s/w builds. It also includes redundant tasks. In this paper we will discuss the different stages at which manual efforts are required and introducing the build enhancement system to remove manual involvement in the build process. by doing that we can,

- Improve product quality
- Accelerate the compile and link processing
- Eliminate redundant tasks
- Minimize "bad builds"
- Eliminate dependencies on key personnel
- Have history of builds and releases in order to investigate issues
- Save time and money

II. TRADITIONAL SOFTWARE BUILD PROCESS

Software projects involve many components that need to be orchestrated together to build a product. Keeping track of all of these is a major effort, particularly when there's multiple people involved. So it's not surprising that over the years software development teams have built tools to manage all this. These tools - called Source Code Management tools, configuration management, version control systems, repositories, or various other names - are an integral part of most development projects[1]. The sad and surprising thing is that they aren't part of all projects.

So as a simple basis make sure we get a decent source code management system. Cost isn't an issue as good quality open-source tools are available. The current open source repository of choice is Subversion. (The older open-source tool CVS is still widely used, and is much better than nothing, but Subversion is the modern choice.) the most commercial source code management tools are liked less than Subversion[3].

Once we get a source code management system, will make sure it is the well known place for everyone to go get source code. Nobody should ever ask "where is the foo-whiffle file?" Everything should be in the repository.

The repository is created for each subsystem(component) of the project, each of the component will have the high level directory structure which includes directories for, source code, libraries, build scripts, test.

Each of the repositories associated with the external properties and tagging, the developer can start working by checking out the repository into local development machine, developer will commit the changes done at the local working copy into the repository Code check-ins are supervised with SVN hooks.

The software build process make use of many tools for example

- Standard Make Tool used for Build
- Continuous integration Tool Ex: HUDSON
- Subversion [SVN] used for Software Configuration Management
- A group server used as the development environment

A special repository(SCM repository) will be created which will be having only the external properties which has component names and urls of associated repositories this repository will be managed by SCM Team . Tags & External Properties are configured in SCM Repository. Developers checkout the sandbox url to obtain the complete work environment.

Developers checkout all the subsystem repositories from the SCM repository/ Developers work on their respective subsystem and commits the change to SVN.

It is recommended for developers to do an Update frequently in order receive the commits from the other repositories.

The SCM engineer will update external properties of the scm repositories with the revision numbers provided by the developer, and trigger the build through CI tool through Hudson,

One of the features of version control systems is that they allow you to create multiple branches, to handle different streams of development. This is a useful, and essential, feature - but it's frequently overused and gets people into trouble.it is suggested to Keep use of branches to a minimum. In particular there will be a mainline(trunk): a single branch of the project currently under development. Pretty much everyone should work off this mainline most of the time (Reasonable branches are bug fixes of prior production releases and temporary experiments)[3].

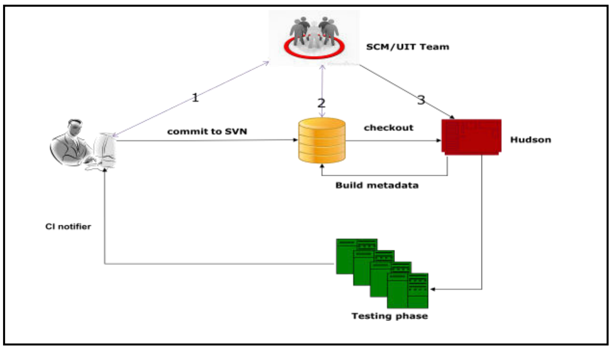


Fig.1: traditional S/W build process

All development and corrections will be happening on trunk/mainline.

Short lived branches will be created for critical/immediate corrections for SysVe.

A separate Hudson(CI Tool) project is created on demand if there are pronto correction required o released loads to SyVe. The Hudson project is created based on the load released.

As per individual component requirements, branches are created on demand. The corrections in the branches should be maintained in the trunk as well.

These branches are short-lived and are terminated with next SyVe release.

As per individual component requirements, the svn revision picked up for that particular Hudson project is managed. The patches for blocking issues will be delivered on demand. This is done through hyphen loads which contains the increments to the load released.

For the creation of branches the developer will request to the SCM engineer, by providing the load name and the particular component to which branch has to be created, through committed logs scm engineer will go through the load details(i,e pick up revision number of the given load) and create the branch based on the corresponding revision number of the component.

III. LOOP HOLES IN THE TRADITIONAL CI BUILD PROCESS

Entities involved in the build process are :

developer

svn

scm engineer

CI Tool

Testing phase

Each component may have a separate svn repository. A developer workspace will contain all the components checkout using a single svn url which has all these component urls configured in svn external properties. Developers work on their respective subsystem and commits the change to SVN.

A CI system at any time picks up all latest code for package creation, but for a fast paced build system where any error/nonworking code cannot be afforded, the SCM team has to ensure the same[10]. scm engineer has the responsibility of taking the revision numbers of the components from the corresponding developer for the next build and update them into the scm repository’s svn:external properties, developer will post revision numbers into wikipage, scm/uit team has to manually pick it up and updates in svn:externals, this involves the manual efforts which involves redundant tasks and error-prone which interns lead to incorrect s/w build. customer released loads will require immediate delta fixes this should be handled by creation of branches for required components,this is handled by the creation of branches for required components developer will request to the scm engineer for branch creation for corresponding component by giving the load name and component name for which branch has to be created, scm will refer to the load details(revision numbers) and replicate the component environment using svn branches, here the developer has to wait for scm engineer to create branch, it's a time consuming, it will delay the fixes[5]. If developer need to know the which revision no for particular component gone through the particular load, has to request to the scm for load details, or search through the not so easy svn commit logs which is time consuming.

Developer is dependent on scm for triggering a build patch build or full s/w build.

IV. SOFTWARE BUILD ENHANCEMENT SYSTEM :

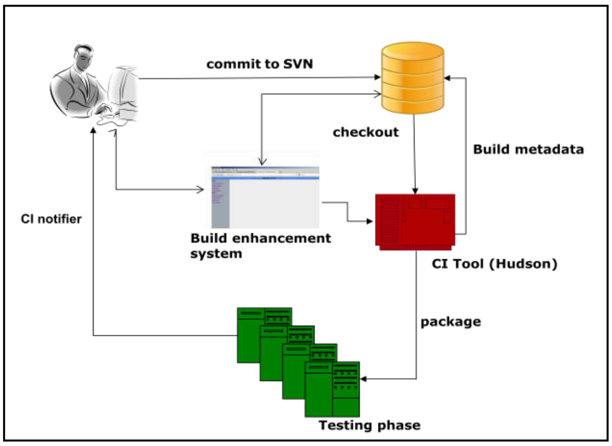


Fig : S/W build process using build enhancement system

Our build enhancement system removes most of manual efforts involved in software build process, the build enhancement system provides web based user interface, built on RHEL platform, through our system developers will be permitted to commit the revision directly into the build environment sandbox[svn external properties] where the software build happens, .This eliminates the error prone rvn no posting and reading process.

Developer itself can create branch by providing load name and the component name, without waiting for scm person. Our system will go through the load details and replicate the component environment using svn branches.

User interface is created to view the old build details, developer can search build details based on the load name, which will also provide the autosuggest feature. Also can search build details by year and month name of the build creation, This eliminates the involvement of an SCM person or other difficult svn commit log references

The privileged user can trigger build directly through build enhancement system.

V. CONCLUSION AND FUTURE WORK

To overcome the manual efforts involved in the continuous integration software build creation we have proposed and implemented software build enhancement system which will make the Agile work flow faster and error less through a single access /operation point for all build related steps.

We see scope for Improvement in Notification area where we can include mail through SMTP.

Can incorporate other tool links. E.g .For NLOC generation, staic/runtime code analysis tools etc.

This can be extended to the E2E perspective of a CI system which includes different functions and stages of the Unit Integration testing

ACKNOWLEDGMENT

This article is derived and project is done as part from Software build process deployed in Femtogateway project in Nokia Siemens Networks(NSN). We thank to all the team members of the Femtogateway project(especially to Cibicherian SCM Engineer at Femtogateway project NSN) for their strong support suggestions.

REFERENCES

[1] Paul M. Duvall with Steve Matyas, Andrew Glover "Continuous Integration" Improving software quality and reducing risk

[2] Kevin Roebuck "Continuous Integration: High-Impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors"

[3] C. Michael Pilato, Ben Collins-Sussman, Brian W. Fitzpatrick "Version Control with Subversion" 2nd edition.

[4] Scott W. Ambler "Agile modeling: effective practices for eXtreme programming and the unified process"

[5] Thomas Stober, Uwe Hansmann "Agile Software Development: Best Practices for Large Software Development Projects"

[6] Scott W. Ambler "The Object Primer:Agile Model-Driven Development with UML 2.0" 3rd ed.

[7] Patrick Cauldwell "Code Leader: Using People, Tools, And Processes To Build Successful Software"

[8] http://en.wikipedia.org/wiki/Software_build

[9] www.google.com

[10] Peter Smith, "Software Build Systems: Principles and Experienc"

[11] Ed Blankenship, Martin Woodward, Grant Holliday, Brian Keller , "Professional Team Foundation Server 2010"

◈◈◈

Statewide Higher Education Counseling System Using Grid Computing

D. Ramesh [1], & A. Krishnan [2]

[1]Dept of CSE, Anna University of Technology, Tiruchirappalli
[2] K.S.Rangasamy college of Technology, Tiruchengode
E-mail : drameshphd@gmail.com & a_krishnan26@hotmail.com

Abstract - Grid computing is a virtualized distributed computing environment which aims at enabling the dynamic “runtime” selection, sharing, and aggregation of distributed autonomous resources based on the availability, capability, performance. The performance of grid is improved in many aspect based on various research direction of last few year. In grid computing, load sharing is the major research issue. This paper further explains various existing techniques, architectures, applications of grid computing. And this paper proposed an application, in which state wise counseling for higher studies is applied. The proposed system has many grid located in many places which spreads geographically in different places. Further these grids are interconnected using inter-grid architecture. The conclusion of the proposed work is proved that the optimal load sharing and efficient transmission of data is obtained in the proposed work.

Keywords: *Grid Computing, Load Sharing, Grid Architecture, Inter Grid Connectivity.*

I. INTRODUCTION

In the mid-1990s the grid metaphor was applied to computing, by extending and advancing the 1960s concept of “computer time sharing.” The grid metaphor strongly illustrates the relation to, and the dependency on, a highly interconnected networking infrastructure.

According to IBM’s definition [1], grid 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 vision is to create virtual dynamic organizations [2] through secure, coordinated resource-sharing among individuals, institutions, and resources. Grid computing is an approach to distributed computing that spans not only locations but also organizations, machine architectures, and software boundaries to provide unlimited power, collaboration, and information access to everyone connected to a grid.

1.1 KEY ISSUES IN GRID COMPUTING

The key issues in the grid computing are:

- benefits of Grid and status of technology;
- Motivations for considering computational grids;
- Brief history of grid computing;
- Is grid computing ready for prime time?;
- Early suppliers and vendors Challenges;
- Future directions;
- What are the components of grid computing systems/architectures?
- Portal/user interfaces
- User security
- Broker function
- Scheduler function
- Data management function
- Job managementand resource management
- Are there stable standards supporting grid computing?
- Virtual organization creation and management
- Service groups and discovery services
- Choreography, orchestration, and workflow
- Transactions
- Metering service
- Accounting service
- Billing and payment service
- Grid system deployment issues and approaches
- Generic implementations

- Security considerations—Can grid computing be trusted?
- What are the grid deployment/management issues?
- Challenges and approaches
- Availability of products by categories
- Business grid types
- Deploying a basic computing grid
- Deploying more complex computing grid
- Grid operation
- What are the economics of grid systems?
- The chargeable grid service
- The grid payment system
- Communication and networking infrastructure
- Communication systems for local grids
- Communication systems for national grids
- Communication systems for global grids

1.2 VIRUATILIZATION IN GRID SYSTEM

From the key issues, the Virtualization [3] [4] is a most important requirement of grid systems, and the virtualization can span the following domains:

1. Server virtualization for horizontally and vertically scaled server environments. Server virtualization enables optimized utilization, improved service levels, and reduced management overhead.
2. Network virtualization, enabled by intelligent routers, switches, and other networking elements supporting virtual LANs. Virtualized networks are more secure and more able to support unforeseen spikes in customer and user demand.
3. Storage virtualization (server, network, and array-based). Storage virtualization technologies improve the utilization of current storage subsystems, reduce administrative costs, and protect vital data in a secure and automated fashion.
4. Application virtualization enables programs and services to be executed on multiple systems simultaneously. This computing approach is related to horizontal scaling, clusters, and grid computing, in which a single application is able to cooperatively execute on a number of servers concurrently.
5. Data center virtualization, whereby groups of servers, storage, and network resources can be provisioned or reallocated on the fly to meet the needs of a new IT service or to handle dynamically changing workloads.

1. ARCHITURAL DESIGN AND RECENT DEVELOPMENT IN GRID COMPUTING

The evolution of grid system and the architecture of grid system are shown in the figure 1 and figure 2.

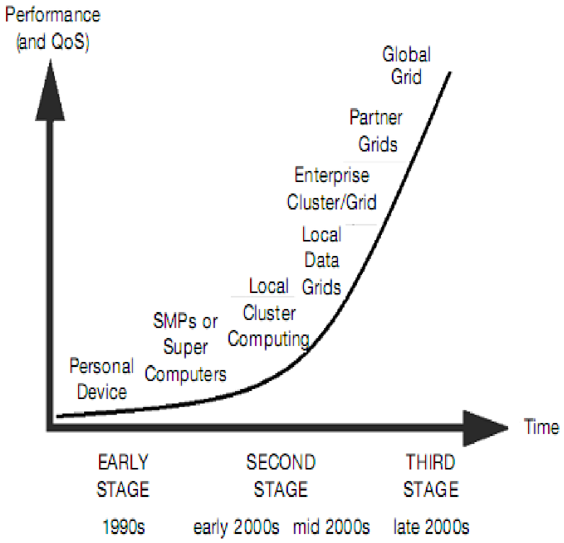


Figure 1: Evolution of Grid system

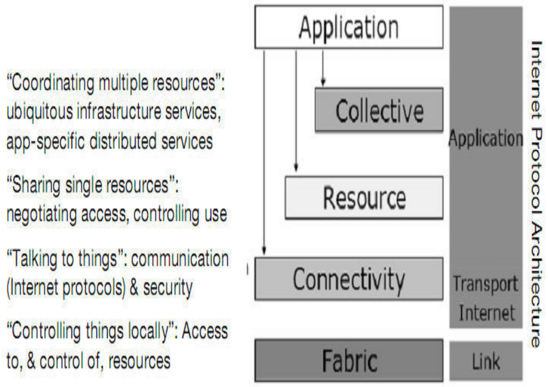

Application
"Coordinating multiple resources": ubiquitous infrastructure services, app-specific distributed services
Collective
"Sharing single resources": negotiating access, controlling use
Resource
"Talking to things": communication (Internet protocols) & security
Connectivity
"Controlling things locally": Access to, & control of, resources
Fabric
Application
Transport
Internet
Link
Internet Protocol Architecture

Figure 2: Layered Architecture of Grid System

The following are the Grid consortium [5] [6] which focuses on grid computing and its engineering applications:

- Asia Pacific Grid
- Australian Grid Forum
- Content Alliance: About Content Peering
- Distributed.net
- eGrid: European Grid Computing Initiative
- EuroTools SIG on Metacomputing

- Global Grid Forum [2]
- Global Grid Forum (GGF)
- Grid Computing Info Centre
- GridForum Korea
- IEEE Task Force on Cluster Computing
- New Productivity Initiative (NPI)
- Peer-to-Peer (P2P) Working Group
- SETI@home
- The Distributed Coalition

And the following are the few Grid Applications in the recent world wide implementation:

- Access Grid
- APEC Cooperation for Earthquake Simulation
- Australian Computational Earth Systems Simulator
- Australian Virtual Observatory
- Cellular Microphysiology
- DataGRID—WP9: Earth Observation Science Application [7]
- Distributed Proofreaders
- DREAM Project: Evolutionary Computing and Agents Applications
- EarthSystemGrid
- Fusion Collaboratory
- Geodise: Aerospace Design Optimisation
- Globus Applications
- GRid seArch & Categorization Engine (GRACE)
- HEPGrid: High Energy Physics and the Grid Network
- Italian Grid (GRID.IT) Applications
- Japanese BioGrid
- NEESgrid: Earthquake Engineering Virtual Collaboratory [8]
- Knowledge Grid [9]
- Molecular Modelling for Drug Design
- NC BioGrid
- Neuro Science—Brain Activity Analysis
- NLANR Distributed Applications
- OpenMolGrid
- Particle Physics Data Grid
- The International Grid (iGrid)
- UK Grid Apps Working Group
- US Virtual Observatory
- Bayanihan Computing Group
- Cetacean acoustic communication study

The figure 3 is shown the architectural designs for connecting the local gird computing systems [10] [11]. The figure 4 is shown the architectural designs for connecting the intra grid computing systems [12] [13]. In which the local grid is located in one place, it means that the entire grid is located geographically in one place. Whereas the intra grid and inter grid architecture are proposed to connect two various grid systems that located two or more different places.

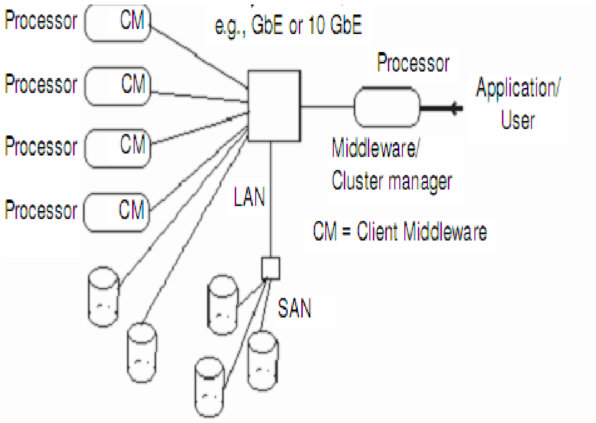


Figure 3. Local Grid Architecture

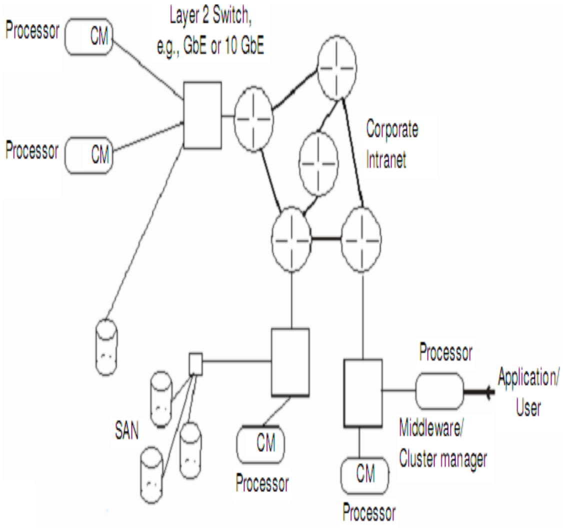


Figure 4. Intra Grid Architecture

II. PROPOSED WORK

In this paper, we proposed a state wise higher education counselling system for engineering admission. Now a days, admission to professional courses become centralized which mostly under the control of affiliated university. In india, tamilnadu become one among best place for professional courses. Merely 500 engineering colleges running under single affiliated university. For admission into all courses and all colleges, counselling based on mark is arranged for students to get their choice. In order to arrange such counselling in the geographically large area, centralized and multiple nodal is required. Therefore, we proposed grid computing for professional course counselling.

The figure 5 and 6 shows that the Grid which available on various places and its interconnection.

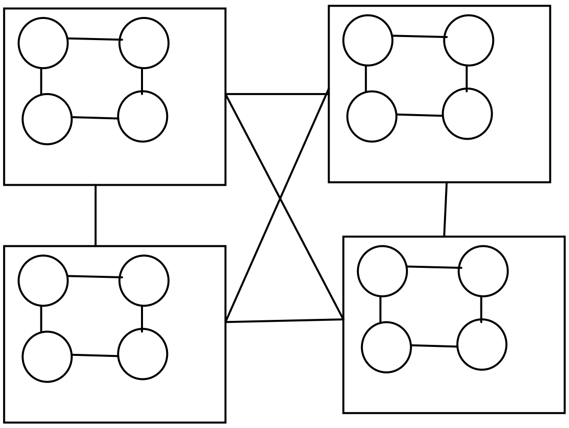


Figure 5. Inter Linked Mesh Grid Architecture

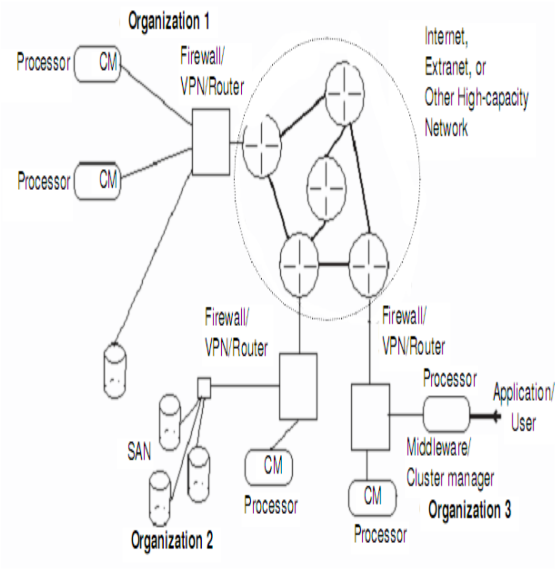


Figure 6. Inter Grid Architecture

In the proposed system, Grid are interconnected and for optimized data transmission, a centralized contol system is proposed which is termed as Grid Control System (GCS). In order to meet more number of request from the users, the proposed system using the Grid Control System (GCS).

The functionalities of the GCS are

1) will receive the query from the user
2) assign effecively the job scheduing,
3) splitting the job as tasks whenever required
4) and scheduling the tasks in proper manner using optimised scheduling algorithm.

Splitter will split jobs into one or more tasks, based on the following equation (1)

$$no\ of\ task = \begin{Bmatrix} 1\ if\ no\ of\ job < no\ of\ grid \\ \frac{no\ of\ job}{no\ of\ grid}\ if\ no\ of\ job < no\ of\ grid \end{Bmatrix} \quad \text{------ ---} \quad (1)$$

The proposed grid is tested on various test case and performance of grid based on execution time, maximum execution time, and idle time are calculated. These recorded information and are shown in the table 1 to table 5.

Table 1. Various Test Case used for testing the grid performance

Test Case	b	H	m1	m2	p1	p2
Case 1	0	1	2	1	0.67	0.33
Case 2	1	0	2	1	0.50	0.50
Case 3	0.5	0.5	2	1	0.56	0.44
Case 4	0	8	2	1	1.00	0.00

Table 2. Performance of Grid on Test Case 1

			POACO- Test Case 1	
Time	T1	T2	G1	G2
0	180	30	140	70
100	90	25	77	38
200	98	10	72	36
300	111	6	78	39
Execution Time			367	183
Max Execution Time			367	
Idle Time			51	156

Table 3. Performance of Grid on Test Case 2

Time	T1	T2	POACO-Test Case2	
			G1	G2
0	180	30	105	105
100	90	25	58	58
200	98	10	54	54
300	111	6	59	59
Execution Time			275	275
Max Execution Time			275	
Idle Time			89	89

Table 4.Performance of Grid on Test Case 3

Table 5. Performance of Grid on Test Case 4

Time	T1	T2	POACO-Test Case3	
			G1	G2
0	180	30	118	92
100	90	25	65	50
200	98	10	61	47
300	111	6	66	51
Execution Time			310	240
Max Execution Time			310	
Idle Time			74	111

Time	T1	T2	POACO-Test Case4	
			G1	G2
0	180	30	209	1
100	90	25	115	0
200	98	10	108	0
300	111	6	117	0
Execution Time			548	2
Max Execution Time			548	
Idle Time			0	298

CONCLUSION:

The proposed work is carried out for professional admission for higher secondary students. In which many number of request from the user of the grid is to be performed. In order to perform such huge request, GCS is proposed and the functionalities of GCS is shown briefly in the previous chapter. For task scheduling, Parameter Optimized Ant Colony Optimization (POACO) is proposed. Hence, the proposed system is tested on various test cases which is shown in the table 1. The performance of test case 1 to test case 4 are tabluated in the tables 2 to 5. From these tables, it is identified that the performance of grid is optimized in the test case 2. Therefore, the parameter of ACO is adjusted based on test case 2.

REFERENCES:

[1]. IBM Press Releases. IBM Corporation, 1133 Westchester Avenue, White Plains, New York 10604, www.ibm.com.

[2]. The Globus Project™, Introduction to Grid Computing, Argonne National Laboratory USC Information Sciences Institute, 2002, http://www.globus.org

[3]. Foster, D. Gannon, H. Kishimoto, and J. Von Reich, Open Grid Services Architecture Use Cases, Global Grid Forum OGSA-WG, Draft draft-ggf-ogsa-usecase-2, 2003.

[4]. Foster, C. Kesselman, and S. Tuecke, “The Anatomy of the Grid: Enabling Scalable Virtual Organizations,” International Journal of Supercomputer Applications, 15 (3), 200–222. 2001

[5]. M. C. Brown, “Grid Computing—Moving to a Standardized Platform,” August 2003, IBM archives, IBM Corporation, 1133 Westchester Avenue, White Plains, New York 10604, www.ibm.com.

[6]. T. Myer, “Grid Computing: Conceptual Flyover for Developers,” IBM Corporation, 1133 Westchester Avenue, White Plains, New York 10604, May 2003.

[7]. Chervenak, I. Foster, C. Kesselman, C. Salisbury, and S. Tuecke, “The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Data Sets,” Journal of Network and Computer Applications: Special Issue on Network-Based Storage Services, 23, 3, 187–200, 2000.

[8]. Foster, C. Kesselman, and S. Tuecke, “The Anatomy of the Grid: Enabling Scalable Virtual Organizations,” International Journal of Supercomputer Applications, 15, 3, 2001.

[9]. R. Buyya et al., “Economic Models for Management of Resources in Peer-to-Peer and Grid Computing,” in Proceedings SPIE International Conference on Commercial Applications for High-Performance Computing, SPIE, Bellingham, WA, 2001.

[10]. Bahrami, M.; Faraahi, A.; Rahmani, A.M., “AGC4ISR, New Software Architecture for Autonomic Grid Computing”, International Conference on Intelligent Systems, Modelling and Simulation (ISMS), 318 – 321, 2010

[11]. Dabhi, V.K.; Prajapati, H.B., “Soft computing based intelligent grid architecture”, International Conference on Computer and Communication Engineering, 574 – 577, 2008.

[12]. Jose, M.V.; Seenivasagam, V. “Object Based Grid Architecture for enhancing security in grid computing “, International Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), 414 – 417, 2011

[13]. Tsung-Li Wang; Chung-Ho Su; Pei-Yun Tsai; Tyng-Yeu Liang; Wen-Hsiung Wu “Development of a GridERP Architecture: Integration of Grid Computing and Enterprise Resources Planning Application”, International Conference on Wireless Communications, Networking and Mobile Computing, 1-4, 2008.

◈◈◈

Action Recognition Using Feature Transform Descriptor from Mined Dense Spatio Temporal

N. R .Vikram & P. M. Ashokkumar

Department of information technology, M.I.T campus, Anna University, Chennai
E-mail : [1]nrvikram89@gmail.com, [2]pmashokk@gmail.com

Abstract - Action recognition in video sequence has been a major challenging research area for number of years. Apriori algorithm and SIFT descriptor based approach for action recognition is proposed in this paper. Here, two phases can be carried out for accurate and updating of action recognition. In the first phase, the input should be the video sequence. For preprocessing the sequences frame can be formatted by background modeling for every successive frame. After modeling the background, the corners are detected for every frame and compound features are extracted. Data mining is performed by using Apriori algorithm as well as with the help of compound features extraction the action can be segregated in this video sequence. In the second phase, the same process is performed as well as analysis for new input frame and new pattern are updated for perfect recognition process. Due to Apriori algorithm, the processing delay is reduced and accuracy is improved.

Keywords: *Feature extraction, action recognition, Apriori, SIFT.*

I. INTRODUCTION

In most of computer vision application identifying moving objects in video sequence is a critical and fundamental task. Background subtraction is the common approach which identifies the moving object in a particular video frame that differs from background model. Good background subtraction algorithm should satisfy these challenges, it must be robust against changes in illumination, avoid detecting non-stationary background objects and internal background model should react immediately for the changes in background. Corner features, sometimes referred to as interest points, are image features characterized by their high intensity changes in the horizontal and vertical directions.

The four corner points are said to be good if the square object is present in the image. Shape and motion analysis can be performed by using these corners. Motion is ambiguous in nature at edges, since here corners are 2D image features ambiguity is not caused in motion analysis.

Within the object recognition, for feature selection learning has proven successful at building classifiers from large sets of possible features e.g. Boosting. Although various approaches similar to our approach have been applied to spatio-temporal activity domain, but those approaches is not good effective with the number of the features and also have issues with time alignment and scaling.

Our approach is based on extracting very low level features from the video sequence and these low level features are combined to form high-level, compound and spatio-temporal features. Data mining is used to accumulate the compound features using the data mining technique, Association rule, which efficiently discovers frequently reoccurring combinations. To group SIFT descriptors for object recognition, association data mining was recently employed. We use it to build high level compound features from a noisy and over-complete set of low-level spatio and spatio-temporal features by which the action recognition can be done. We compare encoding only relative spatial offsets, which provides scale invariance, to the spatial grid proposed by Quack *et al.* and demonstrate that, due to increased scale invariance, higher performance is achieved. Learning is performed with only sequence class labels rather than full spatiotemporal segmentation. The resulting classifier is capable of both recognizing and localising activities in video. Furthermore, we demonstrate that efficient matching can be used to obtain real-time action recognition on video sequences.

II. RELATED WORK

The use of the spatial representation of local features within the field of object recognition has shown considerable success [3], [4], [5] and the temporal recognition of actions have been extended. A sparse selection of local interest points is used in many action recognition methods.. Sch¨uldt *et al.* [6] and Dollar *et al.* [7] make use of sparse spatio-temporal features for the human action recognition. Sch¨uldt takes a codebook and bag-of-words approach practical to single images to turnout a histogram of enlightening words or features for each action. Niebles and Fei-Fei [8] use a hierarchical model which can be characterized as a gathering of bags-of-words. Similarly Dollar take the bag-of-words approach but argue for an even sparser sampling of the interest points which improves the concert on the same video sets. Conversely, the choice of feature used with such a sparse set of points is important. Scovanner *et al.* [9] extended the 2D SIFT descriptor [10] into three dimensions, by adding a further dimension to the orientation histogram that encodes temporal information and significantly outperforms the 2D version. To sulpt motion between frames, optical flow [11] [12] can be applied as was used by Laptev [6] in addition to a shape model to detect drinking and smoking actions. Yang Song *et al.* [13] use a triangular lattice of grouped point features to encode layout.

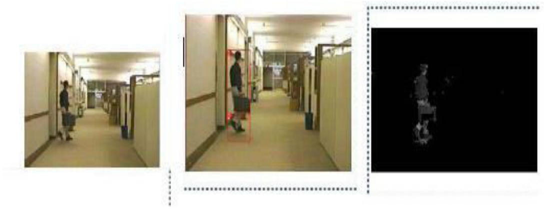
In the early work in action recognition was tested on moderately simple, single person, uniform background sequences [14], [22]. Laptev and Pe´rez [15] expanded the ideas proposed by Ke et al. [16] to apply volumetric features to optical flow [17], [18]. Uemura et al. [15] used a motion model based on optical flow combined with SIFT feature correlation in order to accurately classify multiple actions on a sequence containing large motion and scale changes. A further idea that is being exploited to achieve success on complicated data sets is that of identifying context. Han et al. [19] and Marszalek et al. [20] learn the context of the environment in addition to the actual action. Han applies object recognition to learn relationships such as the number of objects and distance between them in order to boost a standard SIFT-based HoF/HoG [21] bag-of-words approach. Marszalek et al. [20] build on the previous work by Laptev et al. [21] by learning the context in which actions occur. Therefore, by detecting the scene in which the action is occurring, the action classification can be improved. The scene model is learned using 2D Harris corners with SIFT descriptors, while using the HoF and HoG descriptors of Laptev [21] to recognize the action.

The scale of the data sets in temporal-based action recognition directly lends itself to data mining algorithms, especially where only weak supervision is available. However, most previous applications of mining have been within the imaging field. Tesic et al. [28] used a data mining approach to find the spatial associations between classes of texture from aerial photos. While Quack et al. [10] applied Association rule data mining to object recognition by mining spatially grouped SIFT descriptors.

III. FEATURES

1. BACKGROUND MODELLING

An important part of tracking process in video sequences is background subtraction and removal for tracking motion objects. By removing static background we can accurately track the motion object and analyze their movements. Using spatial features of the image and remove background is the older technique for background removal where the new technique is the use of temporal features which improves the background subtraction. Tracking motion objects in video sequences is a multiple process. This processes show in Fig. 1. For motion detection process background subtraction is a part by which we remove background, and only shadows, motion objects and motion noises remains. In the field of background subtraction there are many methods in that the older methods are based on object features such as color, intensity, edges, texture, etc. and relationship between frames is not considered in these methods.



(a) (b) (c)

Fig1. (a) Original frame (b) Finding the moving object (c) The image after background subtraction

2. EXTRACTING TEMPORAL 2D HARRIS INTEREST POINTS

Similar to sparse feature detectors, we fabricate our detection system lead to corner features. The underlying principle for using corners are they are easy to work out, largely invariant to both lighting and geometric transformation, and afford an over-complete feature set from which more complex compound features are constructed. Harris corner detector is the well known method to identify and locate interest points in the image. In our work, 2D Harris corner [24] method is used. Laptev and Lindeberg [25] proposed 3D corners as simple features in (x,y,t). These 3D corners are sparse, so instead, 2D corners are used independently by which the gradient interest points are found independently which yields information on spatial and temporal image changes is much denser than 3D Harris corners [25].

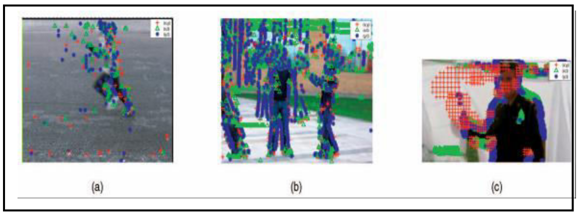


Fig2. 2D Harris corner detection on frames (a) Running (b) Boxing and Handclapping (c) Hug-person

Every corner feature has a dominant gradient orientation used to encode the feature type into single set of discrete corner orientations. Figure 1 shows the example corner detections on three frames. To detect corners at different scales the interest point detector was applied to the video sequences across scale space to overcome the effects of scale. This was achieved by successively 2x2 blocks averaging the image frames. Table 1 shows the scale, image size and effective interest point patch sizes.

Table 1. Table showing the image and relative interest point patch sizes

Scale	1	2	3	4
Image size	160x120	80x60	40x30	20x15
Interest point size	3x3	6x6	24x24	48x48

III. SCALE-INVARIANT GROUPING

The key for object recognition is the scale invariant features which significantly improve the action recognition when modeled from temporal information independently. Quack *et al.* encoded the spatial layout of features by quantising the space around a feature into a grid and assigning features to one of those locations. In order to provide robustness to the scale, the size of the grid is dependent on the scale of the detected SIFT feature. This approach is difficult for achieving less descriptive interest points such as corners, so our approach is to define neighbourhoods centred upon the feature that encode the relative displacement in terms of angle rather than distance hence achieving scale invariance. To perform this, each detected interest points should form the centre of the neighbourhood.

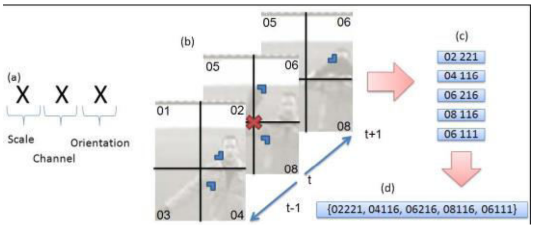


Fig3. a) local feature descriptor, (b) interest points with local features shown as corners, (c) spatial and temporal encoding for local features and (d) for this interest point, local features concatenated into transition vector.

IV. DATA MINING

Data mining allows large amounts of data to be processed to identify any reoccurring patterns within the data in a computationally efficient manner. Association rule [25] mining is the well know mining algorithm. This was initially developed to analyze shopping done by customers in supermarkets, to the find the regularity of those customers in shopping behaviour. Using this millions of transactions the association rule is derived. An association rule, is a relationship of the form {A,B}=> C, where A, B, and C are sets of items. A support and a confidence value plays important role because using this value only the belief of the rule is measured. From numerous transactions possible association rules are produced, to formulate the rules easily and quickly an efficient algorithm is developed. Apriori algorithm developed by Agrawal is the popular algorithm to formulate the rules. The Apriori algorithm is a generative algorithm that uses a breadth-first, bottom-up strategy to explore item sets of increasing size, starting from single item-item sets and increasing the item set.

The frequency of an item set is related to the support and confidence for an association rule. An association rule of the form A =>B is evaluated by looking at the relative frequency of its antecedent and consequent parts, i.e., the item sets A and B. By using the statistical significance, support of the item set is measured, i.e., the probability that the item sets in the Transaction. For A, this is calculated as the size of the set of all T such that T is an element of D and A is a subset of T, normalized by the size of D. This can be formalized using set builder notation as,

$$sup(A) = \frac{|\{T \mid T \in D, A \subseteq T\}|}{|D|} \in \Re \rightarrow [0,1). \qquad (1)$$

The support of the rule A=> B is therefore,

$$sup(A \Rightarrow B) = \frac{|\{T \mid T \in D, (A \cup B) \subseteq T\}|}{|D|}, \qquad (2)$$

and the statistical significance of the rule is measured. Then the confidence of the rule is calculated as,

$$conf(A \Rightarrow B) = \frac{sup(A \cup B)}{sup(A)} = \frac{|\{T \mid T \in D, (A \cup B) \subseteq T\}|}{|\{T \mid T \in D, A \subseteq T\}|}. \qquad (3)$$

The probability of the joint occurrence of A and B is support, i.e., P(A,B), while confidence is the conditional probability P(B/A).

All generated association rules would be maintained and for other action classes the confidence would be used as a discrimination measure. This would be computationally infeasible due to the complete number of rules; therefore, generated rules are filtered using both support and confidence. At the initial level a single support value which is the lowest value is used throughout all of the stages of mining that is computationally feasible.

V. ACTION CLASSIFICATION

The frequently reoccurring distinctive and descriptive compound features for each class, are produced after the training phase is completed. By using the frequent item sets obtained from the data mining, the action classification of the unseen video sequence is done. Each transaction confidence is used to weight the matches and indicates that the Transaction T is distinctive compared to other classes using the high confidence. The use of the confidence ensures that if the transaction is matched with several classes, the confidence will provide a measure of the discrimination between those classes. If there are no matches found in the unlikely event, the model score is zero and the video would be classed as not containing any action.

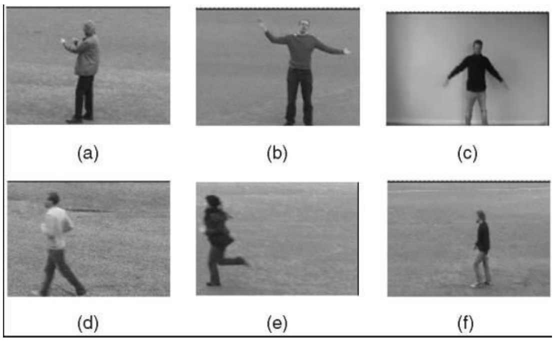


Fig4. Action classification examples (a) boxing, (b) hand clapping, (c) hand waving, (d) jogging, (e) running, and (f) walking.

VI. EXPERIMENTAL RESULTS

To estimate the approach proposed, two different data sets were used. To illustrate the generalization method the focus of each data set is different. The well-known and popular KTH data set by Schuldt et al., to provide a comparison with the other existing techniques containing 6 different actions; boxing, hand-waving, hand-clapping, jogging, running and walking are shown in fig4. The simultaneous multi-action Multi-KTH data set , demonstrates detection of multiple actions in noisy scenes with background confusion and a moving camera.

To provide an additional challenge, the Multi-KTH data set was proposed. It consists of a single 753 frame long sequence, where multiple people perform the KTH actions simultaneously. To increase difficulty, there are large changes in scale, camera motions, and a non uniform background. Some frames from the sequence are shown in Fig.5.

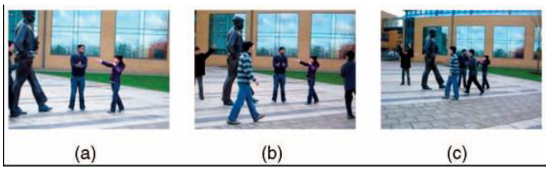
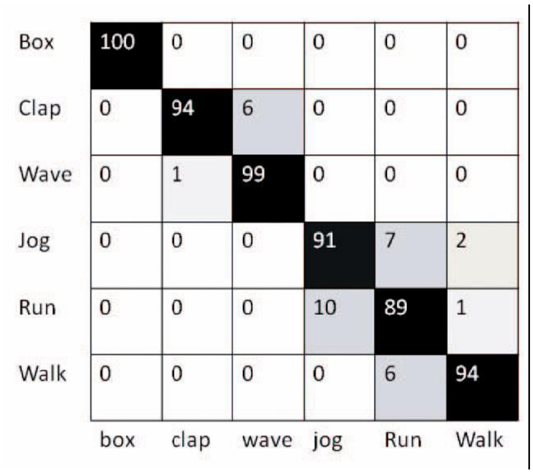


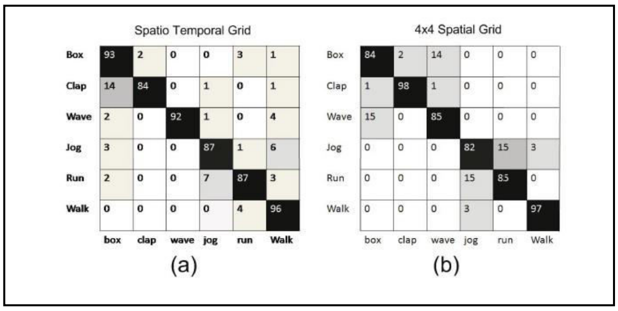
Fig5. Example for the Multi-KTH data set, performing multiple KTH actions



	box	clap	wave	jog	Run	Walk
Box	100	0	0	0	0	0
Clap	0	94	6	0	0	0
Wave	0	1	99	0	0	0
Jog	0	0	0	91	7	2
Run	0	0	0	10	89	1
Walk	0	0	0	0	6	94

Fig6. Convolution matrix results for the KTH data set using training/test partition

To produce a Frequent Mined Configuration vector M the training sequences were used for each of the six actions containing compound features used to classify each of the test sequences. Figure 7(a) shows the classification confusion matrix using the scale invariant grid approach proposed within this paper, where good class separability is exhibited. The results show relatively little confusion compared to other approaches with minor confusion, this is consistent with previous approaches and it’s due to the inherent similarity of the motion. In Figure 4(b) the experiments were repeated using a fixed size 4x4 grid similar, to investigate the importance of the spatial and temporal compounding of individual features.



Spatio Temporal Grid

	box	clap	wave	jog	run	Walk
Box	93	2	0	0	3	1
Clap	14	84	0	1	0	1
Wave	2	0	92	1	0	4
Jog	3	0	0	87	1	6
Run	2	0	0	7	87	3
Walk	0	0	0	0	4	96

(a)

4x4 Spatial Grid

	box	clap	wave	jog	run	Walk
Box	84	2	14	0	0	0
Clap	1	98	1	0	0	0
Wave	15	0	85	0	0	0
Jog	0	0	0	82	15	3
Run	0	0	0	15	85	0
Walk	0	0	0	3	0	97

(b)

Fig7. The confusion of the data mined corner on the Kth dataset (a) Scale invariant spatial grouping (b) non-scale invariant spatial grouping.

Compared to other published methods Mined dense corners has a high classification accuracy, to select optimal low level features for discriminative classification.

Table 2.Comparison of Average precision with other techniques on KTH action recognition Dataset

Method	Average Precision
Nowozin *et al.* [23] Subseq Boost SVM	87.04%
Wong and Cipolla [24] Subspace SVM	86.60%
Niebles *et al.* [25] pLSA model	81.50%
Dollar *et al.* [5] Spat-Temp	81.20%
Schüldt *et al.* [1] SVM Split	71.71%
Ke *et al.* [3] Vol Boost	62.97%
Fixed Grid Mined Dense Corners	88.50%
Scale Invariant Mined Dense Corners	89.92%

The Multi-KTH data set is a more challenging version of the KTH data set. It has the same six actions and training video sequences, but the test sequence consists of multiple simultaneous actions, with significant camera motion. In the multi-KTH data set, the localization is performed to discriminate the multiple actions performed in the video sequence. For static actions the localization is generally focused on person's upper body and face, and is focused on legs for the dynamic action, because the legs consists of the descriptive features.

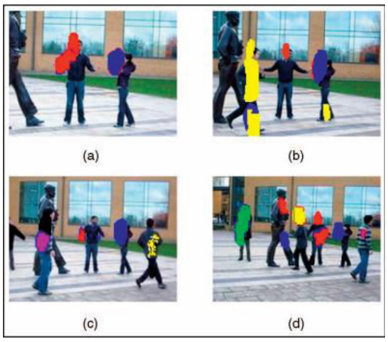


Fig8. Example for multi-KTH localization Red-handclapping, Blue-boxing, Yellow-running, pink-walking and Green-hand waving

Figure 8 shows the various actions performed in the video sequence. To discriminate the multiple actions, different colors are used. Red-handclapping, Blue-boxing, Yellow-running, pink-walking and Green-hand waving. By using these colors the multiple actions performed in the video sequence are discriminated. The sample localization performed in our paper is shown in figure 9.

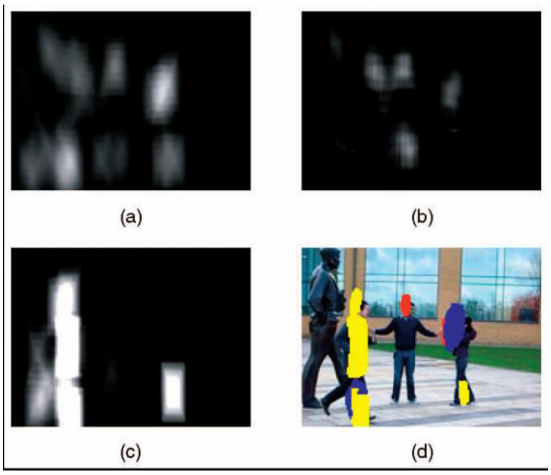


Fig9. Results from multi-KTH data set using localization (a) boxing, (b) handclapping, (c) running and (d) output frame

The main advantage of using the data mining technique is the speed of the learning patterns compared to the other machine language approaches. When compared to other techniques the simple 2D Harris corner detection has relatively a low computational cost. The spatial neighborhood grouping is fast to encode features which has limited neighborhood.

VII.CONCLUSION

This paper has presented an efficient solution to the problem of recognizing actions within video sequences with efficient learning of informative and descriptive local features for actions performed by humans at multiple scales. To form complex discriminative compounds of simple 2D Harris corners mined grouping corners are used which is fast. In a weakly supervised approach the frequently reoccurring patterns are learned by using data mining approach where only class labels are required. Two different data sets have been tested, the Multi-KTH data set required multiple action localization and the KTH data set provides a comparison to other approaches. When tested on the popular KTH and multi-KTH dataset, notable results are obtained which do better than other state-of-the-art approaches. During training object segmentation is not required. To perform activity localization as well as classification the final classifiers can be used.

REFERENCES

[1] Andrew Gilbert, John Illingworth and Richard Bowden, ” Action Recognition Using Mined Hierarchical Compound Features”, Transactions on Pattern Analysis and Machine Intelligence vol 33, no.5, may 2011.

[2] Andrew Gilbert, John Illingworth, and Richard Bowden,“ Scale Invariant Action Recognition Using Compound Features Mined from Dense Spatio-temporal Corners”, CVSSP, University of Surrey, Guildford, GU2 7XH, England

[3] T. Quack, V. Ferrari, B. Leibe, and L. VanGool, “Efficient Mining of Frequent and Distinctive Feature Configurations,” Proc.

[4] S. Lazebnik, C. Schmid, and J. Ponce, “Semi-Local Affine Parts for Object Recognition,” Proc. BMVA British Machine Vision Conf., vol. II pp. 959-968, 2004.

[5] J. Sivic and A. Zisserman, “Video Data Mining Using Configurations of Viewpoint Invariant Regions,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, vol. I, pp. 488-495, 2004.

[6] Schuldt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: a Local SVM Approach. In: Proc. of International Conference on Pattern Recognition (ICPR 2004), vol. III, pp. 32–36 (2004).

[7] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior Recognition via Sparse Spatio-temporal Features. In: ICCCN 2005: Proceedings of the 14th International Conference on Computer Communications and Networks, pp. 65–72 (2005)

[8] . Niebles, J.C., Fei-Fei, L.: A Hierarchical Model of Shape and Appearance for Human Action Classification. In: Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2007) (2007)

[9]. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proc. of MULTIMEDIA 2007, pp. 357–360 (2007)

[10]. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 20, 91–110 (2003)

[11]. Dalal, N., Triggs, B., Schmid, C.: Human Detection using Oriented Histograms of Flow and Apperance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 428–441. Springer, Heidelberg (2006)

[12]. Lucas, B., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In: Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674–679 (1998) Scale Invariant Action Recognition 233

[13]. Song, Y., Goncalves, L., Perona, P.: Unsupervised Learning of Human Motion. Transactions on Pattern Analysis and Machine Intelligence 25, 814–827 (2003)

[14] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing Human Actions: A Local SVM Approach,” Proc. Int’l Conf. Pattern Recognition, vol. 3, pp. 32-36, 2004.

[15] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as Space-Time Shapes,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp.2247-2253, Dec. 2007.

[16] I. Laptev and P. Pe´rez, “Retrieving Actions in Movies,” Proc. IEEE Int’l Conf. Computer Vision, 2007.

[17] Y. Ke, R. Sukthankar, and M. Hebert, “Efficient Visual Event Detection Using Volumetric Features,” Proc. IEEE Int’l Conf. Computer Vision, 2005.

[18] N. Dalal, B. Triggs, and C. Schmid, “Human Detection Using Oriented Histograms of Flow and Appearance,” Proc. European Conf. Computer Vision, vol. II, pp. 428-441, 2006.

[19] B. Lucas and T. Kanade, “An Iterative Image Registration Technique with an Application to Stereo Vision,” Proc. Seventh Int’l Joint Conf. Artificial Intelligence, pp. 674-679, 1998.

[20] M. Marszalek, I. Laptev, and C. Schmid, “Actions in Context,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2009.

[21] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 1-8, 2008.

[22] J. Tesic, S. Newsam, and B.S. Manjunath, "Mining Image Datasets Using Perceptual Association Rules," Proc. SIAM Int'l Conf. Data Mining, Workshop Mining Scientific and Eng. Datasets, p. 7177, 2003.

[23] T. Quack, V. Ferrari, B. Leibe, and L. VanGool, "Efficient Mining of Frequent and Distinctive Feature Configurations," Proc. 11th IEEE Int'l Conf. Computer Vision, 2007.

[24] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," Proc. Alvey Vision Conf., pp. 189-192, 1988.

[25] I. Laptev and T. Lindeberg, "Space-Time Interest Points," Proc. IEEE Int'l Conf. Computer Vision, pp. 432-439, 2003

◈◈◈

Secured Data Transmission using Snort Rules and Mining Techniques

[1]R.Venkatramana & [2]Mrs.M.Sreedevi

Research Scholar Department of CSE
Madanapalli Institute of Technology and Science, Madanapalli, Andhra Pradesh, India.
E-mail : rvramana.r@gmail.com

Abstract - Network traffic analysis becomes more and more crucial in the IP network infrastructure as the amount of IP packets transmitted on the Internet at any given moment of time increases enormously. A thorough understanding of the IP traffic will help us better design our network topology and utilize bandwidth more effectively. From the perspective of security, it can also protect our system from attacks, such as intrusions, our model employs feature selection so that the binary classifier for each type of attack can be more accurate, which improves the detection of attacks that occur less frequently in the training data. Based on the accurate binary classifiers, our model applies a new ensemble approach which aggregates each binary classifier's decisions for the same input and decides which class is most suitable for a given input. During this process, the potential bias of certain binary classifier could be alleviated by other binary classifiers' decision. Our model also makes use of multi boosting for reducing both variance and bias. The clients have some rules to communicate between them using snort rules. Any Communications (such as FTP, SMTP, etc) between the clients are monitored by the snort. If it continues again, then that particular client will be disconnected from this network (means cannot be able to communicate with other clients in that network.) But, that client will be physically connected with the network. The proposed work describes a network traffic analysis software tool, which provides searching, visualization, and preprocessing functions with a user-friendly GUI implemented in Java language. Within the huge network traffic data collected, a user can identify any particular packets using various searching functions provided. Visualization presents the analyzed result in a different setting to further enhance the analysis. The GUI in Java allows the tool to be used in different platforms. This tool is tested and demonstrated through several real network datasets.

Keywords - *Metal Algorithms, filtering algorithms, finite-state automata (FSA), mathematics, packet filters, packet processing, predicate optimization, protocol description languages (PDLs), run-time safety, snort rules and mining techniques.*

I. INTRODUCTION

Packet filters are a class of packet manipulation programs used to classify network traffic in accordance to a set of user-provided rules; they are a basic component of many networking applications such as shapers, sniffers, demultiplexers, firewalls, and more.The modern networking scenario imposes many requirements on packet filters, mainly in terms of processing speed (to keep up with network line rates) and resource consumption (to run in constrained environments). Filtering techniques should also support modern protocol formats that often include cyclic or repeated structures (e.g., MPLS label stacks, IPv6 extension headers). Finally, it is also crucial that filters preserve the integrity of their execution environment, both in terms of memory access safety and termination enforcement, especially when running as an operating system module or on the bare hardware. Although at first sight this aspect might not seem crucial, it is a fact that many of the limitations built into existing packet filters derive directly from safety issues. As an example, the impossibility of automatically proving termination for a generic computer program led the BPF [1] designers to generate acyclic filters only, thus preventing the parsing of packets with multiple levels of encapsulation or repeated field sequences.

Existing packet filters focus invariably on subsets of these issues but, to the best of our knowledge, do not solve all of them at the same time. As an example, two widely known generators, BPF [2] and PathFinder [3], do not support recursive encapsulation; NetVM-based filters [4], on the other hand, have no provision for enforcing termination, either in filtering code or in the underlying virtual machine.

This paper presents Stateless PAcket Filter (SPAF), a finite-state automata (FSA)-based technique to generate fast and safe packet filters that are also flexible enough to fully support most layer-2 to layer-4 protocols, including optional and variable headers and recursive encapsulation. The proposed technique specifically targets the lower layers of the protocol stack and does not directly apply for deep packet inspection nor for stateful filtering in general. Moreover, for the purpose of this paper, we consider only static situations where on-the-fly rule set updates are not required. While these limitations exclude some interesting use cases, SPAF filters are nevertheless useful for a large class of applications, such as monitoring and traffic trace filtering, and can serve as the initial stage for more complex tools such as intrusion detection systems and firewalls.

A stateless packet filter can be expressed as a set of predicates on packet fields, joined by boolean operators; often these predicates are not completely independent from one another, and the evaluation of the whole set can be short-circuited. One of the most important questions in designing generators for high-performance filters is therefore how to efficiently organize the predicate set to reduce the amount of processing required to come to a match/mismatch decision. By considering packet filtering as a regular language recognition problem and exploiting the related mathematical framework to express and organize predicates as finite-state automata, SPAF achieves by construction a reduction of the amount of redundancy along any execution path in the resulting program: Any packet field is examined at most once. This property emerges from the model, and it always holds even in cases that are hard to treat with conventional techniques, such as large-scale boolean composition. Moreover, thanks to their simple and regular structure, finite automata also double as an internal representation directly translatable into an optimized executable form without requiring a full-blown compiler.

Finally, safety (both in terms of termination and memory access integrity) can be enforced with very low run-time overhead.

The rest of this paper is structured as follows. Section II presents an overview of the main related filtering approaches developed to this date. Section III provides a brief introduction to the FSAs used for filter representation and describes the filter construction procedure. Section IV focuses on executable code generation and on enforcing the formal properties of interest, Finally, Section V reports conclusions and also highlights possible future developments.

II. RELATED WORK

Given their wide adoption and relatively long history, there is a large corpus of literature on packet filters. A first class of filters is based on the CFG paradigm; the best-known and most widely employed one is probably BPF [1], the Berkeley Packet Filter. BPF filters are created from protocol escriptions hardcoded in the generator and are translated into a bytecode listing for a simple, *ad hoc* virtual machine. The bytecode was originally interpreted, leading to a considerable run-time overhead impact that can be reduced by employing JIT techniques [5]. BPF disallows backward jumps in filters in order to ensure termination, thus forgoing support for, e.g., IPv6 extension headers; memory protection is enforced by checking each access at run-time. Multiple filter statements can be composed together by boolean operators, but in the original BPF implementation, only a small number of optimizations are performed over predicates, leading to run-time inefficiencies when dependent or repeated predicates are evaluated. Two relevant BPF extensions are BPF and xPF. BPF [2] adds local and global data-flow optimization algorithms that try to remove redundant operations by altering the CFG structure. xPF [6] relaxes control flow restrictions by allowing backward jumps in the filter CFG; termination is enforced by limiting the maximum number of executed instructions through a run-time watchdog built into the interpreter, but its overhead was not measured, and extending this approach to just-in-time code emission has not been proposed and might prove difficult.

Afurther CFG-based approach, unrelated to BPF, is described in [4]. Its main contribution is decoupling the protocol database from the filter generator by employing an XML-based protocol description language, NetPDL [7]. Filtering code is executed on the NetVM [8], a special-purpose virtual machine targeting network applications that also provides an optimizing JIT compiler that works both on filter structure and low-level code. The introduction of a high-level description language reportedly does not cause any performance penalties; this approach, however, delegates all safety considerations to the VM and does not provide an effective way to compose multiple filters.

In general, CFG-based generators benefit from their flexible structure that does not impose any significant restriction on predicate evaluation order; for the same reason, however, they are prone to the introduction of hard-to-detect redundancies, leading to multiple unnecessary evaluations if no further precautions are taken. Even when optimizers are employed and are experimentally shown to be useful, they work on an opportunistic basis and seldom provide any hard guarantees on the resulting code.

A second group of filter generators chooses tree-like structures to organize predicates. PathFinder [3] transforms predicates into template masks (atoms), ordered into decision trees. Atoms are then matched through a linear packet scan until a result is reached. Decision trees enable an optimization based on merging prefixes that are shared across multiple filters. PathFinder is shown to work well both in software and hardware implementations, but it does not take protocol database decoupling into consideration, and no solution to memory safety issues is proposed for the software implementation. FSA-based filters share a degree of similarity with PathFinder as packets are also scanned linearly from the beginning to the end, but predicate organization, filter composition, and safety considerations are handled differently. DPF [9] improves over PathFinder by generating machine code just-in-time and adding low-level optimizations such as a flexible switch emission strategy. Moreover, DPF is capable of aggregating bounds checks at the atom level by checking the availability of the highest memory offset to be read instead of considering each memory access in isolation; our technique, described in Section IV-E, acts similarly but considers the filter as a whole, thus further reducing run-time overhead.

While organizing predicates into regular structures makes it easier to spot redundancies and other sources of overhead, it also introduces different limitations. As an example, generators restricted to the aforementioned acyclic structures do not fully support tunneling or repeated protocol portions. Moreover, it has been noted that performing prefix coalescing is not sufficient to catch certain common patterns, resulting in redundant predicate evaluation [2].

A third approach is to consider packet filtering as a language recognition problem. Jayaram *et al.* [10] use a pushdown automaton to perform packet demultiplexing; filters are expressed as LALR(1) grammars and can therefore be effectively composed using the appropriate rules. This solution improves filter scalability, but there are downsides related to the push-down automaton: A number of specific optimizations are required to achieve good performance. It is also quite unwieldy to express protocols and filter rules as formal grammars that must be kept strictly unambiguous: The authors marginally note that the simpler FSA model would be sufficient for the same task.

Apart from the specialized solutions for fast packet filtering mentioned, one of the most widely used packet filtering programs is the NetFilter framework.1 NetFilter is a component of the Linux kernel that performs packet filtering, firewalling, mangling operations (e.g., network address translation), and more, acting through a set of hooks and callbacks that intercept packets as they traverse the networking stack. In contrast with all the aforementioned approaches, NetFilter uses the relatively simple method of applying all the specified rules in sequence when performing packet filtering, leading to poor performance and scalability; moreover, it appears not possible to specify an arbitrary predicate, filters being limited to preset protocols and statements that are specialized by specifying actual network addresses and ports.

Besides the generation technique, there have also been improvements along other dimensions such as architectural considerations, as demonstrated by xPF, FFPF [19], and nCap [20], or dynamic rule sets support, as shown by the SWIFT tool [21]. We consider these aspects out of scope for the purpose of this paper, being either orthogonal to the technique we present or the object of future works.

Definition of Near-Duplicate

The central idea of near-duplicate spam detection is to exploit reported known spams to block subsequent ones which have similar content. For different forms of e-mail representation, the definitions of similarity between two e-mails are diverse. Unlike most prior works representing e-mails based mainly on content text, we investigate representing each e-mail using an HTML tag sequence, which depicts the layout structure of e-mail, and look forward to more effectively capturing the near-duplicate phenomenon of spams.

Let I ¼ ft1; t2; . . . ; ti; . . . ; tn;

<mytext=>;

<anchor>g be the set of all valid HTML tags

with two types of newly created tags,

<mytext=> and

<anchor>, included. An e-mail abstraction derived from

procedure SAG is denoted as <e1; e2; . . . ; ei; . . . ; em>, which

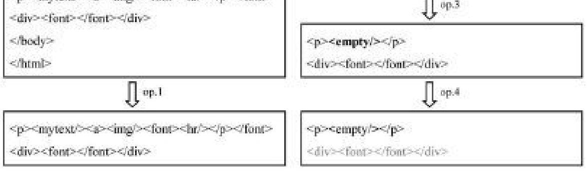
is an ordered list of tags, where ei 2 I. The definition of near duplicate

is: “Two e-mail abstractions _ ¼ <a1; a2; . . . ;

ai; . . . ; an> and _ ¼ <b1; b2; . . . ; bi; . . . ; bm> are viewed as

near-duplicate if 8ai ¼ bi and n ¼ m.”

The tag length of an e-mail abstraction is defined as the number of tags in an e-mail abstraction.



```
<html>
<head><meta/></head>
<body>
<p><mytext/><a><img/><font><hr/></p></font>
<div><font></font></div>
</body>
</html>
```

```
<p><mytext/><a><img/><font><hr/></p></font>
<div><font></font></div>
```

```
<p><mytext/><a><img/><font><hr/></p></font>
<div><font></font></div>
```

```
<p><empty/></p>
<div><font></font></div>
```

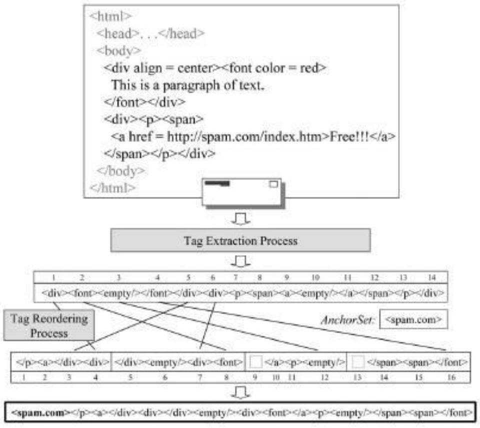
```
<p><empty/></p>
<div><font></font></div>
```

The following sequence of operations is performed in the preprocessing step.

1. Front and rear tags are excluded.

2. Nonempty tags2 that have no corresponding start tags or end tags are deleted. Besides, mismatched nonempty tags are also deleted.
3. All empty tags2 are regarded as the same and are replaced by the newly created <empty=> tag. Moreover, successive <empty=> tags are pruned and only one <empty=> tag is retained.
4. The pairs of nonempty tags enclosing nothing are removed.

Example for Mail

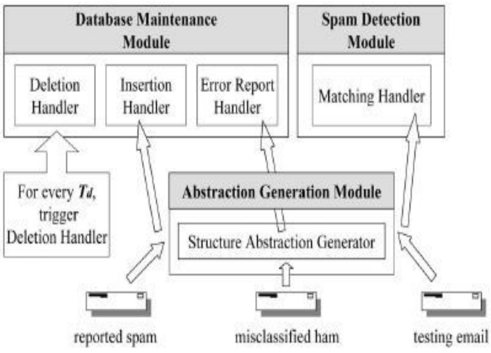


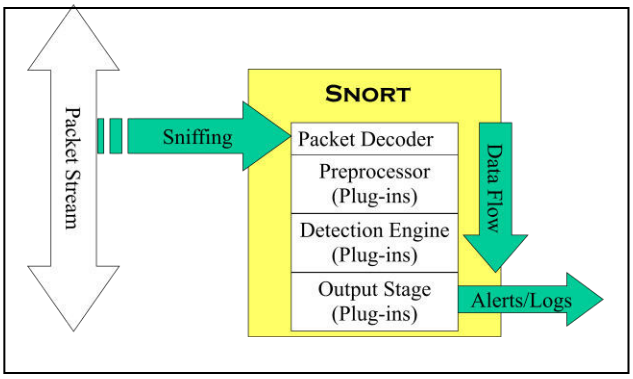
```
<html>
 <head>. . .</head>
 <body>
  <div align = center><font color = red>
   This is a paragraph of text.
  </font></div>
  <div><p><span>
   <a href = http://spam.com/index.htm>Free!!!</a>
  </span></p></div>
 </body>
</html>
```

Tag Extraction Process
1 2 3 4 5 6 7 8 9 10 11 12 13 14
<div><font><empty/></font></div><div><p><span><a><empty/></a></span></p></div>
Tag Reordering Process
AnchorSet: <spam.com>
</p><a></div><div> </div><empty/><div><font> </a><p><empty/> </span><span></font>
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
<spam.com></p><a></div><div></div><empty/><div><font></a><p><empty/></span><span></font>

SYSTEM ARCHITECTURE

In the packet filtering process with the increasing popularity of electronic mail (or e-mail), several people and companies found it an easy way to distribute a massive amount of unsolicited messages to a tremendous number of users at a very low cost. These unwanted bulk messages or junk emails are called spam messages. The majority of spam messages that has been reported recently are unsolicited commercials promoting services and products including sexual enhancers, cheap drugs and herbal supplements, health insurance, travel tickets, hotel reservations, and software products. They can also include offensive content such pornographic images and can be used as well for spreading rumors and other fraudulent advertisements such as make money fast.

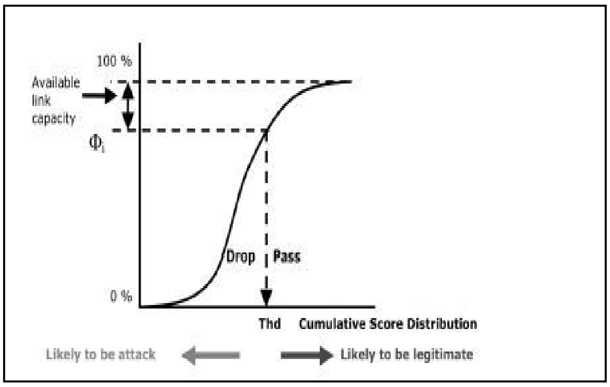




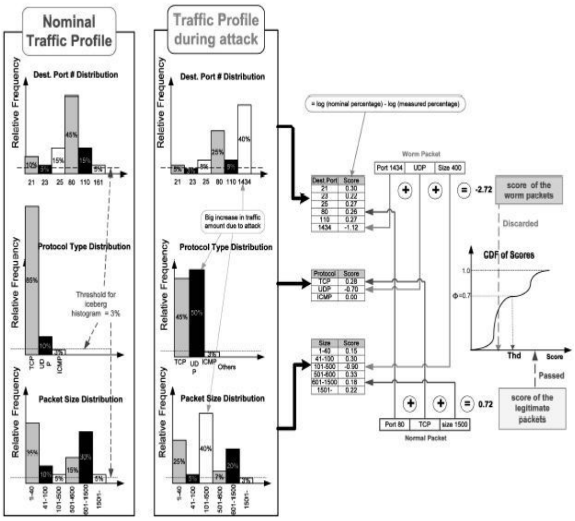
As a result, spam has become an area of growing concern attracting the attention of many security researchers and practitioners. In addition to regulations and legislations, various anti-spam technical solutions have been proposed and deployed to combat this problem. Front-end filtering was the most common and easier way to reject or quarantine spam messages as early as possible at the receiving server. However most of the early anti-spam tools were static; for example using a blacklist of known spammers, a white list of good sources, or a fixed set of keywords to identify spam messages. Although these list-based methods can substantially reduce the risk provided that lists are updated periodically, they fail to scale and to adapt to spammers' tactics.

Selective Packet Discarding

Once the score is computed for a packet, selective packet discarding, and overload control can be performed using the score as the differentiating metric. Since an exact prioritization would require offline, multiple-pass operations, e.g., sorting and packet buffering, the following alternative approach is taken into account. First, the cumulative distribution function (CDF) of the scores of all incoming packets in time period (Ti) is maintained. Second, the cut-off threshold score is calculated. Third, the arriving packets in time period T (i+1) if its score value is below the cut-off threshold are discarded. At the same time, the packets arriving at T (i+1) create a new CDF.



Selective packet discarding



Discarding SQL Slammer Worm attack packets

III. FILTER GENERATION TECHNIQUE

In this section we describe about filtering techniques using snort rules and mining techniques.

One of the key concepts in PacketScore is the notion of “Conditional Legitimate Probability” (CLP) based on Bayesian theorem. CLP indicates the likelihood of a packet being legitimate by comparing its attribute values with the values in the baseline profile. Packets are selectively discarded by comparing the CLP of each packet with a dynamic threshold. The concept of using a baseline profile with Bayesian theorem has been used previously in anomaly-based IDS (Intrusion Detection System) applications, where the goals are generally attack detection rather than real-time packet filtering.

In this research, the basic concept to a practical real-time packet filtering scheme using elaborate processes is extended. In this method, the PacketScore operations for single-point protection is described, but the fundamental concept can be extended to a distributed implementation for core-routers.

To make it more suitable for real-time processing, conversion of floating-point division/multiplication operations into subtraction/addition operations is made. Scoring a packet is equivalent to looking up the scorebooks, e.g., the TTL scorebook, the packet size scorebook, the protocol type scorebook, etc. After looking up the multiple scorebooks, the matching CLP entries in a log-version scorebook are added. This is generally faster than multiplying the matching entries in a regular scorebook. The small speed improvement from converting a multiplication operation into an addition operation is particularly useful because every single packet must be scored in real-time. This speed improvement becomes more beneficial as the number of scorebooks increases. On the other hand, generating a log-version scorebook may take longer than a regular scorebook generation. However, the scorebook is generated only once at the end of each period and it is not necessary to observe every packet for scorebook generation; thus, some processing delay can be allowed. Furthermore, scorebook generation can be easily parallelized using two processing lines, which allows complete sampling without missing a packet.

The purpose of a stateless packet filter generator is to create a program that, given a finite-length byte sequence (a packet) as its input, returns a binary match/mismatch decision. The input of the generator itself consists of a set of filter rules provided by the user that specify the desired properties of matching packets; each rule, in turn, consists of multiple predicates expressed in a simple high-level language (where header fields and protocols appear symbolically), combined together with boolean operators. In older generators, the set of supported protocols was fixed; in modern ones protocol header formats are kept into an external database that can be updated without modifying the generator.

In order to develop a successful FSA-based filtering technique, it is first of all necessary to show that any filter of interest can be expressed as a finite automaton, then provide a method to transform a high-level filter statement and a protocol database into FSA form. Finally, the resulting automaton must be translated into an efficiently executable form.

A. Protocol Database Compilation

The first phase in the SPAF generation process consists of parsing the protocol database and building template automata that recognize all the correctly formatted headers for a given protocol. These automata will be reused and specialized in later phases to create the final filter.

In order to decouple filter generation from the protocol database, we have employed an XML-based protocol description language (NetPDL [7]) designed to describe the on-the-wire structures of network protocols and their encapsulation relationships. NetPDL descriptions are stored in external files that can be freely edited without modifying the generator itself.

A precise description of NetPDL is beyond the scope of this paper. Nevertheless, we shall provide a quick overview of the features supported by the FSA generator. The language provides a large number of primitives that enable the description of header formats of layer-2–7 protocols, but for the scope of this work we have restricted our support to those designed for layer-

2–4 decoding. The basic building block of a protocol format is the header field, a sequence of bytes or bits that can be either fixed or variable in size. Adjacent fields are by default laid out in sequence, but more complex structures such as optional or repeated sections can be created using conditional choices and loops; these statements are controlled by expressions that can contain references to the values of previously encountered fields.

A second NetPDL portion contains a sequence of control flow operations (if, switch) that predicate encapsulation relationships. In general, the control flow is followed until a nextproto tag is encountered, specifying which is the next protocol to be found in the packet. A NetPDL database thus

```
<protocol  name =”ipv6”>
    <format>
        <field>
<field type=”bit” name = ”ver” mask=”0xF0000000” size =”4”/>
<field type=”bit” name = ”tos” mask=”0x0F000000”size =”4”/>
<field type=”bit” name = ”flabel” mask=”0x00FFFFFF”size =”4”/>
<field type=”fixed” name=”plex”  size =”2”/>
<field type=”nexthdr” name=”plex”  size =”1”/>
<field type=”hop” name=”plex”  size =”16”/>
<field type=”src” name=”plex”  size =”16”/>
<field type=”dst” name=”plex”  size =”16”/>
<loop type =”while” expr=”1”>
<switch expr=”nexthdr”>
<case value=”0”><includeblk name=”HBH”/></case>
<case value=”0”><includeblk name=”AH”/></case>
….
<default>
<loopctrl type=”break”/>
</default>
    </switch>
       </loop>
          </fields>
            </format>
<encapsulation>
<switch expr=”nexthdr”>
<case value=”4”> <nextproto proto=”#ip”/></case>
<case value=”4”> <nextproto proto=”#tcp”/></case>
<case value=”4”> <nextproto proto=”#udp”/></case>
….
</switch>
   </encapsulation>
       </protocol>
```

IPV6 NetPDL excerpt

describes an oriented encapsulation graph where the vertices are protocols and the edges are encapsulation relationships. Currently, the graph begins with a single user-specified root that usually represents the link-layer protocol, but an extension to multiple ones would be trivial. Starting from this root, the FSA generator follows the encapsulation graph and builds a FSA for every reachable protocol using the method explained later in this section.

As an example, a simplified NetPDL description of the IPv6 header format is presented in Fig. 1. IPv6 starts with a sequence of fixed-size fields; bitfields (such as ver) are specified by the mask attribute. The initial portion is followed by a set of extension headers, each one containing a “next header” information (nexthdr). This sequence is of unspecified (but implicitly finite, as any packet is finite) length, and it is described using a switch nested within a loop: At each iteration, the newly read nexthdr field is evaluated, and if no more extension headers are present, the loop terminates. Encapsulation relationships are also specified in a similar fashion by jumping to the correct protocol depending on the value of the last nexthdr encountered.

SPAF currently supports the full versions of the most common layer-2–4 protocols in use nowadays, such as Ethernet, MPLS, VLAN, PPPoE, ARP, IPv4, IPv6, TCP, UDP, and ICMP; this set can be easily extended as long as no stateful capabilities are required.

An important point regarding FSA creation from NetPDL descriptions is that, as long as it is correctly performed, it is not be a critical task for filter performance: Any resulting automaton ultimately will be determinized and minimized, yielding a canonical representation of the filter that does not depend on the generation procedure. For this reason, and given the complexity involved, the NetPDL-to-FSA conversion procedure is not fully described in this paper, and it can be regarded as an implementation detail. Nevertheless, in order to exemplify how the conversion can be done, we report the key steps for translating the NetPDL snippets of Fig. 2 into the corresponding automata.

The purpose of this initial conversion step is not to generate automata immediately suitable for filtering. On the contrary, the results are templates for the following generation steps, representing the "vanilla" version of protocol headers, with no other conditions imposed, to be specialized according to the filter rules. Since they are strictly related to header format, any inputconsuming transition in these templates can be related to a specific portion of one3 header field; this information must be preserved to accommodate the imposition of filtering rules. For this reason, template automata are augmented by marking all the relevant transitions with the related field's name.4

The simplest example is generating an automaton that parses a fixed-length header field [Fig. 2(a)]: It is sufficient to build a FSA that skips an appropriate amount of bytes, resulting in Fig. 2(b). During the construction process, header fields are given well-defined start and end5 states that are used as stitching points to join with any predecessors or successors by -transitions, as required. A more complex example involving a conditional choice is shown in Fig. 2(c). The generation procedure starts by creating automata representations for all the initial fields in the NetPDL description; upon encountering the switch construct, however, the generator backtracks the transition graph until it encounters the type field. Once found, all the states/transitions that follow type (the block in the figure) are replicated. The original copy is left as is, while in the replica the transitions for type are specialized to recognize the bytes of interest for the switch, so the right path will be taken depending on the actual input values. Finally, the correct trailing block (or) is joined in the right place. The last example [Fig. 2(e) and (f)] shows the automata generated for a header structure similar to the IPv6 extension headers case. In this case, a loop is interlocked with a switch construct, and a greater amount of block replication is required to ensure that independent paths exist into the automaton for every possible combination of the current nexth value (upon which the outcome of the switch depends) and the next nexth value, which might cause the loop to end.

Encapsulation relationships are handled in a similar fashion by spawning new paths in the automaton graph that end with a special state marked with the protocol that should follow. The exact usage of these marked states is explained in Section III-D. The generation procedure acts to counter the absence of explicit storage locations in the FSA model; when it becomes necessary to use the values of previously encountered fields for subsequent computations, the only solution is to spawn a number of parallel branches within the automaton, each one associated with a specific value of the field under consideration.

B. Multicast Packet Delivery

Here we discuss about packet forwarding to the nodes

Packet sending from the source

After the multicast tree is constructed, all the sources of the group could send packets to the tree and the packets will be forwarded along the tree. In most tree-based multicast protocols, a data source needs to send the packets initially to the root of the tree.

The source node want send the data to the members at that time we perform the security action, i.e. whenever the source node want to send the data , the source node can encrypt the data by using AES (Advanced Encryption Standers) the encrypted data can be transferred to the group members , in the transmission of packets the intermediate nodes want to read the data , if suppose the nodes can access the data that time we don't have any problem because the data is in the encryption form i.e. cipher text , due to this text the intermediate nodes can't get the data it can simply transfer the data to the destination, in the destination side the receiver can decrypt the data using AES algorithm.

For providing the security we use the Advanced Encrypted Standards Algorithm

The algorithm described by AES is a symmetric-key algorithm, meaning the same key is used for both encrypting and decrypting the data. The strength of a 128-bit AES key is roughly equivalent to 2600-bits RSA key. AES data encryption is a more mathematically efficient and elegant cryptographic algorithm the time required to crack an encryption algorithm is directly related to the length of the key used to secure the communication (It takes less time). AES allows you to choose a 128-bit, 192-bit or 256-bit key, making it exponentially stronger than the 56-bit key of DES (RSA). The algorithm was required to be royalty-free for use worldwide .AES has defined three versions, with 10, 12, and 14 rounds. Each version uses a different cipher key size (128, 192, or 256), but the round keys are always 128 bits.

IV. CONCLUSION

We have designed, prototyped, and evaluated SPAF, a packet filter generator based on the creation of finite-state automata from a high-level protocol format database and filter redicates. SPAF aims at emitting fast and efficient filters while preserving all the relevant safety properties, both in terms of memory access correctness and termination. The PacketScore scheme is used to defend against DDoS attacks. The key concept in PacketScore is the Conditional Legitimate Probability

(CLP) produced by comparison of legitimate traffic and attack traffic characteristics, which indicates the likelihood of legitimacy of a packet. As a result, packets following a legitimate traffic profile have higher scores, while attack packets have lower scores. This scheme can tackle never-before-seen DDoS attack types by providing a statistics-based adaptive differentiation between attack and legitimate packets to drive selective packet discarding and overload control at high-speed.

Thus, PacketScore is capable of blocking all kinds of attacks as long as the attackers do not precisely mimic the sites' traffic characteristics. The performance and design tradeoffs of the proposed packet scoring scheme in the context of a stand-alone implementation is studied. By exploiting the measurement/scorebook generation process, an attacker may try to mislead PacketScore by changing the attack types and/or intensities. We can easily overcome such an attempt by using a smaller measurement period to track the attack traffic pattern more closely. We are currently investigating the generalized implementation of PacketScore for core networks.

In order to prove this technique on the field, we have developed a filter generator that creates filters from an external protocol database and user-specified rules. Filter DFAs can be used as they are by existing hardware or software engines or translated into C code by the back end.We also developed an *ad hoc* DFA execution engine that adapts its operations to the word size of the underlying machine instead of processing a byte at a time and enforces memory safety and termination through run-time fully aggregated bound checks.

The run-time performance and memory occupation of SPAF filters have been evaluated both in synthetic and real-world benchmarks. Test results show that FSA-based filters perform on a similar or improved level as other modern approaches such as BPF+, both on simple and complex filters; SPAF filters are also shown to scale better with increasing numbers of filtering rules. The measured overhead of run-time safety checks is small and does not cause any significant penalties both in times of run-times (few checks are executed per packet) and memory occupation (few checks are inserted per filter). Overall, the SPAF approach is an effective and simple way to generate packet filters that are easy to compose and efficient to run, even with increasing complexity.Among the potential problems, a widely known issue affecting specifically DFAs is an explosion occurring in the state space when treating certain critical patterns; this problem is the limiting factor for DFA adoption in other pattern-based detectors such as intrusion detection systems

The SPAF approach can be easily extended to perform packet demultiplexing in addition to packet filtering. This is partial supported by our current generator by labeling final states with identifiers of the matching filtering rules; full support would require dynamic automata creation and code generation, tasks that will be the object of future studies. Another future extension to SPAF could be enabling interactions (e.g., look-ups and updates) with stateful constructs such as session tables, useful for higher-layer filtering and traffic classification. In conclusion, SPAF has been shown as an approach that improves the state of the art by generating packet filters that combine most of the desired properties in terms of processing speed, memory consumption, flexibility and simplicity in specifying protocol formats and filtering rules, effective filter composition, and low run-time overhead for safety enforcement. The development of the filter generator and the test results support the viability of our claims.

REFERENCES

[1] S. McCanne and V. Jacobson, "The BSD packet filter: A new architecture for user-level packet capture," in *Proc. USENIX*, 1993, p. 2.

[2] A. Begel, S. McCanne, and S. L. Graham, "BPF□: Exploiting global data-flow optimization in a generalized packet filter architecture," *SIGCOMM Comput. Commun. Rev.*, vol. 29, no. 4, pp. 123–134, 1999.

[3] M. L. Bailey, B. Gopal, M. A. Pagels, L. L. Peterson, and P. Sarkar, "PathFinder: A pattern-based packet classifier," in *Proc. Oper. Syst. Design Implement.*, 1994, pp. 115–123.

[4] O. Morandi, F. Risso, M. Baldi, and A. Baldini, "Enabling flexible packet filtering through dynamic code generation," in *Proc. IEEE ICC*, May 2008, pp. 5849–5856.

[5] L. Degioanni, M. Baldi, F. Risso, and G. Varenni, "Profiling and optimization of software-based network-analysis applications," in *Proc. 15th Symp. Comput. Arch. High Perform. Comput.*, Washington, DC, 2003, p. 226.

[6] S. Ioannidis and K. G. Anagnostakis, "xPF: Packet filtering for lowcost network monitoring," in *Proc. HPSR*, 2002, pp. 121–126.

[7] F. Risso and M. Baldi, "NetPDL: An extensible XML-based language for packet header description," *Comput. Netw.*, vol. 50, no. 5, pp. 688–706, 2006.

[8] L. Degioanni, M. Baldi, D. Buffa, F. Risso, F. Stirano, and G. Varenni, "Network virtual machine (NetVM): A new architecture for

efficient and portable packet processing applications," in *Proc. 8th Int. Conf. Telecommun.*, Jun. 15–17, 2005, vol. 1, pp. 163–168.

[9] D. R. Engler and M. F. Kaashoek, "DPF: Fast, flexible message demultiplexing using dynamic code generation," in *Proc. ACM SIGCOMM*, New York, 1996, pp. 53–59.

[10] M. Jayaram, R. Cytron, D. Schmidt, and G.Varghese, "Efficient demultiplexing of network packets by automatic parsing," in *Proc. Workshop Compiler Support Syst. Softw.*, 1996.

[11] S. Kumar, J. Turner, and J. Williams, "Advanced algorithms for fast and scalable deep packet inspection," in *Proc. ACM ANCS*, New York, 2006, pp. 81–92.

[12] S. Kumar, S. Dharmapurikar, F. Yu, P. Crowley, and J. Turner, "Algorithms to accelerate multiple regular expressions matching for deep packet inspection," in *Proc ACM SIGCOMM*, New York, 2006, pp.339–350.

[13] M. Becchi and P. Crowley, "An improved algorithm to accelerate regularexpression evaluation," in *Proc. ACM ANCS*, New York, 2007, pp145–154.

[14] F. Yu, Z. Chen, Y. Diao, T. V. Lakshman, and R. H. Katz, "Fast and memory-efficient regular expression matching for deep packet inspection," in *Proc. ACM ANCS*, New York, 2006, pp. 93–102.

[15] M. Becchi and P. Crowley, "A hybrid finite automaton for practical deep packet inspection," in *Proc. ACM CoNEXT*, New York, 2007, pp. 1–12.

[16] R. Smith, C. Estan, and S. Jha, "XFA: Faster signature matching with extended automata," in *Proc. IEEE Symp. Security Privacy*, 2008, pp.187–201.

[17] M. Becchi, M. Franklin, and P. Crowley, "A workload for evaluating deep packet inspection architectures," in *Proc. IEEE Int. Symp. Workload Characterization*, Sep. 2008, pp. 79–89.

[18] T. Hruby, K. van Reeuwijk, and H. Bos, "Ruler: High-speed packet matching and rewriting on NPUs," in *Proc. ACM ANCS*, New York, 2007, pp. 1–10.

[19] H. Bos, W. D. Bruijn, M. Cristea, T. Nguyen, and G. Portokalidis, "FFPF: Fairly fast packet filters," in *Proc. OSDI*, 2004, pp. 347–363.

[20] L. Deri, "nCap: Wire-speed packet capture and transmission," in *Proc. IEEE E2EMON*, Washington, DC, 2005, pp. 47–55.

[21] Z. Wu, M. Xie, and H. Wang, "Swift: A fast dynamic packet filter," in *Proc. 5th USENIX Symp. Netw. Syst. Design Implement.*, Berkeley,CA, 2008, pp. 279–292.

◈◈◈

Iris Feature Extraction Using Wavelet Maxima Components for Biometric Identification Systems

Madhuri Sachane, V.M. Jain

P.G student (E&TC, COEA), Professor (E&TC, COEA), Ambejogai,India
E-mail : msachane@gmail.com,jainvarsh@gmail.com

Abstract - A biometric system provides automatic identification of an individual based on a unique feature or characteristic possessed by the individual. Iris recognition is regarded as the most reliable and accurate biometric identification system available. There is strong scientific demand for the proliferation of the systems, concepts and algorithms for iris recognition and identification. Most commercial iris recognition systems use patented algorithms developed by Daugman, and these algorithms are able to produce perfect recognition rates. Especially it focuses on image segmentation and feature extraction for iris recognition process. The performance of iris recognition system highly depends on edge detection. The Canny Edge Detector is one of the most commonly used images processing tools, detecting edges in a very robust manner. For instance, even an effective feature extraction method would not be able to obtain useful information from an iris image that is not segmented properly. We used a fusion mechanism that amalgamates both, a Canny Edge Detection scheme and a Circular Hough Transform, to detect the iris' boundaries & edges in the eye's digital image in robust manner. Experiments are performed using iris images obtained from CASIA database (Institute of Automation, Chinese Academy of Sciences) and Matlab application for its easy and efficient tools in image manipulation.

Keywords- *Iris recognition, segmentation, image processing, canny edge detection.*

I. INTRODUCTION

A. Overview

Biometrics deals with automated methods of recognizing individuals based on the features derived from their Physiological and behavioral characteristics. A higher degree of confidence can be achieved by using unique physical or behavioral characteristics to identify a person; this is biometrics. A physiological characteristic is relatively stable physical characteristics such as face, fingerprints, palm prints, iris, and hand geometry. Biometric authentication technique based on iris patterns is suitable for high level security systems. Iris recognition systems, in particular, are gaining interest because the iris's rich texture offers a strong biometric cue for recognizing individuals. Applications of these systems include computer systems security, e-banking, credit card, access to buildings in a secure way. The automated personal identity Authentication systems based on iris recognition are reputed to be the most reliable among all biometric methods: we consider that the probability of finding two people with identical iris pattern is almost zero. The uniqueness of iris is such that even the left & right eye of the same individual is very different [1] [2]. That's why iris recognition technology is becoming an important biometric solution for people identification. The uniqueness of every iris parallels the uniqueness of fingerprint, but the iris enjoys further practical advantages over fingerprint & other biometrics for automatic recognition.

Iris is protected from the external environment behind the cornea & eyelid. The iris consists of a number of layers; the lowest is the epithelium layer, which contains dense pigmentation cells. The stromal layer lies above the epithelium layer, and contains blood vessels, pigment cells and the two iris muscles. The externally visible surface of the multi-layered iris contains two zones, which often differ in color. An outer ciliary zone and an inner pupillary zone, and these two zones are divided by the collarette – which appears as a zigzag pattern. The two zones typically have different textural details. A front-on view of the iris is shown in Figure 1. No effects of aging, the small scale radial features of the iris remain stable & fixed from about one year of age throughout life.

B. Outline

This paper, presents an iris recognition system by composing the following four steps. Firstly, an image containing the user's eye is captured by the system. Second step consists of preprocessing the acquired image. Image preprocessing mainly involves iris localization, segmentation, normalization, and image enhancement. Once the preprocessing step is achieved, it is necessary to detect the images [3]. After that, we can extract texture of the iris. Finally, we compare the coded image with the already coded iris in order to find a match.

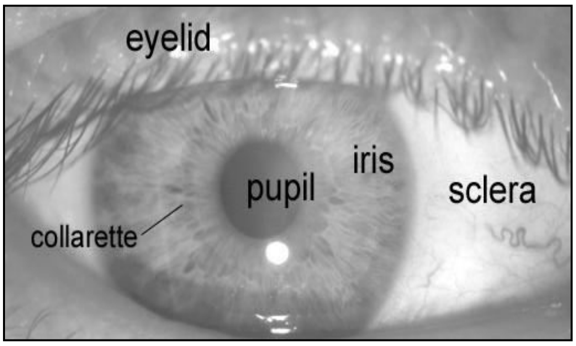


Figure 1 – A front-on view of the human eye.

II. IMAGE PREPROCESSING

An iris image as shown in Figure 1 contains not only the region of interest (iris) but also some ‘unuseful’ parts (e.g. eyelid, pupil etc.). Since it has a Circular shape when the iris is orthogonal to the sensor, iris recognition algorithms typically convert the pixels of the iris to polar coordinates for further processing. An important part of this type of algorithm is to determine which pixels are actually on the iris, effectively removing those pixels that represent the pupil, eyelids and eyelashes, as well as those pixels that are the result of reflections [4]. In this algorithm, the locations of the pupil and upper and lower eyelids are determined first using edge detection. This is performed after the original iris image has been down sampled by a factor of two in each direction. The best edge results came using the canny method [5].The pupil clearly stands out as a circle and the upper and lower eyelid areas above and below the pupil is also prominent. In addition, a change in the camera-to-eye distance may result in the possible variation in the size of the same iris. Furthermore, the brightness is not uniformly distributed because of non-uniform illumination. Before extracting features from the original eye image, the image need to be preprocessed to localize iris, normalize iris, and reduce the influence of the factor mentioned above.

A. Localization

For the preprocessing step i.e., inner and outer boundaries of the iris are located. Integro-differential operators are then used to detect the centre and diameter of the iris, then the pupil is also detected using the differential operators, for conversion from Cartesian to polar transform, rectangular representation of the required area is made [6]. The first preprocessing step consists in locating the inner & outer boundaries of the iris [8]. The first method proposed to localize the iris was proposed by John Daugman who is considered the father of iris recognition technology because his system was the first developed and implemented.. In the Daugman’s system, Integro-differential operators are used to detect the center & diameter of the iris and pupil respectively. These operators exploit both the circular geometry of the iris or the pupil. Indeed they behave as a circular edge detector since the sclera is always lighter than the iris, and pupil generally darker than the iris for healthy eye.

$$max_{(r,x_0y_0)} \left| G_\sigma(r) * \frac{\partial}{\partial x} \int_{r,x_0y_0} \frac{I(x,y)}{2\pi r} ds \right| \qquad (1)$$

Where I(x, y) is the eye image, *r* is the radius to search for, Gσ*(r)* is a Gaussian smoothing function, and *s* is the contour of the circle given by *r*, *x0, y0*. This operator iteratively takes an x,y coordinate and makes circles of various radii centered at that coordinate and sums up (integrates) the intensity values in the circular contour of radius r. It then moves to the next radius and integrates that contour, and the derivative is taken between the changes of the intensity of the two radii, and so on. The circle that is found to have the maximum rate of change between circular contours is then determined to be the circle defining the iris-sclera boundary. This process is then run again with a higher sensitivity just inside the circle containing the iris to find the boundary between the iris and pupil. The sensitivity of the operator is set by the σ factor of the Gaussian function.

B. Segmentation

Segmentation is the important phase in the Iris Recognition process. Segmentation refers to the process of partitioning a digital image into multiple regions (sets of pixels). The main objective of segmentation is to remove nonuseful information, namely the pupil segment and the part outside the iris (sclera, eyelids, skin). 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 the objects and boundaries (lines, curves etc).The segmentation stage is critical to the success of an iris recognition system, since data that is falsely represented as iris pattern data will corrupt the biometric templates generated, resulting in poor recognition rates. The result of image segmentation is a set of regions that collectively cover the entire image or a set of contours extracted from the image. Iris segmentation method is providing more efficient way of locating the Iris boundaries. We applied canny edge detector algorithm. By using this detector, we can easily see the gradient value. If global threshold value is used on that gradient image, the gradient values along potential edge will be lost. In order to avoid that effect we can apply local threshold in the area of interest [9]. The goal of texture segmentation is to partition an image into homogeneous regions and identify the boundaries which separate regions of different textures.

C. Canny Edge Detection

Canny edge detection is the most commonly used image processing tool, detecting edges in very robust manner. The Canny edge detection algorithm is known to many as the optimal edge detector. Canny's intentions were to enhance the many edge detectors already out at the time he started his work. He was very successful in achieving his goal and his ideas and methods can be found in his paper, "*A Computational Approach to Edge Detection*"[10]. In his paper, he followed a list of criteria to improve current methods of edge detection. The first and most obvious is low error rate. It is important that edges occuring in images should not be missed and that there be NO responses to non-edges. The second criterion is that the edge points be well localized. In other words, the distance between the edge pixels as found by the detector and the actual edge is to be at a minimum. A third criterion is to have only one response to a single edge.

Based on these criteria, the canny edge detector first smoothes the image to eliminate and noise. It then finds the image gradient to highlight regions with high spatial derivatives. The algorithm then tracks along these regions and suppresses any pixel that is not at the maximum (nonmaximum suppression). The gradient array is now further reduced by hysteresis. Hysteresis is used to track along the remaining pixels that have not been suppressed. Hysteresis uses two thresholds and if the magnitude is below the first threshold, it is set to zero (made a nonedge). If the magnitude is above the high threshold, it is made an edge. And if the magnitude is between the 2 thresholds, then it is set to zero unless there is a path from this pixel to a pixel with a gradient above T2. The canny operator is optimum even for noisy images as the method bridge the gap between strong & weak egdes of the image by connecting the weak edges in the output only if they are connected to strong edges. Hence the edges are more likely to be the actual ones. Therefore compared to other edge detection methods, this canny operator is less fooled by spurious noise.

D. Hough Transform

The Hough transform is a standard computer vision algorithm that can be used to determine the parameters of simple geometric objects, such as lines and circles, present in an image. The circular Hough transform can be employed to deduce the radius and centre coordinates of the pupil and iris regions. The Hough Transform is considered as the very powerful tool in edge linking for line extraction [11]. The main advantage of the Hough transform technique is that it is tolerant of gaps in feature boundary descriptions and is relatively unaffected by image noise. The standard Hough Transform proposed by duda & Hart, is widely applied for line extraction in natural scenes. The circle is simpler to represent in parameter space, compared to the line, since the parameter of the circle can be directly transfer to the parameter space. The equation of the circle is:

$$(x-a)^2+(y-b)^2=r^2 \qquad (2)$$

Where a & b are the centre of the circle in the direction x & y respectively and r is the radius. Circle has three parameters: two parameters for the centre of the circle and one for the radius of the circle. For ellipse & other parametric objects the algorithm is quite similar. But the computation complexity increases (dimensions of the Hough space) with the number of the variables.

III. NORMALIZATION

Irises from different people may be captured in different size, and even for the iris from the same person, the size may change because of the variation of the illumination and other factors. Such elastic deformation in iris texture affects the result of iris matching. For the purpose of achieving more accurate recognition results, it is necessary to convert the segmented portion of the iris into a standard form, so that two pictures taken at different times and different conditions can still be recognized as from the same eye. This process accomplishes normalization step of iris recognition. Normalization takes the variable-sized circular iris image (because of pupil dilation and image resolution) and transforms it into a fixed size rectangular image to be encoded. This transformation is accomplished by a conversion from polar coordinates into rectangular coordinates, with a fixed number of pixels taken from each angle (φ). The transformation is described by Yu et al and is presented in equation:

$$I\big(x(r,\varphi),y(r,\varphi)\big)=I(r,\varphi) \qquad (3)$$

Where x(r,φ) are the linear combinations of points on the pupil-iris boundary ((φ)Xinner,(φ)Yinner), and points on the pupil-sclera boundary ((φ)Xouter,(φ)Youter).

IV. FEATURE EXTRACTION

In order to provide accurate recognition of individuals, the most discriminating information present in an iris pattern must be extracted. Only the significant features of the iris must be encoded so that comparisons between templates can be made. It must be encode mathematically into some numerical representation of the unique features of the iris. This conversion is called feature encoding. So, it is attractive to search representation methods which can capture local crucial information in an iris. The distinctive spatial characteristics of the human iris are manifest at a variety of scales [12]. To capture this range of spatial detail, it is advantageous to make use of a multi scale representation. Some works have used multi resolution techniques for iris feature extraction [13] and have proven high recognition accuracy. A Gabor filter bank has been shown to be most known multi resolution method used for iris feature extraction and Daugman in his proposed iris recognition system demonstrated the highest accuracy by using Gabor filters. However, from the point of view of texture analysis one can observe that Gabor filter based methods analyzer pretty well the texture orientations. In this paper, we have investigated the use of wavelet maxima components as a multi resolution technique alternative for iris feature extraction.

A. *Wavelet maxima for feature extraction*

Wavelet decomposition provides a very good approximation of images and natural setting for the multi-level analysis. Since wavelet transform maxima provide useful information about textures and edges analysis [20], we propose to use this technique for fast feature extraction by using the wavelet components. Wavelet maxima have been shown to work well in detecting edges which are likely the key features in a query; moreover this method provides useful information about texture features by using horizontal and vertical details.

B. *Algorithm*

As described in [19] to obtain the wavelet decomposition a pair of discrete filters H, G has been used as follows:

Table I

RESPONSE OF FILTERS H, G

H	0	0	0.125	0.375	0.375	0.125	0
G	0	0	0	-2	2	0	0

At each scale s, the algorithm decomposes the normalized iris image I(x,y) into I(x, y, s) ,

$W_v(x,y,s)$ and $W_h(x,y,s)$ as shown in figures (2,3) .

- I(x, y, s): the image smoothed at scale s.

- $W_h(x,y,s)$ and $W_v(x,y,s)$ can be viewed as the two components of the gradient vector of the analyzed image I(x,y) in the horizontal and vertical direction, respectively.

At each scale s (s=0 to s=S-1 where S is the number of scales or decomposition) image I(x, y) is smoothed by a lowpass filter:

$$I(x.y.s+1) = I(x,y,s) * (H_s, H_s) \qquad (4)$$

The horizontal and vertical details are obtained respectively by:

$$W_h(x,y,s) = \frac{1}{\lambda_s} I(x,y,s) * (G_s, D) \qquad (5)$$

$$W_v(x,y,s) = \frac{1}{\lambda_s} I(x,y,s) * (D, G_s) \qquad (6)$$

- We denote by D the Dirac filter whose impulse response is equal to 1 at 0 and 0 otherwise.

- We denote by A * (H, L) the separable convolution of the rows and columns, respectively, of image A with the 1-D filters H and L.

- G_s, H_s are the discrete filters obtained by appending $2^s - 1$ zeros between consecutive coefficients of H and G.

-λ_s, as explained in [19] due to discretization, the wavelet modulus maxima of a step edge do not have the same amplitude at all scales as they should in a continuous model. The constants λ_s compensate for this discrete effect.

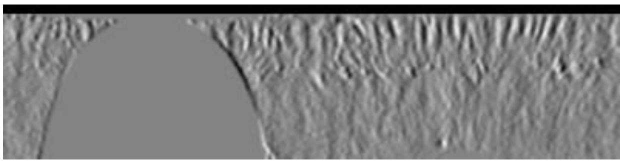


Fig.2. Wavelet maxima vertical components at scale 2 with intensities along specified column.

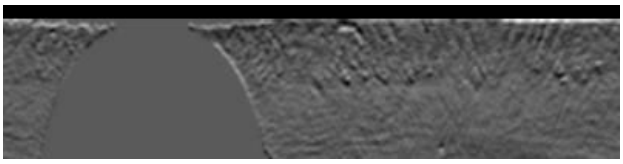


Fig.3. Wavelet maxima horizontal components at scale 2 with intensities along specified column.

In this context, we have analyzed iris textures in both horizontal and vertical directions especially that the iris has a rich structure with a very complex textures so that it makes sense to analyze the iris texture by combining all information extracted from iris region by Considering all orientations in terms of horizontal and vertical details.

V. MATCHING

Several matching algorithms were described in previous research. The most accessible of these is the Hamming distance, which depends upon the XOR operator. As feature encoding outputs a binary matrix, in order to compare two individuals, one can merely perform the XOR operator over the entire matrix. The Hamming distance is then applied in which the number of nonzero entries in the resulting matrix is summed and the result is divided by the number of overall entries in the matrix. The two iris code templates are compared by computing the hamming distance between thewm using equation [14].

$$HD = \frac{1}{N}\left[\sum_{i=1}^{N} X_j (XOR)\, Y_j\right] \qquad (7)$$

Where, Xj and Yj are the two iris codes (bit patterns), and N is the number of bits in each template. The Hamming distance gives a measure of how many bits are the same between two bit patterns. Using the Hamming distance of two bit patterns, a decision can be made as to whether the two patterns were generated from different irises or from the same one. The Hamming Distance is a fractional measure of the number of bits disagreeing between two binary patterns The Hamming distance approach is a matching metric employed by Daugman for comparing two bit patterns and it represents the number of bits that are different in the two patterns. And hence Hamming Distance matching classifier is chosen as it is more reasonable [15] compared with Weighted Euclidean Distance and Normalized correlation matching classifiers, as it is fast and simple.

VI. CONCLUSION

We describe in this paper efficient techniques for iris recognition system with high performance. The iris recognition system is tested using CASIA image database. The segmentation is the crucial stage in iris recognition. We have used the global threshold value for segmentation. In the above algorithm we have not considered the eyelid and eyelashes artifacts, which degrade the performance of iris recognition system. We have presented a novel method for iris recognition. Further development of this method is under way and the results will be reported in the near future. Judging by the clear distinctiveness of the iris patterns we can expect iris recognition system to become the leading technology in identity verification.

REFERENCES

[1] J. G. Daugman, “*Complete discrete 2-D Gabor transforms by neural network for image analysis and Compression*,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 36, pp. 1169–1179, 1988.

[2] L. Ma, Y.Wang.T. Tan. “*Iris recognition using circular symmetric filters*.” National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 2002

[3] Tisse C.L.;Martin L.;Torres L.;Robert M.,”*Person Identification Technique Using Human Iris Recognition*”,St Journal of System Reearch, Vol.4,pp.67_75,2003.

[4] Daugman, J,”*High Confidence Visual Recognition of Persons by a Test of Statistical Independence*, “IEEE Transactions on pattern analysis and Machine intelligence, vol. 15, no. 11, November 2, June 2001, pp. 1148-1161.

[5] Gonzalez,R.C., Woods,R.E,Digital Image Processing, 2nd ed., Prentice Hall (2002)

[6] Lim, S.,Lee, K., Byeon, O., Kim, T, “*Efficient Iris Recognition through Improvement of Feature Vector and Classifier*”, ETRJ Journal, Volume 23, Number 2, June 2001, pp. 61-70.

[7] Bowyer K.W., Kranenburg C., Dougherty S. \”*Edge Detector Evaluation Us-ing Empirical ROC Curves*" IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 354-359, 1999.

[8] Canny J.F., \”*A computational approach to edge detection*", IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI), Vol. 8(6), pp. 769{798, 1986.

[9] Devernay F., \”*A Non-Maxima Suppression Method for Edge Detection with Sub- Pixel Accuracy*", Research report 2724, INRIA Sophia-Antipolis, 1995.

[10] Deriche R., \”*Using canny's criteria to derive a recursively implemented optimal edge detector*", International Journal of Computer Vision (IJCV), Vol. 1(2), pp. 167{187, 1987.

[11] Heath M., Sarkar S., Sanocki T., Bowyer K.W. \”*A Robust Visual Method for Assessing the Relative Performance of Edge Detection Algorithms*", IEEE Trans- actions on Pattern Analysis and Machine Intelligence (TPAMI),Vol.19 (12), pp. 1338-1359, 1997.

[12] Jain R., Kasturi R., and Schunk B.G., *Machine Vision*, McGraw-Hill, 1995.

[13] Marr D., Hildreth E., "*Theory of Edge Detection*", Proceedings of Royal Society Of London, Vol. 207, pp. 187{217, 1980.

[14] Shin M., Goldgof D., Bowyer K.W., \”*An Objective Comparison Methodology of Edge Detection Algorithms for Structure from Motion Task*", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 190-195, 1998.

[15] Y.Zhu, T. Tan, and Y. Wang, *“Biometric Personal Identification Based on Iris Patterns*”, International Conference on Pattern Recognition (ICPR’00)-Volume2.p.2801, Sept. 2000.

[16] S. R. Ganorkar, Dr. A. A. Ghatol, "*Iris recognition: An emerging biometric technology*", Proceedings of international Conference ICSCI 2007, Hyderabad, pp. 596-600, January 03-07, 2007.

[17] Proceedings of the 6th WSEAS International Conference on Signal Processing, Robotics and Automation, Corfu Island, Greece, February 16-19, 2007 96

[18] Makram Nabti , Ahmed Bouridane,\" Wavelet Maxima And Moment Invariants Based Iris Feature Extraction", 1-4244-1437-7/07/$20.00 ©2007 IEEE II - 397 ICIP 2007.

[19] S. Mallat, W. Hwang. "*Singularity Detection and Processing with Wavelets*", IEEE Trans. Information Theory, volume.38, no. 2, pp. 617-643, 1992.

[20] S. Mallat. "A *Wavelet Tour of Signal Processing*", Academic Press, Second Edition, 1998.

◈◈◈

Determining the Positioning Algorithm for Fingerprinting Using WLAN

Indu Maurya , Vivek Singh Sengar & Priyanka Maurya

Department of CSE(GCET,GR.NOIDA)
E.mail: indumaurya42@gmail.com , sengar.vivek@gmail.com &: priyanka28maurya@gmail.com

***Abstract* -** The effectiveness of Location Based Systems depends on the correct location of users and mobile devices , the outdoor location can be easily calculated, while using technologies such as GPS (Global Positioning System), it is more difficult to obtain when the location scenario is an indoor environment. Several technologies and location techniques can be used in this field. One of these techniques is FINGERPRINTING which consists in two different phases:
The first phase is the: CALIBRATION PHASE: when data is collected and the Fingerprint Map is generated.
The second phase is the :
ON-LINE PHASE: where data collected by the mobile device and the data collected in the calibration phase are used to estimate the location of the mobile node.
From the several wireless communications technologies IEEE802.11 is probably the most used in Wireless Local Area Networks.

KEYWORDS----- *Existing WLAN location method, IEEE802.11b, Introduction, Location determination technique, Proposed algorithm, Triangulation and KNN approach*.

I. INTRODUCTION

All the Location Based Services (LBS) depends on the correct estimation of the users' location. While in outdoor environments technologies such as GPS (Global Positioning System) can be successfully used, the same is not true when the operating scenario is indoor environments. In such scenarios alternative location technologies and methodologies must therefore be used, making this a very challenging research area where in the last years several different types of solutions have been developed.

Some of the most used technologies used for indoor location include the use of infra-red, ultrasonic waves Pressure sensors, RFI (Radiofrequency Identification) and wireless communications networks.

In what concerns to the methodologies used to obtain the location, they can be divided into three main areas:

Triangulation, Proximity, Scene Analysis: Here, we focused on a particular location technique, which uses wireless communications networks as location technology, methodology based on scene analysis.

Location using fingerprinting:

Location Estimation Algorithms will be used, and their performance was analyzed. The following LEA were considered to do this analysis:

- ➢ Nearest Neighbor – which considers the coordinates of the nearest reference as coordinates of the actual location
- ➢ k-Nearest Neighbor – which uses the average of the coordinates of the k nearest neighbors.
- ➢ Weighted k-Nearest Neighbor_– which uses a weighted average of the coordinates of the k nearest neighbors.

IEEE 802.11b:

IEEE802.11b supports data rates up to 11Mbps, much higher than what IEEE 802.11 supports, increasing the variety of application that were feasible as compared to the previous IEEE 802.11 protocol. A number of different rates were defined; these can be used in different situations. These rates are: 11 Mbps, 5.5Mbps, 2Mbps, and 1Mbps. The actual user data throughput which users experience is usually around 5Mbps, similar to the actual throughput of an IEEE 802.3 10Base-T wired local network. IEEE802.11b uses the 2.4GHz Industrial, Scientific, and Medical (ISM) frequency band, which does not require a license for the user – but does require that the manufacturers meet the requirements for wireless local area network devices for their products. IEEE 802.11b WLANs can be used as a supplement to LANs or as an independent network. Today most laptop computers have a built-in IEEE 802.11b WLAN interface in addition to a IEEE 802.3 1000 base T interface. This WLAN interface helps users to avoid having to carry and connect cables. Today, most corporate and academic sites provide (legitimate) users with WLAN connectivity.

While IEEE 802.3 (Ethernet) utilizes a carrier sense multiple access (CMSA) mechanism to control when packets are sent. Specifically, IEEE 802.3 employs Carrier Sense Multiple Access with Collision Detect (CSMA/CD). With this media access protocol all stations that wish to send messages listen for when the channel is idle. As only one station can be allowed to send at a time, the others need to wait until the channel is free. If two or more stations send messages at the same time, then a collision occurs, in this case all of the messages which were being set are lost and all of the stations attempting to transmit will stop trying to send and will wait for a random amount of time

Before listening to see if the channel is idle – at which time they will attempt to transmit their message again. Because in a wired network it is possible to determine if someone else is attempting to transmit at the same time as you are, this method works quite well. Unfortunately, this is not easy to do in the case of a WLAN interface, thus IEEE802.11b uses another technique CSMA with collision avoid (CSMA-CA). Collision avoidance is necessary because the radio transmitter cannot detect a collision (unlike the wired LAN case). To avoid collisions each interface that wishes to transmit waits for a random time after detecting that the channel is idle before attempting to transmit. This period of time can be divided into different ranges of time in order to enable a mix of higher priority traffic and lower priority traffic.

Application:

WLANs provide additional features. For example, in many historic sites, it is quite hard and expensive to run new wires or cables; additionally it is better to avoid cables in order to protect the historic building’s appearance, construction, and so on. The most important feature of a wireless network is their flexibility. In many indoor scenarios, the network configuration must be changed frequently, which for a wired network would require an expensive (re-)deployment; where as in a wireless network there is little wiring (as only the access points are attached to the LAN) significantly reducing the installation cost. Furthermore, WLANs also support other desirable features, such as roaming. Roaming enables users to move between APs; while retaining the ability to communicate, despite attaching to a new subnet via a new AP.

II. LOCATION DETERMINATION TECHNIQUE

3.1 LOCATION USING WIRELESS NETWORKS

Location using wireless networks is based on the properties of wireless signals. Any property of a wireless signal can be used in location systems, as long as there is a relation between it and the current location of the mobile terminal. The signal properties that usually are used in location systems are:-

Time-of-Flight – the time needed by the information to travel from the transmitter to the receiver.

Received Signal Strength (RSS) – which indicates the power received by the wireless networks.

Technique comes under location using wireless network

3.1.1. Location using Triangulation

Triangulation uses the geometric properties of triangles to determine the location of the mobile node.

It can be divided into Lateration and Angulation.

1. Lateration uses the distances to determine the location. It takes into account the distances between the mobile node to be located and the references.
2. Angulation the angle of incidence of a signal must be known. By analyzing the Angle of Arrival of a wireless signal relatively to a given reference, it is possible to determine the location of the mobile node.

3.1.2 Localization using Proximity

Location using this methodology consists in discovering the nearest reference to the mobile terminal; therefore its spatial resolution is dependent on the number of used references.

3.2. LOCATION USING FINGERPRINTING

Fingerprinting is a scene analysis technique. In scene analysis a scene is observed and its patterns and variations along the time are observed. The information about a scene in the case of fingerprinting is obtained from one or more properties of electromagnetic signals from the references.

IV. EXISTING WLAN LOCATION METHOD

4.1. Cell-ID

In this method the serving cell identifier (cell-ID) is used to locate the user. The accuracy in this method depends upon the radius of the cell. For urban areas, e.g.

in a large city, this may be a few hundred meters; in rural areas it could be up to 30km.

4.2. Cell-ID and RxPowerLevels

This information is used to locate the mobile subscriber with good accuracy and high speed. The mobile terminal gathers information concerning the serving cell and the power level received from it. Along with the same information about other cells in the locality, this data is passed back to a server within the network operator's network. The network server then calculates the position of the user based on the positions of the cell base stations and the power at which they are transmitting.

4.3. Global Positioning System (GPS)

The GPS positioning method measures the distance from the satellites to the receiver by determining the pseudo ranges (code phases). The system extracts the time of arrival of the signal from the contents of the satellite transmitted message. It then computes the position of the satellites by evaluating the ephemeris data at the indicated time of arrival. Finally it is possible to calculate the position of the receiving antenna and the clock bias of the receiver by using this information.

4.4. Angle of Arrival (AOA)

This requires a minimum of two base stations with directional antenna. It measure the angle of arrival of signals, coming from a particular mobile subscriber, at the two base stations, and from this can calculate the users position.

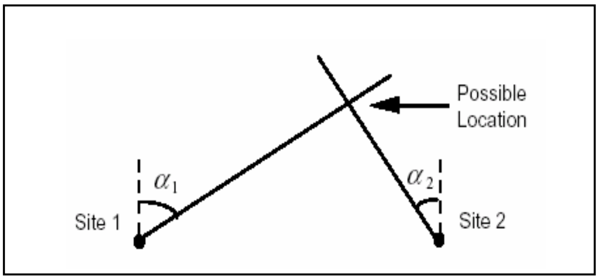


fig1: AOA

4.5. Time of Arrival (TOA)

The Time of Arrival method locates the mobile terminal by triangulation from a minimum of three base stations. Because the speed of electromagnetic waves is known, it is possible to calculate the distance from each base station by observing the time taken to arrive. This method assumes that all transmitters and receivers are perfectly synchronized and ignores reflections or interference that will affect the position accuracy.

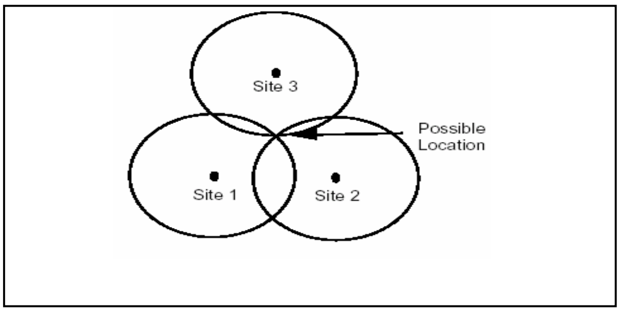


fig2: TOA

4.6. Time difference of arrival (TDOA)

Time difference of arrival (TDOA) is an algorithm based on TOA, which determines the position by measuring the time difference of signal arrival. This significantly decreases the requirement for time synchronization. This technique is used in a wide range of applications ranging from wireless communication to electronic warfare. Receivers are located at known fixed positions; the transmitter's position can then be determined by a hyperbolic function.

4.7. Received signal strength (RSS)

The power density of an electromagnetic wave is proportional to transmitted power and inversely proportional to the square of the distance to the source. This physical law as well as the vectorial combination of waves that reach a receiver over different paths is the basis for estimating distance and location from signal strength measurements.

4.8. Time of flight (TOF)

The distance between a transmitter and receiver equals the time of flight, or electromagnetic propagation time, of the transmitted signal times the speed of propagation, which is the speed of light. Distance can be determined from measurement of time of arrival (TOA) of a signal at a receiver when transmission time is known, or from differences of reception time at different locations (time difference of arrival—TDOA).Another expression of time of flight is the phase of the received signal, which may be observing the time taken to arrive. This method assumes that all transmitters and receivers are perfectly synchronized and ignores reflections or interference that will affect the position accuracy.

V. PROPOSED METHODOLOGY

A dynamic KNN algorithm is proposed, which uses triangulation and a KNN (k nearest neighbor) as a mathematical model to estimate location and a positioning algorithm for location estimation is designed by merging the TRIANGULATION and KNN approach to estimate the location of mobile devices and to improve accuracy.

7.1 LOCATION ESTIMATION USING TRIANGULATION: The transmitted signal (Tx) and the received signal (Rx) are two of the most important parameters used for location prediction of wireless node. Tx is used to calculate Available Signal Strength (ASS) and Rx is used to calculate Receive Signal Strength (RSS). These ASS and RSS are used to calculate the distance between the sending node and the receiving node. If there are three or more access point in a room or in an area then it is possible to build a triangulation positioning technique. Signal level drops when the distance between the antenna and the mobile device increases. Under ideal conditions the contours of signal level around the antenna are circles. If we know the relation between signal level and distance, from the signal level measured we can get the distance from the mobile device to the antenna. The mobile device is on the circle around the antenna with the distance as semi-diameter. Once the signal levels of the mobile device from three antennas are measured, using triangulation we can estimate the location of the mobile device.

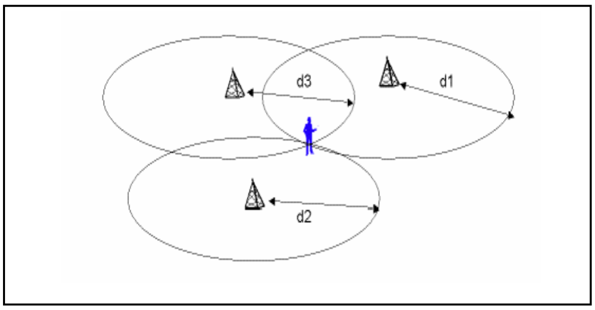


fig3: Location estimation using triangulation

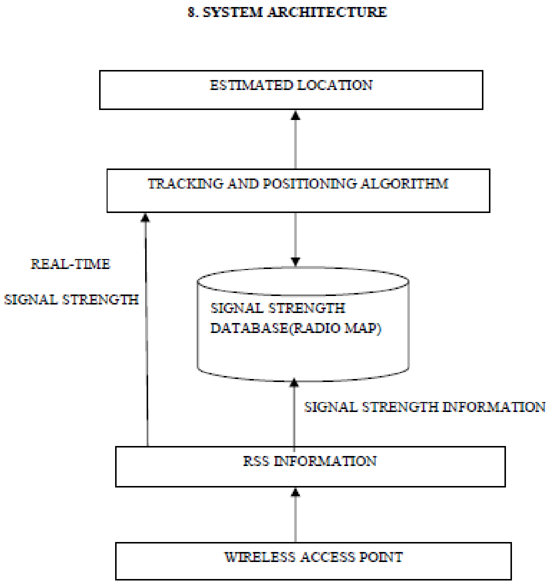
7.2 LOCATION ESTIMATION USING KNN ALGORITHM: In the KNN algorithm the mobile device measures the signal strength of each of the access point within range, then searches through the radio map database to determine the signal strength tuple that best matches the signal strength, it has measured. The system estimates the location associated with the best matching signal strength tuple (i.e. nearest neighbour) to be the location of mobile. We use the following Euclidean distance to measure----

$$Sd = \sqrt{\Sigma}\,(si- si')^2$$

Using this measure, we investigate the signal space to find the position in the physical space. In this formula---

Si is signal matches in the database for APi and Si' is the measured signal strength in the real-time operation for APi.

This technique basically calculates the Euclidean distance(sd) in the signal space and pick the signal tuple that minimize this distance in signal space and declares the corresponding physical coordinates as its estimate of the mobile location.



System architecture: We are using three access point in our project**.** We first collect multiple values of RSS (received signal strength) from the three access point and create a signal strength database and report it to the location determination server. The location determination server matches the RSS value with the stored database (radio map) using tracking and positioning algorithm i.e. here we are using triangulation and knn approach and estimate location of mobile device by merging the Triangulation and KNN approach.

9. PROPOSED ALGORITHM (POSITIONING ALGORITHM FOR LOCATION ESTIMATION)

Step1: Begin

Step2: Input: Real-time signal strength information.

Step3: Open signal strength strength database.

Step4: Read a first record from signal strength database.

Step5: If direction of current record = direction of real-time signal information.

Step6: Then calculate Euclidean distance of current record and signal strength information, otherwise Loop: next record in the signal strength database and go to step4.

Step7: After calculating Euclidean distance compare that current Euclidean distance < minimum distance.

Step8: If this condition holds then minimum distance = current Euclidean distance, otherwise Loop: next record

```
in the signal strength database and go to step4.
Step9: estimated location = location of current record.
Step10: Close the signal strength database.
Step11: Output: estimated location.
Step12: End
```

CONCLUSION

The best candidate to determine a user location in indoor environment is by using IEEE802.11 (Wi-Fi) signals, since it is most widely available installed on most mobile devices used by users. Unfortunately, the signal strength, signal quality and noise of Wi-Fi in worst scenario, fluctuate up to 33% because of the reflection, refraction, temperature, humidity and dynamic environment etc. This makes problem in determining a user location indoor. This study present our current development on a light-weighted algorithm, which is designed to be easy, simply but robust in producing the determination of user location. We study determination of estimated location and improve accuracy. A dynamic KNN algorithm is proposed, which uses triangulation approach and KNN algorithm to estimate the location of mobile device and a positioning algorithm for location estimation is designed by merging the TRIANGULATION and KNN approach to estimate the location of mobile devices and to improve accuracy.

REFERENCES

1. [802.11WIN] IEEE 802.11 Network Adapter Design Guidelines for Windows XP. http://www.microsoft.com/whdc/device/network/802x/80211_netadapt.mspx.
2. ON INDOOR POSITION LOCATION WITH WIRELESS LANS P. Prasithsangaree1, P. Krishnamurthy1, P.K. Chrysanthis2 1 Telecommunications Program, University of Pittsburgh, Pittsburgh PA 15260, {phongsak, prashant}@mail.sis.pitt.edu 2 Department of Computer Science, University of Pittsburgh, Pittsburgh PA 15260, panos@cs.pitt.edu.
3. Y. Gu, A. Lo, and I. Niemegeers, "A survey of indoor positioning systems for wireless personal networks," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 1, pp. 13–32, 2009.
4. M. Kjaergaard, "A taxonomy for radio location fingerprinting," pp. 139– 156, 2007.
5. J. Hightower and G. Borriello, "Location systems for ubiquitous computing" *Computer*, vol. 34, no. 8, Aug. 2001.
6. Heberling, D.: Indoor positioning using WLAN, In: Conf. Proc. Wireless Congress 2005: Systems & Applications, München, 2005.
7. Ladd, A. M.; Bekris, K. E.; Rudys, A.; Kavraki, L. E.; Wallach, D. S.: Roboticsbased location sensing using wireless Ethernet, Wireless Networks, vol 11, no 1 - 2, pp. 189 - 204, 2005.
8. K.; Sanghi D.; Bhagwat P. Saha, S.; Chaudhuri. Location determination of a mobile device using ieee 802.11b access point signals. pages 1987–1992. WCNC 2003 - IEEE Wireless Communications and Networking Conference, 2003.
9. Moustafa Youssef and Ashok Agrawala. The horus wlan location determination system. In MobiSys '05: Proceedings of the 3rd international conference on Mobile systems, applications, and services, pages 205–218, New York, NY, USA, 2005.

GLRT DESIGN OF MIMO RADAR IN SPHERICALLY INVARIANT RANDOM VECTOR CLUTTER

Pradeep G B & B. Roja Reddy
Department of Telecommunication, R.V.College of Engineering, Bangalore

Abstract- **The problem of target detection is analyzed using MIMO radar in SIRV clutter. The clutter process is realized as a product of two independent random processes can be called as 'texture' and 'speckle'. First, the Generalized Likelihood Ratio Test (GLRT) is derived assuming known covariance structure and then a suitable estimation of covariance matrix based on secondary data is introduced to make the derived GLRT detector fully adaptive. Some numerical results are given, showing that the derived GLRT outperforms Gaussian clutter GLRT in spikier clutter and the adaptive loss is acceptable.**

Keywords: **Covariance matrix, Compound Gaussian Clutter, GLRT, K-distribution, SIRV.**

I. INTRODUCTION

The Multiple-Input Multiple-Output (MIMO) radar employs multiple transmitting waveforms and has the ability to jointly process signals received at multiple receiving antennas, which uses the widely separated transmitters and receivers such that the target is observed from many different aspects simultaneously, resulting in spatial diversity, which can improve radar detection performance [1-3], and support high-resolution target locations and can provide more degrees of freedom. The space-time coding (STC) has been largely investigated as a viable means to achieve spatial diversity, and thus to contrast the effect of fading [4], [5]. Upon suitably space-time encoding the transmitted waveforms, a maximum diversity can be achieved. Jian Li et al [9] have pointed out that iterative generalized likelihood ratio test (iGLRT) can provide excellent detection and estimation performance at low computations cost. The potential advantages of MIMO radars are thoroughly considered [7, 8].

Recently the generalized likelihood ratio test (GLRT) detector is shown to yield excellent performance and it is very much attractive for radar detection in the presence of correlated non-Gaussian clutter modeled as multivariate compound-Gaussian form. In many of the contributions in the existing literature, the compound-Gaussian clutter model has been well accepted not only for its suitability in formulation of the GLRT detection scheme [3] but also its consistency with the K-distribution which gives deep insight into the scattering mechanism of low-grazing angle land clutter [4] and high resolution sea clutter [5]. Various papers pointed about the non-Gaussian clutters which are suitable for low resolution radars. In high resolution radars case, the disturbance of clutter is better modeled as Compound Gaussian process which is the family of non Gaussian processes. Compound Gaussian process is the product of two components: nonnegative random variable (texture) and multivariate Gaussian random vectors (speckle)[9-11].

For high resolution MIMO radars, the design of GLRT detector against the compound-Gaussian clutter is considered. Here, the Space time coding model is well suited in case of the compound-Gaussian clutter. The power level of the clutter and RCS of the target are viewed as determinate unknown parameters and are estimated with maximum likelihood method. Since the covariance matrix of the clutter is unknown, it is impossible to admit closed-for expressions. To avoid this drawback, we first derive the GLRT based on the primary data assuming that the covariance structure is known. A suitable estimate of the covariance matrix based on the secondary data is derived and plugged into the derived detectors in place of exact covariance matrix.

The rest of the paper is organized as follows. Section 2 presents the Problem formulation and section 3 presents the GLRT detector design. The performance of the derived GLRT is analyzed in section 4 and section 5 presents conclusions.

II. PROBLEM FORMULATION

Consider a narrow band MIMO radar system with s transmitters and r receivers and assume that the antennas are with widely separated to provide uncorrelated reflection coefficients between each transmit/receive pairs of sensors, N denotes the number of pulse train for each transmit antennas, as shown in Fig.1.

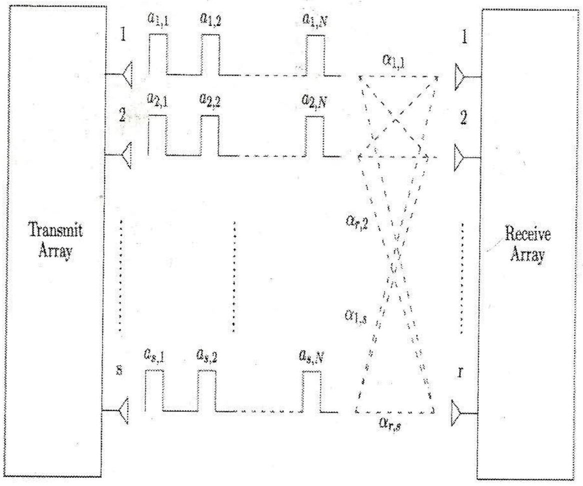


Fig. 1. Schematic representation of considered transmit/receive system.

Moreover, suppose that K (K >2N) secondary data vectors, sharing the same covariance structure of the primary data are available. Denote that r_i and, r_{ik}, i =1,.... r, k =1,..... K , are the received signal from the primary and secondary data, respectively. Then, the problem of detecting a target with MIMO radar can be formulated in terms of the following binary hypotheses test:

$$\begin{cases} H_0 \begin{cases} r_i = n_i & i = 1, \dots\dots .r \\ r_{ik} = n_{ik} & i = 1,..r, k = 1, \dots K \end{cases} \\ H_1 \begin{cases} r_i = A\alpha_i + n_i & i = 1, \dots .r \\ r_{ik} = n_{ik} & i = 1, \dots r, k = 1, \dots K \end{cases} \end{cases}$$
(1)

where, A_{NXs}=□$[a_1, \dots \dots a_s] \epsilon C^{NXs}$ is the transmit code matrix which defines s different code words of length N,

$a_l = [a_{l1}, \dots \dots . a_{lN}]^T$□□ i=1,....s are referred as the code word of the 1 th antennas and N is the length of the codeword;

$\alpha_i = [\alpha_{i1}, \dots .. \alpha_{is}]^T$, i=1,.....r are the complex values accounting for both the target backscattering and the channel propagation effects between the transmitters and receivers;

$r_i = [r_{i,1}, \dots .., r_{1,N}]^T$ *denotes* the N dimensional column vectors and are the echo signals of i th receive antennas contaminated by the clutter;

The clutter vectors n_i, i=1,....□r are assumed as compound-Gaussian random vectors, or SIRVs, i.e.,

$$n_i = \sqrt{\sigma_i} g_i \qquad \text{i=1,....r}$$

The textures ($\sigma_i, i = 1, \dots r$) are non-negative random variables and the speckle components ($g_{i,} i = 1, \dots .r$) are correlated N-dimensional complex circular Gaussian vectors and independent each other. At the design stage, we model,(σ_i, i =1,...□r) as the unknown deterministic parameters. This is tantamount to assuming independent zero-mean complex circular Gaussian vectors with covariance matrix

$$R_i = E[n_i n_i^{\dagger}] = \sigma_i R_0$$

where $R_0 = E[g_i g_i^{\dagger}]$ is the covariance structure.

According to the Neyman-Pearson criterion, the optimum solution to the hypotheses testing problem is the likelihood ratio test, but, for the case at hand, it cannot be implemented since total ignorance of the parameters $\propto_i$ is assumed. A possible way to circumvent this drawback is to resort to the GLRT which is tantamount to replacing the unknown parameters with their maximum likelihood (ML) estimates under each hypothesis.

III. GLRT DESIGN

The GLRT of MIMO radar with an unknown covariance matrix against compound-Gaussian clutter is derived here. More specifically, first assume that the clutter covariance structure is known and derives a GLRT maximizing the likelihood function of the primary data over the remaining unknown parameters. Then a suitable estimate of the unknown covariance based on the secondary data is inserted to make the detector fully adaptive.

A straightforward way to determine the threshold T given a false-alarm rate is to use Monte-Carlo simulation. The number of simulations and computation load are usually huge because of the small value of *P*FA.

$$\frac{\underset{\alpha_1,..\alpha_r,\sigma_1,...\sigma_r}{max} f(r_1,....,r_r|H_1,\alpha_1,..\alpha_r,\sigma_1,..\sigma_r)}{\underset{\sigma_1,....\sigma_r}{max} f(r_1,....r_r|H_0,\sigma_1,...\sigma_r)} \underset{H0}{\overset{H1}{\gtrless}} T \qquad (2)$$

Where $f(r_1,...r_r|H_1) = f(r_i,..r_r|H_1,\alpha_1,..\alpha_r\sigma_1,...\sigma_r)$ and $f(r_1,...r_r|H_0) = f(r_i,.....r_r|H_0,\sigma_1,...\sigma_r)$ denote the pdfs of the data under *H1* and *H0*, respectively. More specifically, they are given by

$$f(r_1,.r_r|H_0) = \frac{1}{\pi^{Nr}\prod_{i=1}^{r} det(R_i)} exp\left\{-\sum_{i=1}^{r} r_i^\dagger R_i^{-1} r_i\right\} \qquad (3)$$

Under H0 and

$$f(r_1,...r_r|H_1) = \frac{exp\{-\sum_{i=1}^{r}(r_i - A\alpha_i)^\dagger R_i^{-1}(r_i - A\alpha_i)\}}{\pi^{Nr}\prod_{i=1}^{r} \det(R_i)} \qquad (4)$$

under H1,

where det(.) denotes the determinant.

To determine the maximum likelihood estimators of□□ $\sigma_1,...,\sigma_r$ under H0, the log-likelihood function of (3) is

$$lnf(r_1,....r_r|H_0) = -Nrln\,\pi - N\sum_{i=1}^{r} ln\sigma_i - rln\det(R_0) - \sum_{i=1}^{r}\frac{r_i^\dagger R_0^{-1} r_i}{\sigma_i} \qquad (5)$$

It can be shown that (5) admits the following solution

$$\hat{\sigma}_{i0} ⍰ = \frac{r_i^\dagger R_0^{-1} r_i}{N} \qquad (6)$$

i=1,….r

As to the estimators of $\alpha_1,.....\alpha_r$ and $\sigma_1,...,\sigma_r$ under H1, the log-likelihood function of (4) is

$$\ln f(r_1,......r_r|H_1) = Nrln\,\pi - N\sum_{i=1}^{r} ln\sigma_i - rln\det(R_0) - \sum_{i=1}^{r}\frac{(r_i - A\alpha_i)^\dagger R_0^{-1}(r_i - A\alpha_i)}{\sigma_i} \qquad (7)$$

Thus it is easy to obtain the maximum likelihood estimate of the complex amplitude α_i as

$$\hat{\alpha}_{i1} ⍰ = (\mathbf{A}^\dagger \mathbf{R}_0^{-1}\mathbf{A})^{-1}\mathbf{A}^\dagger \mathbf{R}_0^{-1}\mathbf{r}_\mathrm{i}$$

i=1,….,r (8)

$$\hat{\sigma}_{i1} ⍰ = \frac{\mathbf{r}_\mathrm{i}^\dagger(\mathbf{R}_0^{-1} - \mathbf{R}_0^{-1}\mathbf{A}(\mathbf{A}^\dagger\mathbf{R}_0^{-1}\mathbf{A})^{-1}\mathbf{A}^\dagger\mathbf{R}_0^{-1})\mathbf{r}_\mathrm{i}}{\mathbf{N}}$$

i= 1,…,r

$$\prod_{i=1}^{r}\frac{r_i^{\dagger}R_0^{-1}r_i}{r_i^{\dagger}(R_0^{-1}-(R_0^{-1}A(A^{\dagger}R_0^{-1}A)^{-1}A^{\dagger}R_0^{-1})r_i}\underset{H0}{\overset{H1}{\gtrless}}T \qquad (9)$$

where the detection threshold T is a suitable modification of the original threshold in (2)

Adaptive detection:

In order to make the derived detectors fully adaptive, we replace the covariance matrix R_0 by a suitable estimate in the LHS of (9) based on the secondary data, which shares the same correlation properties with the cell under test and free of signal. To make the detectors ensure the CFAR property w.r.t texture statistics, a normalized sample covariance matrix is adopted[13], based on the secondary data collected by the receiver antennas, that is,

$$\hat{R}_{0i}\,\boxed{?}=\frac{\mathbf{N}}{\mathbf{K}}\sum_{\mathbf{k=1}}^{\mathbf{K}}\frac{n_{i,k}n_{i,k}^{\dagger}}{n_{i,k}^{\dagger}n_{i,k}} \qquad (10)$$

Substituting (10) in (9), we come up with the following adaptive detectors, i.e.,

$$\prod_{i=1}^{r}\frac{r_i^{\dagger}\hat{R}_{0i}^{-1}r_i}{r_i^{\dagger}(\hat{R}_0^{-1}-\hat{R}_{0i}^{-1}\mathrm{A}(\mathrm{A}^{\dagger}\hat{R}_{0i}^{-1}\mathrm{A})^{-1}\mathrm{A}^{\dagger}\hat{R}_{0i}^{-1})r_i}\underset{\boldsymbol{H0}}{\overset{\boldsymbol{H1}}{\gtrless}}\boldsymbol{T1} \qquad (11)$$

Where the detection threshold T1 s are a suitable modification of the original values in (9).

We highlight that, with given N , the proposed adaptive detector end up coincident with (9) as K diverges. However, for finite K , the performance of the estimate and, eventually, of the adaptive detector itself depends upon the actual values of N . It is thus necessary to quantify the loss of the proposed decision strategy with respect to its "non adaptive" counterpart under situations of exact covariance matrix. This is one of the objects of the next section.

IV. PERFORMANCE ASSESSMENT

To compare the performance of derived detector with the detector derived by A.De Maio in Gaussian noise, we simulate the GLRT detector derived by A.De Maio namely GC-GLRT, that is

$$\sum_{i=1}^{r}r_i^{\dagger}\,\hat{\mathrm{R}}_{0\mathrm{i}}^{-1}\mathrm{A}(\mathrm{A}^{\dagger}\,\hat{\mathrm{R}}_{0\mathrm{i}}^{-1}\mathrm{A})^{-1}\mathrm{A}^{-1}\mathrm{A}^{\dagger}\,\hat{\mathrm{R}}_{0\mathrm{i}}^{-1}r_i\underset{H0}{\overset{H1}{\gtrless}}T2 \qquad (12)$$

We assume a clutter-dominated scenario, and the clutter is sampled from K-distribution with pdf

$$f(z)=\frac{\sqrt{2v/\mu}}{\Gamma(v)}\left(\sqrt{\frac{2v}{\mu}}z\right)^{v}K_{v-1}\left(\sqrt{\frac{2v}{\mu}}z\right) \qquad (13)$$

the texture component $\sqrt{\sigma_i}$ is gamma distribution, with pdf

$$f(\sqrt{\sigma_i})=\frac{1}{\Gamma(v)}\left(\sqrt{\frac{v}{\mu}}\right)^{v}\sqrt{\sigma_i}^{\,v-1}e^{\frac{-v}{\mu\sqrt{\sigma_i}}}u(\sigma_i) \qquad (14)$$

where $\Gamma(.)$ is the Eulerian Gamma function, $\nu > 0$ is the parameter ruling the shape of the distribution, $u(.)$ denotes the unit step function, and $K_v(.)$ is the modified second kind Bessel function with order ν, which rules the clutter spikiness, namely smaller the value of ν , higher the tails of the distribution. The distribution will become Gaussian for $\nu \to \infty$.

The clutter has exponential correction structure covariance matrix R0, the (i,j) element of which is $\rho^{|i-j|}$, where ρ is the one-lag correlation coefficient and is set to 0.9 in the simulations.

Finally, the transmit code matrix A is the orthogonal space time codes, and the signal-to-clutter ratio (SCR) is defined as

$$\mathbf{SCR}=\frac{\sigma^2}{Ns}tr[A^{\dagger}R_0^{-1}A]$$

In Fig.2, we analyze the shape parameter v of the clutter that effect the detection performance. The p_d s of derived GLRT and of GC-GLRT are plotted versus SCR with $P_{fa}=10^{-4}$, N=8, r=4, s=4, ρ=0.9,K=32 for several values v. The curves show that the performance of derived GLRT is better in more spikier clutter with smaller v, however, as to GC-GLRT, the situation is reverse. It is because that the derived GLRT is devised in compound-Gaussian clutter, and the GC-GLRT is devised in Gaussian clutter, the performance is better for more matched case. More specifically, the gap in the case P_d=0.9 is about 8dB between derived GLRT and GC-GLRT for v=0.5, however the performance of GC-GLRT is better than derived GLRT in the case P_d>0.9 for v=5, and the gaps is about 0.7db. It is because that the distribution becomes nearly to Gaussian for high value of v.

The effect of the number of transmit antennas is studied in fig 3 and the P_d s are plotted versus SCR with several values of s. The curve of derived GLRT shows that the performance of s=2 is better than that of s=4 and s=6. As to the GC-GLRT, the increase in the value s can lead to a significant performance improvement.

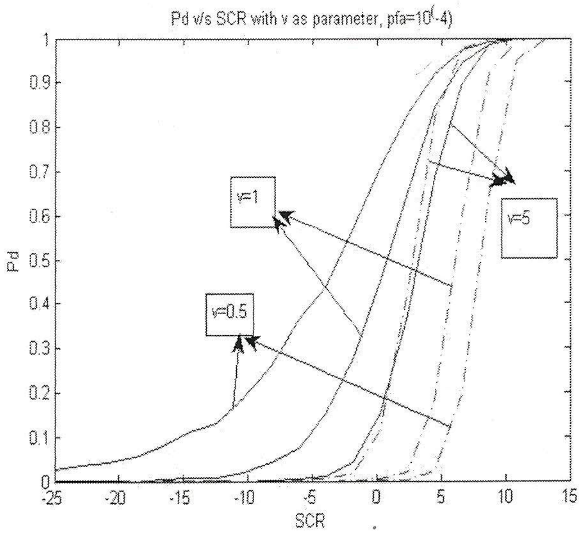


Fig 2. P_d versus SCR plots of derived GLRT(solid curves) and GC-GLRT (dashed curves) receivers, for $P_{fa}=10^{-4}$, N=8, r=4, s=4, ρ = 0.9, K=32, v as a parameter.

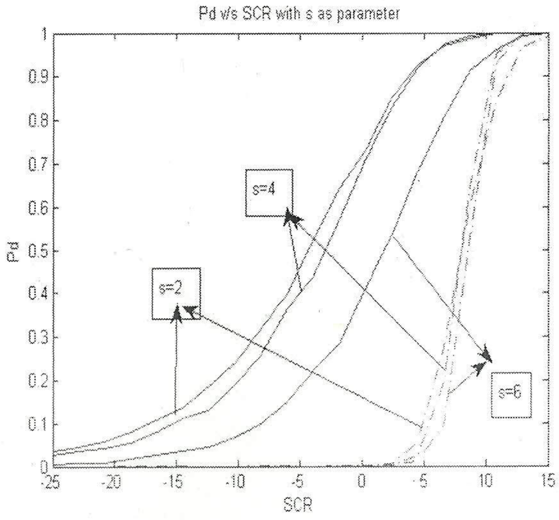


Fig 3. P_d versus SCR plots of derived GLRT (solid curves) and GC-GLRT (dashed curves) receivers, for $P_{fa}=10^{-4}$, N=8, r=4, , ρ = 0.9 , K=32, v=0.5, s as parameter.

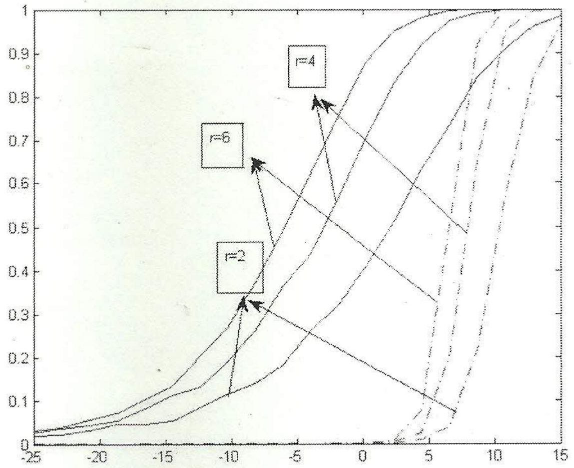


Fig 4. P_d versus SCR plots of derived GLRT (solid curves) and GC-GLRT (dashed curves) receivers, for $P_{fa}=10^{-4}$, N=8, s=4, ρ=0.9, K=32, v=0.5, r as a parameter.

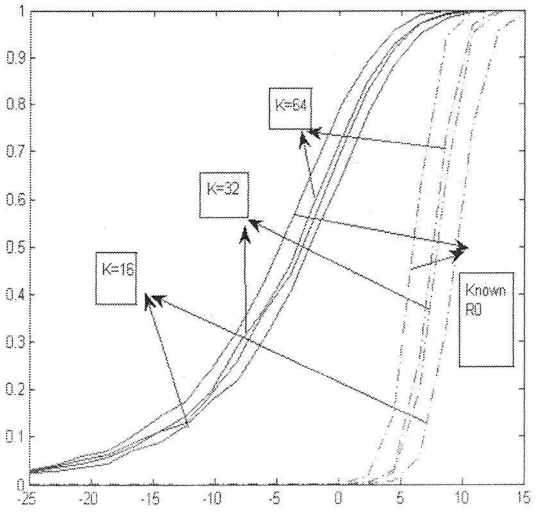


Fig 5. P_d versus SCR plots of derived GLRT (solid curves) and GC-GLRT (dashed curves) receivers, for $P_{fa}=10^{-4}$, N=8, s=4, r=4,ρ=0.9, v=0.5, K as a parameter.

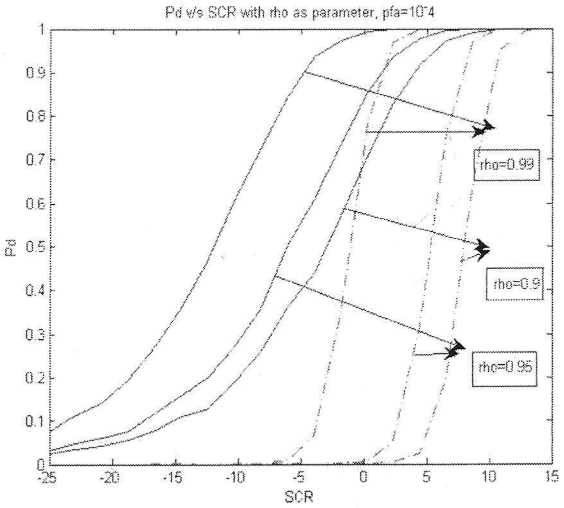
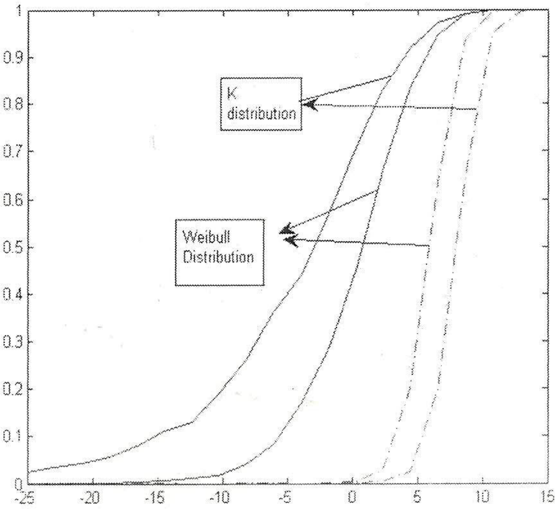


Fig 6. P_d versus SCR plots of derived GLRT (solid curves) and GC-GLRT (dashed curves) receivers, for $P_{fa}=10^{-4}$, N=8, s=4, r=4, ρ=0.9, K=32, v=0.5, rho as a parameter.



The number of receive antennas that effect the performance of detection is analyzed in fig 4, and the P_d s are plotted versus SCR with several values of r. The results show that the performance are increased steadily with increasing the number of r for both derived GLRT and GC-GLRT. Specifically, the gaps in the case P_d=0.9 are about 4dB, 6.2dB between r=2 and r=6 for derived GLRT and GC-GLRT, respectively. The gaps between the two receivers is about 6.4dB in the case r=4.

The effect of the Parameter Rho is considered in fig6 and it shows that the performance increases with increase in rho for derived detector where as in GC-GLRT, the performance is poor for the value of rho=0.9.

To make the derived GLRT fully adaptive, the estimated covariance matrix using the secondary data is inserted into which are obtained with known covariance matrix. The effect of the size K of the secondary data on the performance of derived GLRT and GC-GLRT is analyzed in fig 5. The curves show that the increase in the size K can lead to a significant performance improvement for the derived GLRT. The performance gaps in the case P_d=0.9 between K=16 and K=64 are about 0.9dB and 2.6 dB for derived GLRT and GC-GLRT, respectively. The performance with exact covariance matrix is also accessed, and the results show that the adaptive loss is acceptable.

The comparision of the performance of K-distribution with the Weibull distribution is done in fig 7 and the graph shows that the performance of K-distribution is better compared to the weibull distribution.

V. CONCLUSIONS

This paper has mainly developed the MIMO radar detection problem to compound-Gaussian case, and designed the GLRT detector.The design procedure has been adopted. It should be pointed out that the normalized sampled covariance matrix can ensure CFAR property with respect to textures, however, does not guarantee CFARness with respect to the structure of the covariance matrix.

The performance of the derived GLRT and together with GC-GLRT is studied by several numerical results. The resuts show that the derived GLRT has the better performance in spikier clutter. It has demonstrated that the loss due to the prior uncertainty as to clutter covariance result in acceptable losses, as compared to the case of exact statistics. We should point out that the performance is not increased steadily with increasing the number of transmit antennas and there should be an optimal values with given parameters.

REFERENCES

[1] A. Haimovich, R. S. Blum, L. J. Cimini, "MIMO radar with widely separated antennas," IEEE Signal Processing Magazine, pp: 116 – 129, Jan. 2008.

[2] E. Fishler, A. Haimovich, R. S. Blum, L. Cimini, D. Chizhik, and R. Valenzuela, "Spatial diversity in radars-Models and detection performance," IEEE trans. on Signal Processing, Vol. 54, No. 3, pp: 823-838, Mar. 2006.

[3] A. de Maio, M. Lops, "Design principles of MIMO radar detectors," IEEE Trans. on Aerospace and Electronic Systems, Vol. 43, No. 3, pp: 886-898, Jul. 2007.

[4]. Tarokh, V., Seshadri, N., and Calderbank, A. R. Space-time codes for high data-rate wireless communication: Performance criterion and code construction. *IEEE Transactions on Information Theory*, 44 (Mar. 1998), 744—765. 896 IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS VOL. 43, NO. 3 JULY 2007

[5]. Hochwald, B. M., and Marzetta, T. M. Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading. *IEEE Transactions on Information Theory*, 46 (Mar. 2000),543—564

[6] L. Xu, J. Li, "Iterative generalized-likelihood ratio test for MIMO radar,", IEEE Trans. on Signal Processing, Vol. 55, No. 6, pp: 2375-2385, , Jun. 2007.

[7] Fishler, E., Haimovich, A., Blum, R., Cimini, L., Chizhik, D., and Valenzuela, R.
Performance of MIMO radar systems: Advantages of angular diversity. In *Proceedings of Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, Nov. 2004, 305—309.

[8] Fishler, E., Haimovich, A., Blum, R., Cimini, L., Chizhik, D., and Valenzuela, R. Spatial diversity in radars–Models and detection performance. *IEEE Transactions on Signal Processing*, 54 (Mar. 2006),823—838.

[9] E. Conte, M. Longo, "Modelling and simulation of non-Rayleigh radar clutter," IEEE Proceedings-F, Vol. 138, No. 2, pp: 121-130, , Apr. 1991.

[10] K. J. Sangston, K. and Gerlach, "Coherent detection of radar targets in a non-Gaussian background," IEEE Trans. on Aerospace and Electronic Systems, Vol. 30, No. 2, pp: 330-340, Aug. 1994.

[11] K. Gerlach, "Spatially distributed targets detection in non-Gaussian clutter," IEEE Transactions on Aerospace and Electronic Systems, Vol. 35, No. 3, pp: 926-934, , Jul. 1999.

[12] F. Gini; M.V. Greco; L. Verrazzani; "Detection problem in mixed clutter environment as a Gaussian problem by adaptive preprocessing," Electronics Letters, Vol. 31, No. 14, pp: 1189-1190, , Jul. 1995.

[13] F. Gini; M.V. Greco; "Covariance Matrix Estimation for CFAR Detection in Correlated Heavy Tailed Clutter," Signal Processing, Vol.82, No. 12, pp: 1847-1895, , Dec 2002.

[14] Guolong cui; "2 step GLRT design of MIMO radar in compound Gaussian Clutter," 2010.

[15] Sea clutter: "Scattering, the K-distribution and Radar performance" by Keith D. Ward, Robert J.A. Tough and Simon Watts.

Predicting Medical Diseases Using Data Mining

Shweta A.Gode & G.R.Bamnote
Department of CSE, PRMITR,Badnera, Amravati (M.S)

ABSTRACT

Existing research in association with mining has focused mainly on how to expedite the search for frequently co-occurring groups of symptoms in "medical diseases" type of transactions; less attention has been paid to methods that exploit these "frequent symptoms list" for prediction purposes. This project contributes to the latter task by proposing a technique that uses partial information about the contents of a medical diseases for the prediction of what else the physician is likely to diagnose. Using the recently proposed data structure of item set trees (IT-trees), we obtain, in a computationally efficient manner, all rules whose antecedents contain at least one symptom from the incomplete disease. This project combine these rules by uncertainty processing techniques, including the classical Bayesian decision theory and a new algorithm based on the Dempster-Shafer (DS) theory of evidence combination.

Key Words

Disease, symptoms, association mining, DS theory.

I. INTRODUCTION

The primary task of association mining is to detect frequently co-occurring groups of items in transactional databases. The intention is to use this knowledge for prediction purposes: if bread, butter, and milk often appear in the same transactions, then the presence of butter and milk in a shopping cart suggests that the customer may also buy bread. More generally, knowing which items a shopping cart contains, we want to predict other items that the customer is likely to add before proceeding to the checkout counter. This paradigm can be exploited in diverse applications. For example, in the domain discussed in each "shopping cart" contained a set of hyperlinks pointing to a Web page in medical applications, the shopping cart may contain a patient's symptoms, results of lab tests, and diagnoses; in a financial domain, the cart may contain companies held in the same portfolio; and Bollmann- Sedorra et al proposed a framework that employs frequent item sets in the field of information retrieval [2]. In all these databases, prediction of unknown items can play a very important role. For instance, a patient's symptoms are rarely due to a single cause; two or more diseases usually conspire to make the person sick. Having identified one, the physician tends to focus on how to treat this single disorder, ignoring others that can meanwhile deteriorate the patient's condition. Such unintentional neglect can be prevented by subjecting the patient to all possible lab tests. However, the number of tests can undergo is limited by such practical factors as time, costs, and the patient's discomfort. A decision-support system advising a medical doctor about which other diseases may accompany the ones already diagnosed can help in the selection of the most relevant additional tests. The prediction task was mentioned early as in the pioneering association mining paper by Agrawal et al., but the problem is yet to be investigated in the depth it deserves. In our work, we wanted to make the next logical step by allowing any symptom to be treated as a class label its value is to be predicted based on the presence or absence of other symptoms. Put another way, knowing a subset of the disease symptoms, we want to "guess" (predict) the rest. In our work, we sought to solve both of these problems by developing a technique that answersuser's queries in a way that is acceptable not only in terms of accuracy, but also in terms of time and space complexity.

II. EXISTING SYSTEM

Existing research in association mining has focused mainly on how to expedite the search for frequently co-occurring groups of items in "shopping cart" type of transactions; less attention has been paid to methods that exploit these "frequent item sets" for prediction purposes.

Existing system Disadvantages:

•They didn't find missing items in frequently used item set.

•Couldn't find number of users per item set.

•Time complexity

•Lack of viewing items to the user.

III. RESEARCH AND FINDINGS

A. Data Mining

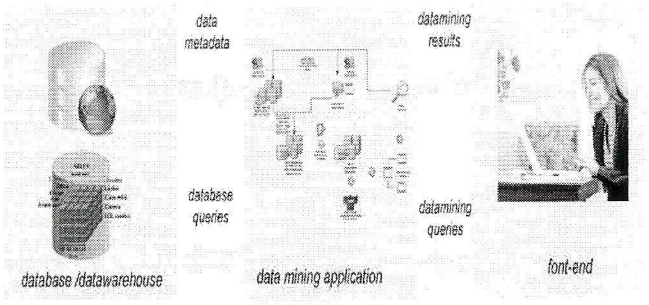


Figure 1:- Data Mining Architecture

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets.[1] These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

B. Data Mining Techniques

Five Techniques in Data Mining

- Association
- Classification
- Clustering
- Prediction
- Sequential

c. Data mining Architecture

Data mining is described as a process of discover or extracting interesting knowledge from large amounts of data stored in multiple data sources such as file systems, databases, data warehouses and etc[10]. This question leads to four possible architectures of a data mining system as follows:

- No Coupling
- Loose Coupling
- Semi Tight Coupling
- Tight Coupling

IV. PROPOSED SYSTEM ANALYSIS AND DESIGN

A. Analysis

Proposed system:

✓Finding missing symptoms using apriority algorithms in frequently used symptoms list.

✓Counting number of users per symptoms.

✓ Calculating total number of visitor's in our websites

Advantages of Proposed system:

✓ Reducing time complexity. User can easily view the symptoms.

✓Missing symptoms can easily find in the symptoms list.

B. Problem Statement

1) Let $I=\{i_1,.......i_n\}$ be a set of distinct symptoms.
2) Let a database consist of transactions $T_1,\ldots T_n$ such that $T_i \underline{c}\, I, \forall_i$
3) Let X be a group of symptoms such that $X \underline{c}\, I$
4) An association rule has the form $r^{(a)} => r^{(c)}$ where , $r^{(a)}$ and $r^{(C)}$ are the set of symptoms. The $r^{(a)}$ is the antecedent and $r^{(c)}$ is consequent.
5) The rule reads:- If all symptoms from $r^{(a)}$ are present in a transaction from $r^{(c)}$ are also present in same transaction.

6) This rule does not have absolute reliability.

7) The probabilistic confidence in rule $r^{(a)} => r^{(c)}$ Can be defined with the help of supports (relative frequencies) of antecedents and consequent as the percentage o transaction that contain $r^{(c)}$ among those transactions that contain $r^{(a)}$:

   Conf = support ($r^{(a)} \cup r^{(c)}$ / support $r^{(a)}$ ----------(1)

8) Let us assume that an association mining program has already discovered all high support set of symptoms.

9) For each such symptom ,X, any pair of subsets $r^{(a)}$ and $r^{(c)}$ such that

   $r^{(a)} \cup r^{(c)} = X$ and $r^{(a)} \cap r^{(c)} = \emptyset$

10) We can define an association rule

   r: $r^{(a)} => r^{(c)}$

11) The number of rules implied by X grows exponentially in the number of symptoms in X

12) We usually consider only high confidence rules derived from high support list of symptoms [4].

C. Detail designed

Data Flow Diagrams:

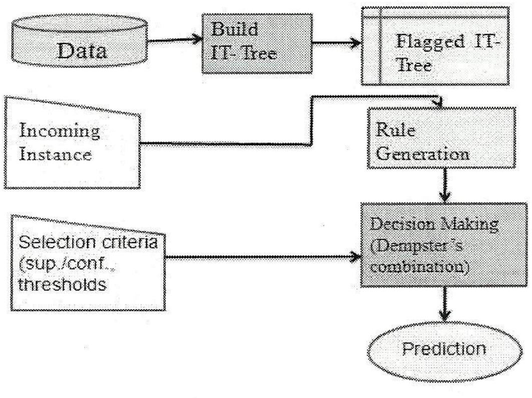

Data
Build IT- Tree
Flagged IT-Tree
Incoming Instance
Rule Generation
Selection criteria (sup./conf., thresholds
Decision Making (Dempster's combination)
Prediction

Fig.2. Data flow of our proposed system

D. The Bayesian approach

- In this approach suppose we want to establish the presence of list of symptoms, [5],[6], $\mathbf{S=\{i_1}^{(s)},......, \mathbf{i_k}^{(s)}\}$ increases the chance that item $i_j \in$ S is also present.
- Bayes rule yields

$$P(i_j | \mathbf{i_1}^{(s)},........., \mathbf{i_k}^{(s)}) = \frac{P(\mathbf{i_1}^{(s)},........., \mathbf{i_k}^{(s)} | i_j) P(i_j)}{P(\mathbf{i_1}^{(s)},..............., \mathbf{i_k}^{(s)})}$$

- We select all the symptoms for which; $P(i_j | \mathbf{i_1}^{(s)},..........., \mathbf{i_k}^{(s)}) > P(\neg i_j | \mathbf{i_1}^{(s)},..........., \mathbf{i_k}^{(s)})$ where, $P(\neg i_j | \mathbf{i_1}^{(s)},..............., \mathbf{i_k}^{(s)})$ is the probability of the symptom i_j being absent given the list of symptoms S is present.
- Since, the denominator is the same for any given list of symptoms S, it is enough if the classifier chooses the symptoms that maximizes the value of numerator.

E. The Proposed Solution

- The proposed Rule Generation Algorithm makes use of the flagged IT Tree created from the training data set.
- The algorithm takes an incoming list of symptoms as the input and returns a graph that denes association rules entitled by the incoming list of symptoms [3],[4].

Itemset Tree Construction(IT Tree)

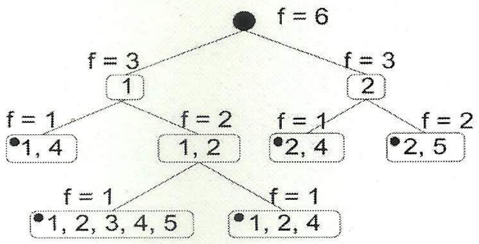


Fig. 3. The IT-tree constructed from the database .

F. Employing the DS Theory

When searching for a new way to predict the presence or absence of an item i_j in a partially observed medical disease S, we wanted to use association rules. However many rules with equal antecedents differ in their consequents- some of these consequents contain i_j , others do not. The question is how o combine the potentially conflicting evidence. One possibility is to rely on the DS theory of evidence combination. Let us now describe our technique, which we refer to by the acronym DS-ARM (Dempster-Shafer-based Association Rule Mining).

V. SYSTEM IMPLEMENTATION & TESTING

A. Modules

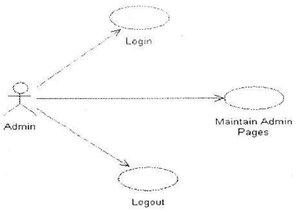
1) Admin

- Login
- Upload symptoms to the user pages

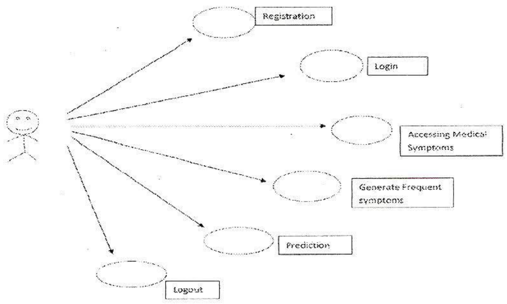
2) User

- Registration
- Login
- Access Symptoms
- Frequent symptoms Generation
- Prediction of Missing symptoms

Use Case Admin:



Use Case User:



B. Implementation details

Inputs

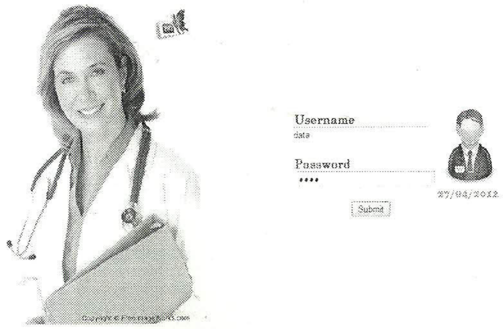
Data Set on the Web [Details of symptoms in medical disease] Data collection on client, server sides and anywhere in between Goal determine actual disease Tracking patients data Web logs, E-Commerce logs, cookies, explicit login Data then used to provide personalized content to site users to: Assist customers in locating their target selections "Encourage" practioner to make certain selections

- Automated Recommender Systems
- Networks and Recommendations
- Web Path Analysis for disease Prediction

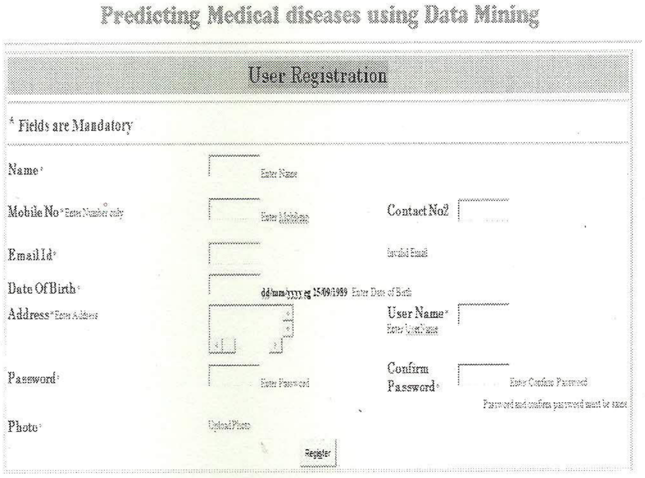
C. System Execution details

➢ **Screen Shots**

1] Login Page


Username
data
Password
••••
27/04/2012

2] User Registration


Predicting Medical diseases using Data Mining

User Registration

* Fields are Mandatory

Name* Enter Name

Mobile No* Enter Number only Enter Mobileno

Contact No2

Email Id* Invalid Email

Date Of Birth* dd/mm/yyyy eg 15/09/1989 Enter Date of Birth

Address* Enter Address

User Name* Enter UserName

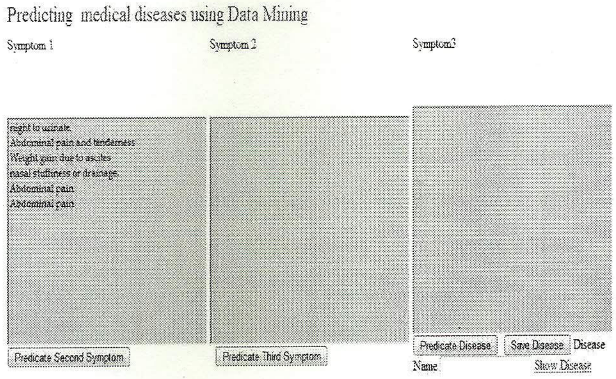
Password* Enter Password

Confirm Password* Enter Confirm Password

Password and confirm password must be same

Photo* Upload Photo

3] Symptoms List


Predicting medical diseases using Data Mining

Symptom 1

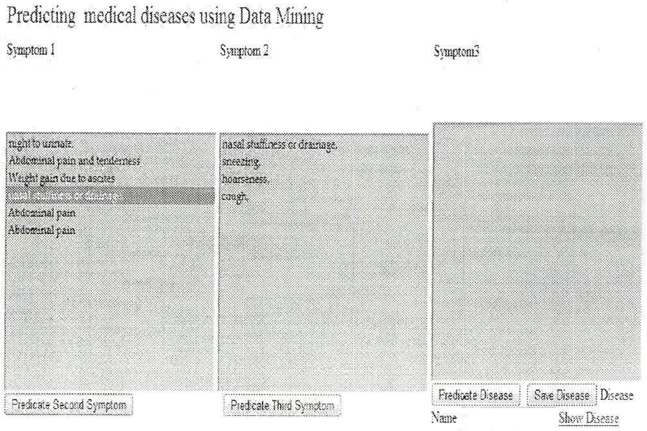
night to urinate
Abdominal pain and tenderness
Weight gain due to ascites
nasal stuffiness or drainage.
Abdominal pain
Abdominal pain

Symptom 2

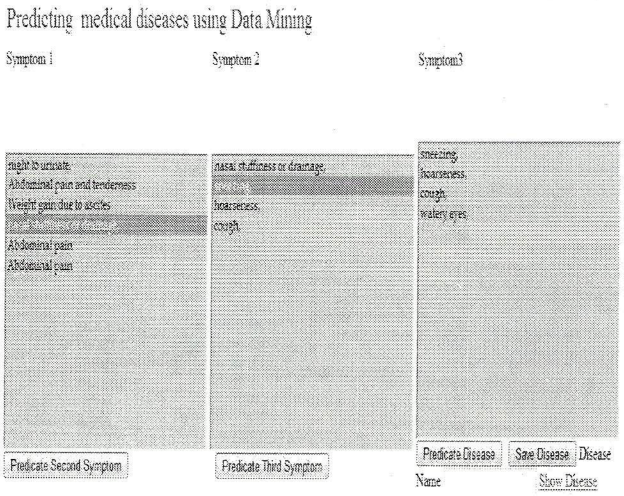
Symptom3

Disease Name

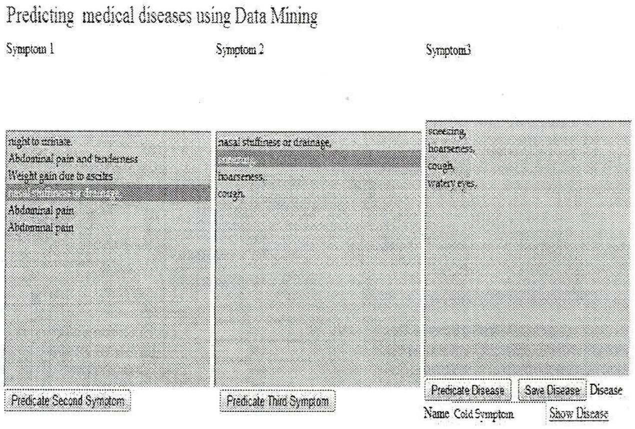
4] Predicting Second Symptom List



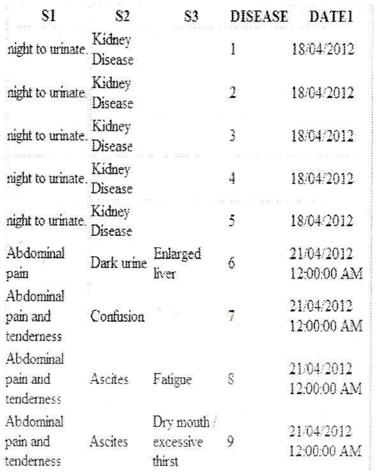
5] Predicting Third Symptom List



6] Finally the disease is being predicted on the basis of all the symptoms



7] After saving the disease we have the records of all the patients



S1	S2	S3	DISEASE	DATE1
night to urinate.	Kidney Disease		1	18/04/2012
night to urinate.	Kidney Disease		2	18/04/2012
night to urinate.	Kidney Disease		3	18/04/2012
night to urinate.	Kidney Disease		4	18/04/2012
night to urinate.	Kidney Disease		5	18/04/2012
Abdominal pain	Dark urine	Enlarged liver	6	21/04/2012 12:00:00 AM
Abdominal pain and tenderness	Confusion		7	21/04/2012 12:00:00 AM
Abdominal pain and tenderness	Ascites	Fatigue	8	21/04/2012 12:00:00 AM
Abdominal pain and tenderness	Ascites	Dry mouth / excessive thirst	9	21/04/2012 12:00:00 AM

VI. CONCLUSION

The mechanism reported in this system focuses on one of the oldest tasks in association mining: based on incomplete information about the medical symptoms, can we predict which other symptom the disease contains? Our literature survey indicates that, while some of the recently published systems can be used to this end, their practical utility is constrained, for instance, by being limited to domains with very few distinct symptoms. Bayesian classifier can be used too, but we are not aware of any systematic study of how it might operate under the diverse circumstances encountered in association mining. We refer to our technique by the acronym DS-ARM [11]. The underlying idea is simple: when presented with an incomplete list s of symptoms in a disease, our program first identifies all high-support, high-confidence rules that have as antecedent a subset of s. Then, it combines the consequents of all these (sometimes conflicting) rules and creates a set of symptoms most likely to complete the disease. Two major problems complicate the task: first, how to identify the relevant rules in a computationally efficient manner; second, how to combine (and quantify) the evidence of conflicting rules. We addressed the former issue by the recently proposed technique of IT-trees and the latter by a few simple ideas from the DS theory. Our experimental results are promising: DS- ARM compares favorably with the Bayesian approach and out performs more traditional approaches even in domains designed in a manner meant to be "tailored" to them. In Particular,DS-ARM performs well in applications where the older approaches incur intractable computational costs (e.g., if there are many distinct items). Besides the encouraging results, our experiments have also identified ample room for further improvements. Also our implementation of the Bayesian classifier can perhaps be found suboptimal. Finally, completely different approaches (beyond Bayesian classification and DS theory) should be explored—a research strand that we strongly advocate.

VII. FUTURE SCOPE

- ✓ This system can be made online.
- ✓ This system can be expanded such that it will predict the disease deficiency cause.
- ✓ This system will be very beneficial for rural physician.
- ✓ This system can also suggest to undergo the disease related tests, surgeries and preventive measures and drugs.
- ✓ Also, if the disease is complicated in future it will also suggest the proper specialist.

VIII. REFERENCES:

[1] S. Noel, V.V. Raghavan, and C.H. Chu, "*Visualizing Association Mining Results through Hierarchical Clusters,"* Proc. Int'l Conf. Data Mining (ICDM '01) pp. 425-432, Nov./Dec. 2001.

[2] P. Bollmann-Sdorra, A. Hafez, and V.V. Raghavan, *"A Theoretical Framework for Association Mining Based on the Boolean Retrieval Model,"* Data Warehousing and Knowledge Discovery: Proc. Third Int'l Conf. (DaWaK '01), pp. 21-30, Sept. 2001.

[3] R. Agrawal, T. Imielinski, and A. Swami, *"Mining Association Rules between Sets of Items in Large Databases,"* Proc. ACM Special Interest Group on Management of Data (ACM SIGMOD), pp. 207-216, 1993.

[4] M. Kubat, A. Hafez, V.V. Raghavan, J.R. Lekkala, and W.K. Chen, *"Itemset Trees for Targeted Association Querying,"* IEEE Trans. Knowledge and Data Eng., vol. 15, no. 6, pp. 1522-1534, Nov./Dec. 2003.

[5] A. Rozsypal and M. Kubat, *"Association Mining in Time-Varying Domains,"* Intelligent Data Analysis, vol. 9, pp. 273-288, 2005.

[6] V. Raghavan and A. Hafez, *"Dynamic Data Mining,"* Proc. 13th Int'l Conf. Industrial and Eng. Applications of Artificial Intelligence and Expert Systems IEA/AIE, pp. 220-229, June 2000.

[10] C.C. Aggarwal, C. Procopius, and P.S. Yu, *"Finding Localized Associations in Market Basket Data,"* IEEE Trans. Knowledge and Data Eng., vol. 14, no. 1, pp. 51-62, Jan./Feb. 2002.

[11] R. Bayardo and R. Agrawal, *"Mining the Most Interesting Rules,"* Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 145-154, 1999.

[12] J. Zhang, S.P. Subasingha, K. Premaratne, M.-L. Shyu, M. Kubat, and K.K.R.G.K. Hewawasam, *"A Novel Belief Theoretic Association Rule Mining Based Classifier for Handling Class Label Ambiguities,"* Proc. Workshop Foundations of Data Mining (FDM '04), Int'l Conf. Data Mining (ICDM '04), Nov. 2004.

Density-Based Clustering with Statistical Functions in Spatial Dataset

[1]Sahar Zarei & [2]Ehsan Moradi Motlagh Jonaghani

Information Technology , Department CSE, JNT University Hyderabad, India
Spatial Information Technology, Department SIT, JNT University Hyderabad
E-mail: sahar.zarei.jntu@gmail.com, ehsan.moradi.m@gmail.com

Abstract— **A density-based algorithm is a well-known clustering method to classified data according to density in arbitrary shapes of point dataset or polygon dataset. However, DBSCAN is a clustering method that suitable for point dataset, P-DBSCAN is suitable for clustering polygonal shapes in geometric dataset. DBSCAN and P-DBSCAN requires minimum domain of knowledge and two main parameters (Eps,MinPts) that support the user to have high inter cluster distance in density-based clustering method. This paper shows that How selecting Epas and MinPts as input parameters have relation with statistical functions like Variance, Covariance and Standard Deviation of points.**

Keywords-component; formatting; Density-based Clustering Algorithm , Arbitrary Shape of Clustering , Spatial Dataset.

I. Introduction

The problem of detecting clusters of points in data is challenging when the clusters are in different size, density and shape. Many of these issues become even more significant when the data is high-dimensionality and when it includes noise and outlier.

DBSCAN (Density-based Spatial Clustering of Application with Noise) based on connected regions with high density that: Firstly, Density of an object can be measured by the number of objects close to it and, Secondly, It connect core objects and neighbourhoods to form dense region as cluster, and also DBSCAN algorithm can discover of cluster with arbitrary shape, because the shape of clusters in spatial databases may be spherical, drown-out, linear, elongated etc. Eps and MinPts are two parameters that use for determining core objects and outlier objects in DBSCAN clustering method.

The rest of paper is organized as follows. We discuss density-based notion of cluster and qualified the neighbourhood of an object by DBSCAN in section 2 and evaluating them. In section 3, we present our algorithm which discovers such clusters in spatial database. In section 4, we performed an experimental evaluation of the effectiveness and efficiency of DBSCAN using real datasets of The USA city Network. Section 5 concludes with a summary and some directions for future research.

II. Literature Review

A. Density-base notion of clusters

A density-based clustering methods have been developed based on the notion of density and the idea is to continue growing a given cluster as long as the density(number of points or objects)in the neighbourhood exceeds some discover clusters of arbitrary shape.

In the following, we try to formalize this intuitive notion of "clusters" and "noise" in database of points of some k-dimension space. Note that, the shape of a neighbourhood is determined by the choice of a distance function for two points denoted by distance (Point1,Point2)that we use the Euclidean Distance function.

B. Qualified the neighbourhood of an objects by DBSCAN

- *Eps > 0*: Eps is a parameter use for specify the radius of a neighborhood we consider for every object.
- *MinPts:* To determine neighborhood is dense or not, we use MinPts and it specify density threshold of dense region. An object is core object if the Eps of an object contains at least MinPts objects and It Core objects are the pillars of dense regions.
- *Directly-density Reachable:* Point (p) is directly-density reachable from Point (q) if Point (p) is within the Eps of Point (q) and Point (q) is core object.

- *Density Reachable:* An object (p) is density-reachable from object (q) if there is a chain of objects between them like p(1),p(2),…,p(n) , with p1=q , pn=p such that p(i)+1 is directly-density reachable from pi for all 1<=i<=n.
- *Density-connectivity:* An object (p) is density connected to object (q), if there is an object (o) such that both (p) and (q) are density reachable from (o).
- *Variance:* A Variance (x) is a measure that how far a set of numbers are spread around average(x) (or Mean value).
- *Covariance :* Is measure of how much two random variables change together and it shows the tendency of points in the linear relationship between the variables. Covariance represents the dependence between the two categories is given. If the result is positive, it means two points will change in same direction. But, if the result is negative, the change will be in opposite directions. There is no dependency between the points if the Covariance is zero (or close to zero).
- Standard Deviation: Widely-used measure of variability used in statistics. It shows how much variation existed from average. A low standard deviation indicates that the data points tend to very close to the Mean value, whereas high standard deviation indicated that the data points are spread out over a large range of values.

  The shape of the neighborhood is determined by the choice of a distance function for two points that denoted by distance (object(p),object (q))that we use Euclidean Distance that is straight distance between two points.

III. DBSCAN ALGORITHM

The steps in DBSCAN algorithm are:

1. Select the dataset domain.
2. Show the Variance, Covariance and STD.
3. All object marked as "unvisited".
4. Randomly select an unvisited objects(p) and mark object(p) as visited.
5. Check the Eps of object(p) contain at least MinPts objects:

   If "No" object is Noise.

   If "Yes"
6. New cluster created for (p).
7. All object in E-neighbourhood (p) add to (N) that is candidate set.
8. Those objects in [N] that not belong to any cluster iteratively adds to our new cluster.

In DBSCAN, the positive or negative Covariance function illustrate that dataset of objects are not distributed in a square or a rectangle area and normally located in one side of the Average value. However, if it is nearby zero value, represent the normal scattering.

IV. EXPERIMENTAL DATASER

Experiments were conducted on two sets of real data the California road network and points of interest dataset, and city of San Joaquin County (TG) road network dataset. The California road network contain 21,047 objects and TG road network contain 18,262 objects. Both datasets are available in http://www.cs.fsu.edu/~lifeifei/SpatialDataset.htm .

V. EXPRIMENTAL RESULT & ANALYSIS

A. California road network and pounts of interest datast

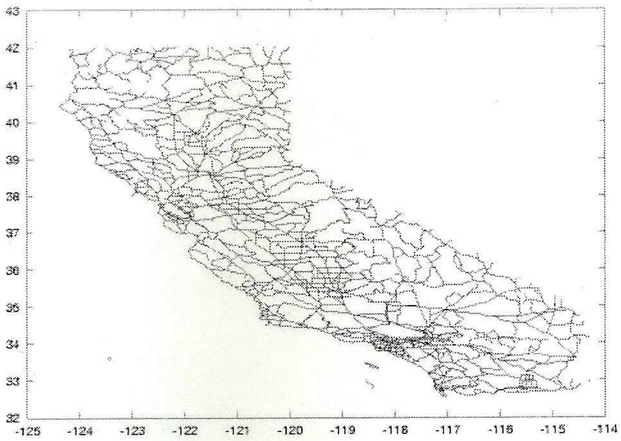


Figure 1-1 California Road Network.No.points:21,048

We applied algorithm on dataset and these are results:

X= (114 to 124) , Y=(32 to 42) , Var=5.9 , Cov=5.1 , STD=2.4.As we have in above table

Eps	MinPts	Time	No.clusters	Noise
50	50	34	1	No
6	59	27	1	No
6	6	27	1	No
50	6	33	1	No

B. City of Oldenburg(OL) Road Network

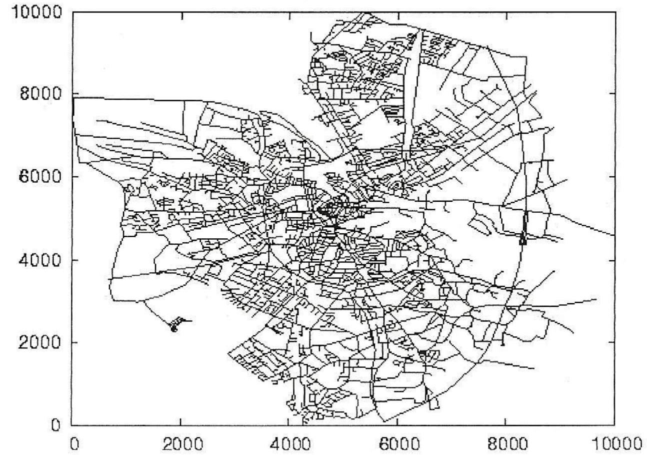


Figure 2-1 Oldenburg Road network.No.points:6,104

We applied the algorithm on dataset and these are result:

X= (0 to 10000) , Y= (0 to 10000) , XSTD=1724 , YSTD=2236. As we have in above table.

Eps	MinPts	Time	No.clusters	Noise
50	50	1:179	No Cluster	All
6	50	1:234	No Cluster	All
6	6	1:255	8	5272
2000	50	1:572	1	No
50	6	1:208	No cluster	All

VI. CONCLUSION

In DBSCAN, the positive and negative covariance illustrate that dataset of objects are not distributed in a square or rectangle area and normally located in one side of the AVG value (Mean value).But, if it is close to zero value, represented the normal scattering.

Given the previous description about impact of statistical function in DBSCAN, We can see that as we select our Eps value by Variance , STD and Covariance, the time will be less than normal DBSCAN algorithm. Sometimes even the number of clusters and noise will changed.

VII. REFERENCES

[1] K.Santhisree, A.Damodaram and SV Appaji, “An Enhanced Dbscan Algorithm to Cluster Web usage Data using Rough Sets and Upper Approximations” , IJCSC Vol.1,January-June 2010,pp.263-265.

[2] Martin Ester, Hans-Peter Kriegel, Gorg sander and Xiaowei Xu , “A Density-based Algorithm for discovering Clusters in Large Spatial Databases with Noise”, published in 2nd International Conference on knowledge Discovery.

[3] Deepti Joshi, Ashok K.Samal, Member IEEE and Leen-Kiat Soh, “Density-Based Clustering of Polygons”, IEEE 2009.

[4] R.T Ng and J.Han, “Efficient and effective clustering methods for spatial data mining” in processing of 20th International conference on very large databases,Santiago,Chile,1994.144-155.

[5] Deepti Joshi, Ashok K.Samal, Member IEEE and Leen-Kiat Soh, “Density-based Clustering of Polygon”, 2009 IEEE.

[6] Anriano Moreira, Maribel Y.Santos and Sofia Carneiro, “Density-based Clustering Algorithms-DBSCAN and SNN”, 2005.

[7] Osmar R.Zaiane, Chi-Hoon Lee, “Clustering Spatial Data in the Presence of Obstacles:a Density-based Approache”, IEEE 2002.

[8] Jiawei Han, Micheline Kamber and Jain Pei, “Data mining Concept and Techniques”,3rd Edition2012

Visual Visitor Verifier

Kishore G. R & Y Manjula

Department of Electronics & Communication
Sri Siddhartha University, Sri Siddhartha Institute of Technology, Tumkur, Karnataka, India

ABSTRACT: ***The main aim of the project is to design a technology where the user can get the benefit of advanced security in the real time environment. The user can use the keypad and with the LCD choose a particular person and the micro controller will read the data and it will then send the same to the GSM modem , which will send the SMS to the cell phone. This cell phone is camouflaged and kept in a place which cannot be seen by the visitor but the cell after getting the SMS will start to record the video of the person for a particular duration of time and after which it will send the SMS to that user. He can then see the live video as MMS and then can decide whether to allow the person.***

1 Introduction

Visual Visitor verifier is used to provide the residents of an apartment visual verification regarding the visitor who wishes to visit their premises. Anytime there is any visitor in the reception of the apartment, the intended resident would get to see the video clipping of the visitor on his cell. After seeing the video, one can decide either to open the entrance of the apartment automatically or let it shut sitting in the apartment itself.

The door of the apartment is always in the closed mode. Every time a visitor enters the reception of the apartment, he is expected to choose the name of the person whom he'd like to visit. This is done with the help of keypad and LCD. Once the selection is done, an auto SMS is sent by the GSM modem to the camera phone hidden in the reception. Content of the SMS would be the name of the resident of the apartment chosen by the visitor. Once the camera phone receives SMS, it invokes camera automatically, starts video recording the visitor's movements for around 2minuts time. Auto sends it as MMS to the intended apartment resident. This camera phone in turn maintains the mapping of names and phone numbers of all the residents of the apartment. After seeing the video if in turn the resident replies back to GSM modem saying "Yes", stepper motor moves clockwise meaning, authentic visitor.

2. Literature survey

A *Implementation Methodology for Interoperable Personal Health Devices With Low-Voltage Low-Power Constraints*

Traditionally, e-Health solutions were located at the point of care (PoC), while the new ubiquitous user-centered paradigm draws on standard-based personal health devices (PHDs). Such devices place strict constraints on computation and battery efficiency that encouraged the International Organization for Standardization/IEEE11073 (X73) standard for medical devices to evolve from X73PoC to X73PHD.

In this context, low-voltage low-power (LV-LP) technologies meet the restrictions of X73PHD-compliant devices. Since X73PHD does not approach the software architecture, the accomplishment of an efficient design falls directly on the software developer. Therefore, computational and battery performance of such LV-LP-constrained devices can even be outperformed through an efficient X73PHD implementation design. In this context, this paper proposes a new methodology to implement X73PHD into microcontroller-based platforms with LV-LP constraints. Such implementation methodology has been developed through a patterns-based approach and applied to a number of X73PHD-compliant agents (including weighing scale, blood pressure monitor, and thermometer specializations) and microprocessor architectures (8, 16, and 32 bits) as a proof of concept.

As a reference, the results obtained in the weighing scale guarantee all features of X73PHD running over a microcontroller architecture based on ARM7TDMI requiring only 168 B of RAM and 2546 B of flash memory.

B: Wireless Control of Humanoid Robot Using 3G

The increase of greed in people has paved way to civil wars and natural disasters. A swift action has to be taken in the relief work of the aftermath of earthquake affected areas, such that any delay in the rescue could lead the death toll to rise. The same can be applied to war fields too. The proposed research focuses on human beings who are alive and struggling for their lives either in the war field or due to natural disasters like earthquakes, to be recognized and rescued in a much faster pace. The robot senses the alive condition of human and sends a notification to the mobile to capture the images of the same. The captured image is then sent to the server to view and act accordingly. The administrator at the server's end has options to move the robot in any required direction for more accurate detection of human beings alive.

C. The Implementation of DTMF signals

As explained, DTMF signals are thus analog, and consist of two sine waves which are

Independent of each other. It is therefore not possible to generate them with only digital components. The digital signals must instead be converted by means of DACs (Digital-to Analog Converters) and/or filters, into the desired sinusoidal waveforms

If DTMF signals are generated from square-waves, then the demands for hardware and software will be at a minimum. Every recurrent waveform having a cycle duration of T can be represented by a Fourier series consisting of the infinite sum of individual sine and cosine waveforms [2], as follows:

$$y(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_0 * \cos(n\omega_0 t) + b_0 * \sin(n\omega_0 t)]$$

$A_0/2$ is the direct component of the signal. The partial component with lowest angular frequency (w_0) is termed the fundamental, and the others are known as overtones or harmonics. A recurrent waveform which can be very easily generated with a microcontroller is the square wave, of which the Fourier series is as follows:

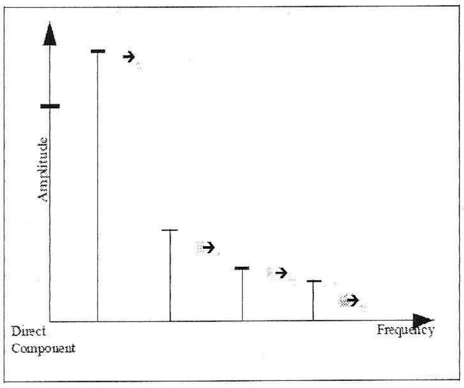


Fig.1: Amplitude Spectrum of Square Wave.

D: Software for the generation of square-wave signals

The software for the generation of the square-wave signals must meet the following requirements:

- It must be able to generate two square-wave signals which are independent of each other.
- In order to separate the signals, two output pins are needed, which provide the outputs of the Hi-Group and the Lo-Group signals respectively.

The MSP430 is provided with various timers which are suitable for generating square-wave signals. In the configuration '31x/'32x, the 8-Bit and Timer Port timers are used, in order to generate both square-wave signals. This software is tested with a MCLK of 1.048 MHz.

Timer_A in the configuration '33x can generate both of the signals which are needed. The second software package uses this timer for the generation of the square-wave signals, and is also be tested with other MCLKs.

3. System design

3.1 Software and Hardware parts

The reception of the apartment has the following setup GSM Modem connected. To GSM Modem, ARM microcontroller, Dial-pad, LCD, stepper motor. Also, a camera enabled cell phone is hidden in the reception. **GSM Modem** used for sending and receiving SMS. **ARM microcontroller** maintains the names of all the residents of the apartment. **Dial pad** Used to scroll up and down the names stored in ARM. **LCD** selected name is displayed on LCD.

Stepper motored ARM's the door of an apartment. Clockwise motion of stepper motor means opening of the door.

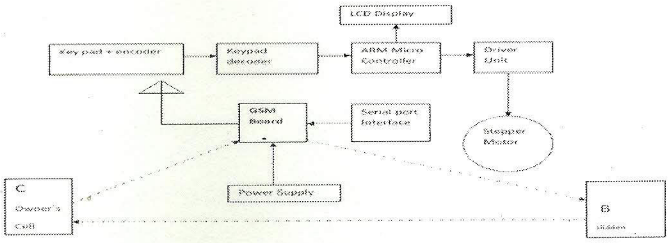


Fig.2 Block diagram of Visual Visitor Verifier

Door of the apartment is always in the closed mode. Every time a visitor enters the reception of the apartment, he is expected to choose the name of the person whom he'd like to visit. This is done with the help of keypad and LCD. Once the selection is done, an auto SMS is sent by the GSM modem to the camera phone hidden in the reception. Content of the SMS would be the name of the resident of the apartment chosen by the visitor. Once the camera phone receives SMS, it invokes camera automatically, starts video recording the visitor's movements for around 2minuts time. Auto sends it as MMS to the intended apartment resident. This camera phone in turn maintains the mapping of names and phone numbers of all the residents of the apartment. After seeing the video if in turn the resident replies back to GSM modem saying "Yes", stepper motor moves clockwise meaning, authentic visitor

3.2 Hardware Design Technology

3.2.1 Brief history of ARM

- ARM is short for Advanced Risc Machines Ltd.
- Founded 1990, owned by Acorn, Apple and VLSI.

- Known before becoming ARM as computer manufacturer
- Acorn which developed a 32-bit RISC processor for it's own use (used in Acorn Archimedes).

3.2.2 ARM architecture

- 32-bit RISC-processor core (32-bit instructions).
- 37 pieces of 32-bit integer registers (16 available).
- Pipelined (ARM7: 3 stages).
- Cached (depending on the implementation).
- Von Neuman-type bus structure (ARM7), Harvard (ARM9).
- 8 / 16 / 32 -bit data types.
- 7 modes of operation (usr, fiq, irq, svc, abt, sys, und).
- Simple structure -> reasonably good speed / power consumption ratio.
- ARM core modes of operation:
- User (usr): Normal program execution state
- FIQ (fiq): Data transfer state (fast irq, DMA-type transfer)
- IRQ (iqr): Used for general interrupt services.
- Supervisor (svc): Protected mode for operating system support.
- Abort mode (abt): Selected when data or instruction fetch is aborted.
- System (sys): Operating system 'privilege'-mode for user.
- Undefined (und): Selected when undefined instruction is fetched

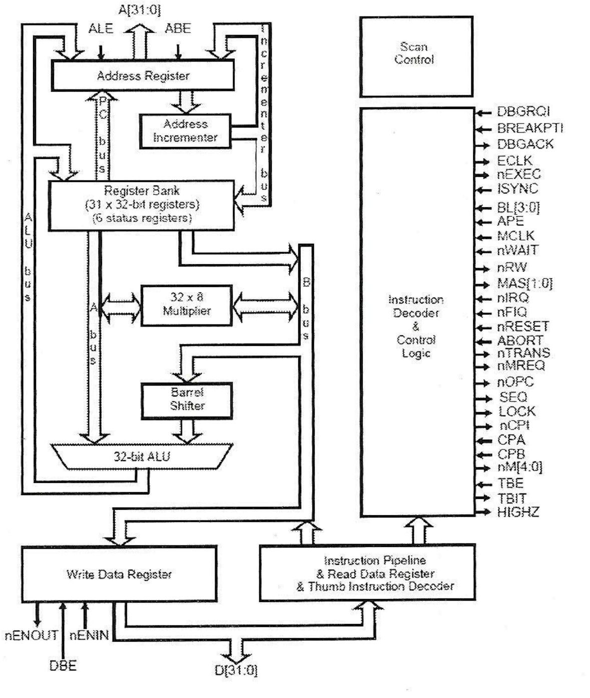


Figure 3 ARM architecture

3.3 Liquid Crystal Display

A x16 Parallel LCD (#603-00006) :General Information

- ARM is one of the most licensed and thus widespread processor cores in the world.
- Used especially in portable devices due to low power consumption and reasonable performance (MIPS / watt).
- Several interesting extensions available or in development like Thumb instruction set.

B LCD Control from a BASIC Stamp

Parallax (www.parallax.com) publishes many circuits and examples to control LCDs. Most of these examples are available for download from our web site. These examples are featured in Stamp Works, the Nuts and Volts of BASIC Stamps books, the free LCD Character Creator Software, and the BS2p plus Pack. Example codes are listed below for the BASIC Stamp 1 and 2 modules.

C Operation of the HD44780 Registers

The HD44780 has two 8 bit registers, an instruction register (IR) and a data register (DR). The IR stores instruction codes such as display clear and cursor shift, and address information for display data RAM (DD RAM) and character generator RAM (CG RAM). The IR can be written from the MPU but not read by the MPU

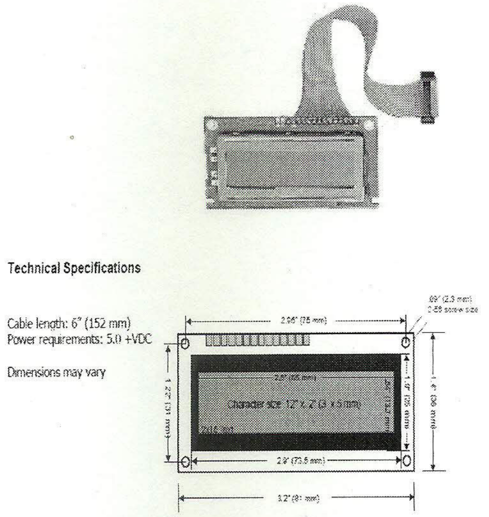


FIGURE 4 Liquid Crystal Display

The DR temporarily stores data to be written into the DD RAM or the CG RAM and data to be read out from the DD RAM or the CG RAM. Data written into the DR from the MPU is automatically written into the DD RAM or the CG RAM by internal operation. The DR is also used for data storage when reading from the DD RAM or the CG RAM. When address information is written into the IR, data is read into the DR from the DD RAM or the CG RAM by internal operation. Data transfer to the MPU is then completed by the MPU reading DR. After the MPU reads the DR, data in the DD RAM or CG RAM at the next address is sent to the DR for the next read from the MPU. Register selector (RS) signals make their selection from these two registers.

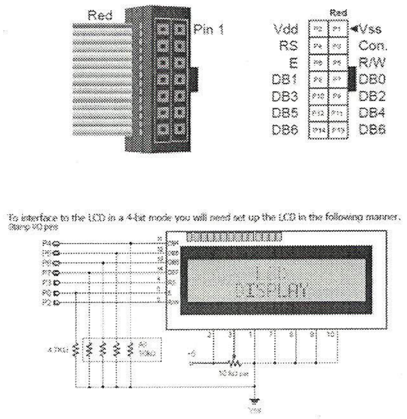


Figure 5 LCD CONTROL DISPLAY

Register selection

RS	R/W	Enable	Operation
0	0	H,H->L	IR as internal operation (clear, etc.)
0	1	H	Read busy flag (DB7) and Address counter
1	0	H,H->L	DR write as internal operation
1	1	H	DR read as internal operation

Busy Flag

When the busy flag is "1", the HD44780 is in the internal operation mode, and the next instruction will not be accepted. As the Register selection table above shows, the busy flag is output to DB7 when RS = 0 and R/W = 1. The next instruction must be written after ensuring that the busy flag is "0".

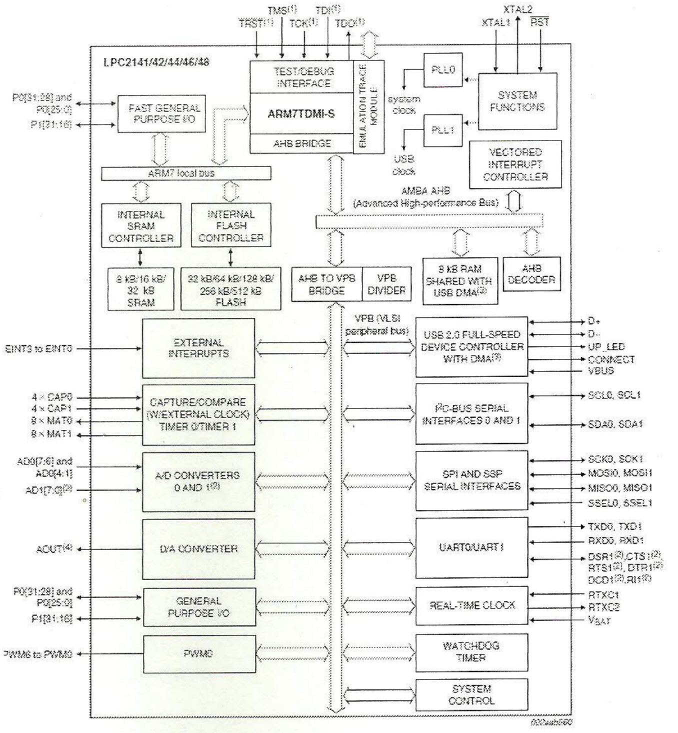
Address counter (AC)

The address counter (AC) assigns addresses to DD and CG RAMs. When an instruction for address is written in IR, the address information is sent from IR to AC. Selection of either DD or CG RAM is also determined concurrently by the instruction. After writing into (or reading from) DD or CG RAM display data, AC is automatically incremented or decremented by 1. AC contents are output as DB0-DB6 when RS = 0 and R/W = 1, as shown in the Register selection table above.

3.4 LPC 2148

The LPC2141/2/4/6/8 microcontrollers are based on a 32/16 bit ARM7TDMI-S CPU with real-time emulation and embedded trace support, that combines the microcontroller with embedded high speed flash memory ranging from 32 kb to 512 kb. A 128-bit wide memory interface and unique accelerator architecture enable 32-bit code execution at the maximum clock rate. For critical code size applications, the alternative 16-bit Thumb mode reduces code by more than 30 % with minimal performance penalty.

Due to their tiny size and low power consumption, LPC2141/2/4/6/8 are ideal for applications where miniaturization is a key requirement, such as access control and point-of-sale. A blend of serial communications interfaces ranging from a USB 2.0 Full Speed device, multiple UARTs, SPI, SSP to I2Cs, and on-chip SRAM of 8 kb up to 40 kb, make these devices very well suited for communication gateways and protocol converters, soft modems, voice recognition and low end imaging, providing both large buffer size and high processing power. Various 32-bit timers, single or dual 10-bit ADC(s), 10-bit DAC, PWM channels and 45 fast GPIO lines with up to nine edge or level sensitive external interrupt pins make these microcontrollers particularly suitable for industrial control and medical systems.



(1) Pins shared with GPIO.

(2) LPC2144/46/48 only.

(3) USB DMA controller with 8 kB of RAM accessible as general purpose RAM and/or DMA is available in LPC2146/48 only.

Figure 6 LPC 2148

Even the GSM SIM card advantage, that allows you to change your cell phone and keep your phone list, is being surplice by some CDMA operators with a service that allows you to store your phone book on the operator's database, allowing you to recover your phone book even if your cell phone is stolen (which is not possible with GSM, since if your cell phone is stolen, your SIM card will be stolen together). Notice that recently a new accessory called SIM backup was released, which allows you to backup the data stored in your SIM card. Also some GSM operators are offering a similar backup service.

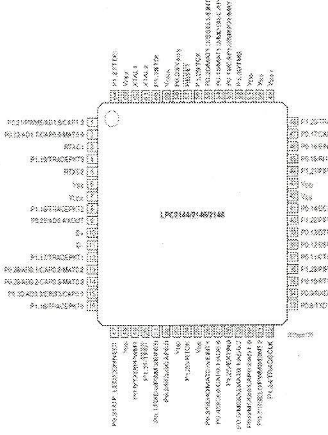
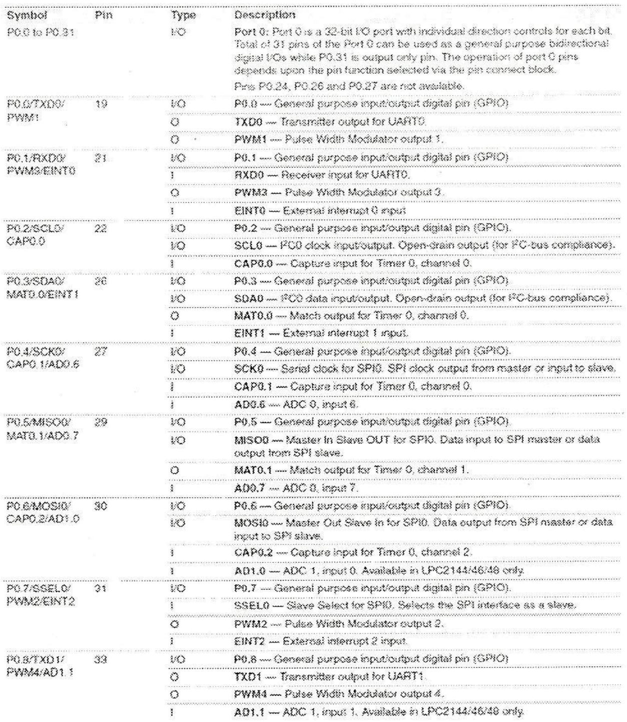


Figure 7 Pin Diagram of LPC 2148



Symbol	Pin	Type	Description
P0.0 to P0.31		I/O	Port 0: Port 0 is a 32-bit I/O port with individual direction controls for each bit. Total of 31 pins of the Port 0 can be used as a general purpose bidirectional digital I/Os while P0.31 is output only pin. The operation of port 0 pins depends upon the pin function selected via the pin connect block. Pins P0.24, P0.26 and P0.27 are not available.
P0.0/TXD0/ PWM1	19	I/O	**P0.0** — General purpose input/output digital pin (GPIO)
		O	**TXD0** — Transmitter output for UART0.
		O	**PWM1** — Pulse Width Modulator output 1.
P0.1/RXD0/ PWM3/EINT0	21	I/O	**P0.1** — General purpose input/output digital pin (GPIO)
		I	**RXD0** — Receiver input for UART0.
		O	**PWM3** — Pulse Width Modulator output 3.
		I	**EINT0** — External interrupt 0 input
P0.2/SCL0/ CAP0.0	22	I/O	**P0.2** — General purpose input/output digital pin (GPIO).
		I/O	**SCL0** — I²C0 clock input/output. Open-drain output (for I²C-bus compliance).
		I	**CAP0.0** — Capture input for Timer 0, channel 0.
P0.3/SDA0/ MAT0.0/EINT1	26	I/O	**P0.3** — General purpose input/output digital pin (GPIO).
		I/O	**SDA0** — I²C0 data input/output. Open-drain output (for I²C-bus compliance).
		O	**MAT0.0** — Match output for Timer 0, channel 0.
		I	**EINT1** — External interrupt 1 input.
P0.4/SCK0/ CAP0.1/AD0.6	27	I/O	**P0.4** — General purpose input/output digital pin (GPIO).
		I/O	**SCK0** — Serial clock for SPI0. SPI clock output from master or input to slave.
		I	**CAP0.1** — Capture input for Timer 0, channel 0.
		I	**AD0.6** — ADC 0, input 6.
P0.5/MISO0/ MAT0.1/AD0.7	29	I/O	**P0.5** — General purpose input/output digital pin (GPIO).
		I/O	**MISO0** — Master In Slave OUT for SPI0. Data input to SPI master or data output from SPI slave.
		O	**MAT0.1** — Match output for Timer 0, channel 1.
		I	**AD0.7** — ADC 0, input 7.
P0.6/MOSI0/ CAP0.2/AD1.0	30	I/O	**P0.6** — General purpose input/output digital pin (GPIO).
		I/O	**MOSI0** — Master Out Slave In for SPI0. Data output from SPI master or data input to SPI slave.
		I	**CAP0.2** — Capture input for Timer 0, channel 2.
		I	**AD1.0** — ADC 1, input 0. Available in LPC2144/46/48 only.
P0.7/SSEL0/ PWM2/EINT2	31	I/O	**P0.7** — General purpose input/output digital pin (GPIO).
		I	**SSEL0** — Slave Select for SPI0. Selects the SPI interface as a slave.
		O	**PWM2** — Pulse Width Modulator output 2.
		I	**EINT2** — External interrupt 2 input.
P0.8/TXD1/ PWM4/AD1.1	33	I/O	**P0.8** — General purpose input/output digital pin (GPIO)
		O	**TXD1** — Transmitter output for UART1.
		O	**PWM4** — Pulse Width Modulator output 4.
		I	**AD1.1** — ADC 1, input 1. Available in LPC2144/46/48 only.

3.5 Symbian OS v7.0 architecture

Symbian is a mobile operating system (OS) and computing platform designed for smart phones and currently maintained by Accenture. The Symbian platform is the successor to Symbian OS and Nokia Series 60; unlike Symbian OS, which needed an additional user interface system, Symbian includes a user interface component based on S60 5th Edition. The latest version, Symbian^3, was officially released in Q4 2010, first used in the Nokia N8. In May 2011 an update, Symbian Anna, was officially announced, followed by Nokia Belle (previously Symbian Belle) in August 2011.

Symbian OS was originally developed by Symbian Ltd. It is a descendant of Psion's EPOC and runs exclusively on ARM processors, although an unreleased x86 port existed.

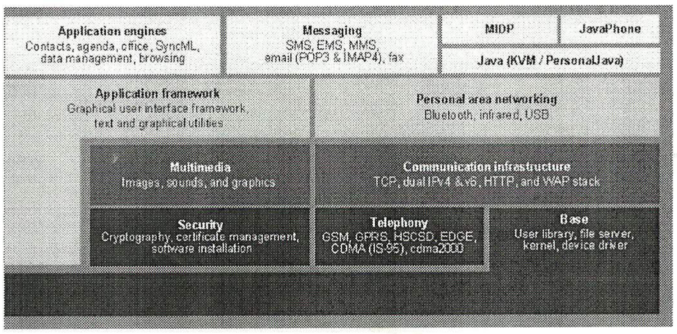

Application engines
Contacts, agenda, office, SyncML, data management, browsing
Messaging
SMS, EMS, MMS, email (POP3 & IMAP4), fax
MIDP
JavaPhone
Java (KVM / PersonalJava)
Application framework
Graphical user interface framework, text and graphical utilities
Personal area networking
Bluetooth, infrared, USB
Multimedia
Images, sounds, and graphics
Communication infrastructure
TCP, dual IPv4 & v6, HTTP, and WAP stack
Security
Cryptography, certificate management, software installation
Telephony
GSM, GPRS, HSCSD, EDGE, CDMA (IS-95), cdma2000
Base
User library, file server, kernel, device driver

Figure 8– Symbian OS v7.0 architecture

3.5.1 Technical Details

Symbian believes that the mobile phone market has five key characteristics that make it unique, and result in the need for a specifically designed operating system:

Mobile phones are both small and mobile. Mobile phones are ubiquitous – they target a mass-market of consumer, enterprise and professional users. Mobile phones are occasionally connected – they can be used when connected to the wireless phone network, locally to other devices, or on their own .Manufacturers need to differentiate their products in order to innovate and compete in a fast-evolving market The platform has to be open to enable independent technology and software vendors to develop third-party applications, technologies and services

Symbian OS has a rich set of APIs for independent software developers, partners and licensees to write their applications. Symbian OS offers an extensive set of messaging APIs. These provide facilities for writing applications that have integrated messaging functionality without having to use low-level APIs. In addition, the messaging framework is open so that developers have the freedom to extend it and create their own protocols for sending and receiving messages. Some possible application messaging enhancements include the following:

The ability to send and receive e-mails, SMS messages, and faxes with the addition of a minimal amount of code. The ability to access the Contacts database from within the application, using the Contacts database API. The ability to use the Global Find API, to search for a text string in messages stored on the phone. The reason is Symbian allows the developers to access the system level services. For example, the default port number for inbox is 0 is restricted to be accessed by the third party application until and unless it has the authentic certificate, but the Symbian allows to access it. So using the third party Symbian application the inbox can be accessed.

The multitasking nature of Symbian OS architecture makes it the ideal application platform for mobile phones. The single-tasking Palm OS is a much more challenging environment for developers of communications applications. Under Symbian OS, each program runs as a separate process, and applications can run concurrently. Each process under Symbian OS contains one or more threads, and the system scheduler allocates processor use to threads through prioritized pre-emptive multitasking. This fully multitasking design allows tasks to be run in the background at the same time an application is running. For example, users can download e-mail and browse the Web at the same time, maintaining an FTP connection.

The Symbian application supports an asynchronous call. Using an asynchronous call means that the application can do something else while waiting for the server to complete the task the same functionality achieved by multithreading. The server then signals to the application when it completes the task. With multithreading the signal is pre-emptive, but with an asynchronous call, the application must check to see if the task has completed. Asynchronous calls therefore provide cooperative multitasking rather than preemptive multitasking. Using asynchronous calls instead of multithreading gives two significant benefits:

4. Advantages and Limitations

Advantages

This project can be used to provide security to the residents of the apartment. One can decide either to open the entrance of the apartment automatically or let it shut sitting in the apartment itself. Camera phone in turn maintains the mapping of names and phone numbers of all the residents of the apartment.

Limitation:

We can dream of establishing live video communication between the owner and the visitor. Mobile phones are ubiquitous they target a mass-market of consumer, enterprise and professional users

5. Conclusion

In this paper, the problem stated earlier is over-come by sending the live video to the user as MMS and this will eliminate having a front office person to monitor the same and many a times humans are prone to make errors or communicate late or give wrong messages. Results are encouraging which has made me to extend this project with online devises.

In extension of this project I am working on MATLAB and the ARM programming to implement the same on cell phone using emulator programming.

ACKNOWLEDGMENT

I am deeply indebted and I would like to express my sincere thanks to our beloved Principal Dr. K.A.Krishnamurthy, for providing me an opportunity to do this project. My special gratitude and sincere thanks to my guide Mrs Y. Manjula, Lecturer, Department of E & C, S.S.I.T for her guidance, constant encouragement and wholehearted support.

Reference:

[1] Bundesamtfür Post und Telekommunikation (Federal Office for Post and Telecommunications): BAPT 223 ZV 5, ZulassungsvorschriftfürEndeinrichtungenzur

[2] Tietze / Schenk: *Halb leiters chaltung stechnik*; (Semiconductor Circuit Design), 10th.Edition; Springer Verlag, Berlin 1993

[3] Marven / Ewers: A Simple Approach to Digital Signal Processing; Texas Instruments,1994

[4] Sauvagerd, Ulrich: A Ten-Channel Equalizer for Digital Audio-Applications; IEEE Transactions on Circuit and Systems, Vol. CAS-36, No 2, February 1989, pp. 276-280

[5]Eine digitale Signal prozesor architektur mitreduzier tem Befehls satzfür Wellen-Digital filter (Digital Processor Architecture with Reduced n Instruction Set for Wave Digital Filters); Dissertation, Fakultätfür Elektrotechnik, Ruhr-Universität Bochum 1991

[6] Lutz Bierl / Texas Instruments: MSP430 Family, Metering Application Report, Texas Instruments, Version 2.1, Jan 1996

Linear Face Recognition Techniques

Kavita Bramhankar[1], Nitin Mishra[2] & Priti Subramanium[3]
[1&2]Dept. of Information Tech, NRIIST, Bhopal, India
[3]SSGB COET. Bhusawal
E-mail : bramhankar.krutika@gmail.com, nitin.nriist@gmail.com, pritikanna559@gmail.com

Abstract - Automated face recognition has become a major field of interest. Face recognition algorithms are used in a wide range of applications viz., security control, crime investigation, and entrance control in buildings, access control at automatic teller machines, passport verification, identifying the faces in a given databases. This paper discusses different face recognition appearance based techniques. They are classified as linear and non-linear. This paper gives the study of linear techniques and comparison between them and focuses on the major problems of face recognition technique.

Keywords - *PCA, Eigen faces, LDA, Fisher faces, Face recognition, Eigen vectors.*

I. INTRODUCTION

The face is our primary focus of attention in social intercourse, playing a major role in conveying identity and emotion. Although the ability to infer intelligence or character from facial appearance is suspect, the human ability to recognize faces. Face recognition has become an important issue in many applications such as security systems, credit card verification and criminal identification. For example, the ability to model a particular face and distinguish it from a large number of stored face models would make it possible to vastly improve criminal identification. A large number of systems has emerged that are capable of achieving recognition rates of greater than 90% under controlled conditions. Successful application under real world conditions remains a challenge though. In field settings, face images are subject to a wide range of variations. These include pose or view angle, illumination, occlusion, facial expression, time delay between image acquisition, and individual differences. The scalability of face recognition systems to such factors is not well understood. Most research has been limited to frontal views obtained under standardized illumination on the same day with absence of occlusion and with neutral facial expression or slight smile. There are three major research groups which propose three different approaches to the face recognition problem. The largest group [1, 2, 3] have dealt with facial characteristics which are used by human beings in recognizing individual faces. The second group [4, 5, 6, 7, 8] performs human face identification based on feature vectors extracted from profile silhouettes. The third group [9, 10] uses feature vectors extracted from a frontal view of the face. Although there are three different approaches to the face recognition problem, there are two basic methods from which these three different approaches arise. The first method is based on the information theory concepts, in other words, on the principal component analysis methods. The second method is based on extracting feature vectors from the basic parts of a face such as eyes, nose, mouth, and chin. The third method is to combine the two approaches i.e. first and second. The outline of a typical face recognition system [13] is given in figure 1.1.

Face image Normalized Face Image Feature vector

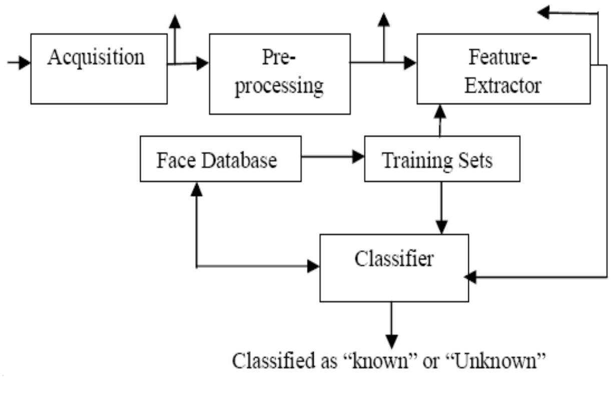


Fig 1.1 : Outline of a typical face recognition system.

II. PRINCIPAL COMPONENT ANALYSYS

PCA also known as Karhunen Loeve projection. PCA Calculates the Eigen vectors of the covariance matrix, and projects the original data onto a lower dimensional feature space, which is defined by Eigen vectors with large Eigen values. PCA has been used in face representation and recognition where the Eigen vectors calculated are referred to as Eigen faces. In gel images, even more than in human faces, the dimensionality of the original data is vast compared to the size of the dataset, suggesting PCA as a useful first step in analysis. The Eigen face algorithm uses the Principal Component Analysis (PCA) for dimensionality reduction to find the vectors which best account for the distribution of face images within the entire image space [11]. These vectors define the subspace of face images and the subspace is called face space. All faces in the training set are projected onto the face space to find a set of weights that describes the contribution of each vector in the face space. To identify a test image, it requires the projection of the test image onto the face space to obtain the corresponding set of weights. By comparing the weights of the test image with the set of weights of the faces in the training set, the face in the test image can be identified. The key procedure in PCA is based on Karhumen-Loeve transformation [12]. If the image elements are considered to be random variables, the image may be seen as a sample of a stochastic process. PCA based on information theory concepts, seek a computational model that best describes a face, by extracting the most relevant information contained in that face. Goal is to find out the eigenvectors (eigenfaces) of the covariance matrix of the distribution, spanned by a training set of face images. Later, every face image is represented by a linear combination of these eigenvectors. Evaluation of these eigenvectors is quite difficult for typical image sizes but, an approximation that is suitable for practical purposes is also presented. Recognition is performed by projecting a new image into the subspace spanned by the eigenfaces and then classifying the face by comparing its position in face space with the positions of known individuals. A face recognition system, based on the eigenfaces approach is proposed. Eigenfaces approach seems to be an adequate method to be used in face recognition due to its simplicity, speed and learning capability. The Principal Component Analysis basis vectors are defined as the eigenvectors of the scatter matrix *ST,*

$$ST = \sum_{i=1}^{N} (x_i - \mu)\,(x_i - \mu)^T \qquad (1)$$

where μ is the mean of the data. PCA determines the orthogonal projection ϕ in

$$Y_k = \phi^T \; x_k, \; k=1,\ldots..,N$$

that maximizes the determinant of the total scatter matrix of the projected samples y1,………yN.

$$\Phi_{OPT} = \underset{\Phi}{\arg\max} \;\; \phi^T \; \left| ST\,\phi \right| \qquad (2)$$

The transformation matrix W_{PCA} is composed of the eigenvectors corresponding to the *d* largest eigenvalues. A 2D example of PCA is demonstrated in Fig. 1. After applying the projection, the input vector (face) in an n-dimensional space is reduced to a feature vector in a d-dimensional subspace. For most applications, these eigenvectors corresponding to very small eigenvalues are considered as noise, and not taken into account during identification. Several extensions of PCA are developed, such as modular eigenspaces [20] and probabilistic subspaces [21]. PCA involves finding a linear subspace (referred to as the Eigenspace) that maximizes the variance between images in a training set. The high-dimensional image can be projected onto the principal components (referred to as eigenimages), and classification can be carried out by performing a nearest neighbor search in the eigenspace [2]–[6].

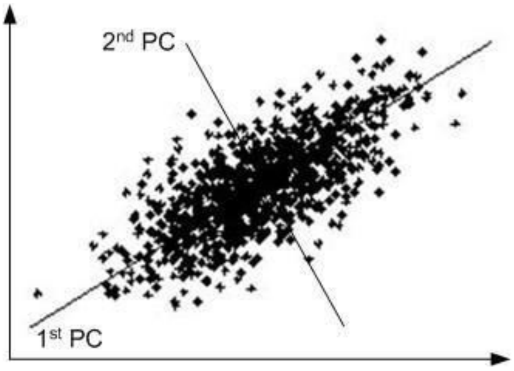


Fig. 1: Principal components (PC) of a two-dimensional set of points. The first principal component provides an optimal linear dimension reduction from 2D to 1D, in the sense of the mean square error.

An alternate approach for classification (particularly when dealing with multiple classes) is to use a class-specific linear projection such as FLDA [16]. All of these techniques require the computation of the principal components before classification can be performed. Computing the principal components of a large set of images is prohibitively expensive and thereby discourages the use of PCA-based techniques in real-world applications. Reducing the computational burden associated with computing the principal components has been addressed using several different approaches based on either iterative power methods, conjugate gradient algorithms, or eigenspace updating. A fundamentally different approach was proposed by Chang *et al.* [17], where the authors show that the Fourier transform can be used to approximate the desired subspace dimension, as well as the principal eigenimages, if the image data set is correlated in one dimension. This result has recently been extended to correlations in higher dimensions that are due to a change in orientation (assuming constant lighting conditions). PCA is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension. The basic goal is to implement a simple face recognition system, based on well-studied and well-understood methods. One can choose to go into depth of one and only one of those methods. The method to be implemented is the PCA (Principle Component Analysis). It is one of the more successful techniques of face recognition and easy to understand and describe using mathematics. This method involves using Eigen faces. The first step is to produce a feature detector (dimension reduction). Principal Components Analysis (PCA) was chosen because it is the most efficient technique, of dimension reduction, in terms of data compression. This allows the high dimension data, the images, to be represented by lower dimension data and so hopefully reducing the complexity of grouping the images.

Benefits of PCA

- The basic Benefit in PCA is to reduce the dimension of the data.
- No data redundancy as components is Orthogonal.
- With help of PCA, complexity of grouping the images can be reduced.
- Application of PCA in the prominent field of criminal investigation is beneficial.
- PCA also benefits entrance control in buildings, access control for computers in general, for automatic teller machines in particular, day-to-day affairs like withdrawing money from bank account, dealing with the post office, passport verification and identifying the faces in a given databases.

PCA Features

- PCA computes means, variances, covariance's, and correlations of large data sets
- PCA computes and ranks principal component and their variances.
- Automatically transforms data sets.
- PCA can analyze datasets up to 50,000 rows and 200 columns.

III. LINEAR DISCRIMINANT ANALYSIS

LDA is widely used to find linear combinations of features while preserving class separability. Unlike PCA, LDA tries to model the differences between classes. Classic LDA is it requires data points for different classes to be far from each other, while point from the same class are close. Consequently, LDA obtains differenced projection vectors for each class. Multi-class LDA algorithms which can manage more than two classes are more used. Suppose we have m samples x1,...,xm belonging to c classes; each class has mk elements. We assume that the mean has been extracted from the samples, as in PCA. Both PCA and ICA construct the face space without using the face class (category) information. The whole face training data is taken as a whole. In LDA the goal is to find an efficient or interesting way to represent the face vector space. But exploiting the class information can be helpful to the identification tasks, see Fig 2. for an example. The Fisher face algorithm is derived from the Fisher Linear Discriminant (FLD), which uses class specific information. By defining different classes with different statistics, the images in the learning set are divided into the corresponding classes. Then, techniques similar to those used in Eigenface algorithm are applied. The Fisher face algorithm results in a higher accuracy rate in recognizing faces when compared with Eigenface algorithm The Linear Discriminant Analysis finds a transform W_{LDA}, such that

$$W_{LDA} = \arg\max_{W} \frac{W^T S_B W}{W^T S_W W} \quad (3)$$

Where *SB* is the between-class scatter matrix and *SW* is the within-class scatter matrix, defined as

$$S_B = \sum_{i=1}^{c} N_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (4)$$

$$S_W = \sum_{i=1}^{C} \sum_{x_k \in x_t} (x_k - \mu_i)(x_k - \mu_i)^T \quad (5)$$

In the above expression, *Ni* is the number of training samples in class *i*, *c* is the number of distinct classes, μ_i is the mean vector of samples belonging to class *i* and X_i represents the set of samples belonging to class *i*. The LDA basis vectors are demonstrated in Figure 3.

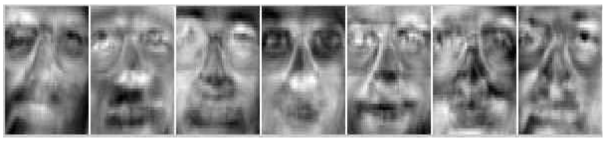


Fig. 2 : First seven LDA basis vectors shown as *p X p* images.

This uses the available class information to compute a projection better suited for discrimination tasks. We define the within-class scatter matrix S_W in equation (5) and we define the between-class scatter matrix S_B in equation (4). Φ_{OPT} is found by solving the generalized eigenvalue problem

$$S_{B\ \phi} = \lambda\ S_W\, \phi \qquad (6)$$

Due to the structure of the data the within-class scatter matrix SW is always singular. We can overcome this problem by first using PCA to reduce the dimensionality and then applying LDA .The overall projection is therefore given by

$$W^{T}_{OPT} = \phi^{T}_{OPT} \quad \phi^{T}_{OPT} \qquad (7)$$

FLDA determines the linear subspace that maximizes the variance between different classes in the data while minimizing the variance within each class. A common issue associated with FLDA is the small-sample size problem. Fortunately, this issue can be overcome by using an intermediate space that is computed using the principal components of the image data. Both PCA and FLDA find a linear subspace for classification using the global information contained in the data set. FLDA attempts to find a linear subspace that minimizes the within-class scatter while maximizing the between class scatter. The idea is that the set of images *Ir* lies close to a low-dimensional linear subspace and is therefore linearly separable [8], [22], [31], [32]. Unfortunately, in most practical applications, $n < m$, which implies that the within-class scatter matrix is singular. To overcome this issue, the within-class scatter matrix is typically projected onto an intermediate subspace using PCA.

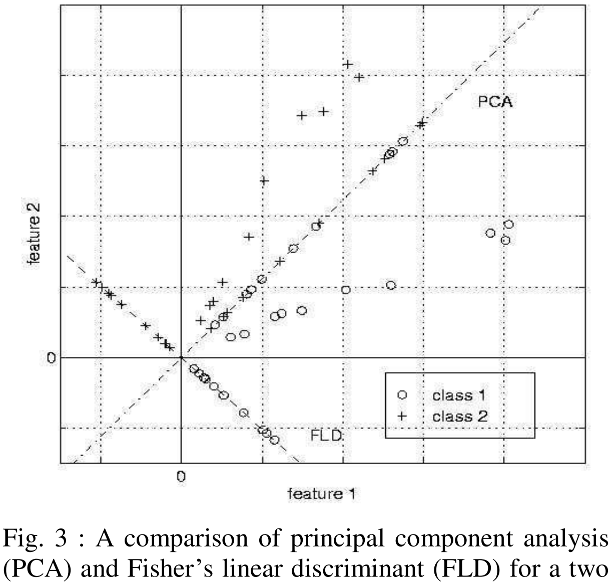


Fig. 3 : A comparison of principal component analysis (PCA) and Fisher's linear discriminant (FLD) for a two class problem where data for each class lies near a linear subspace. It shows that FLD is better than PCA in the sense of discriminating the two classes.

IV. ICA (INDEPENDENT COMPONENT ANALYSIS)

Independent Component Analysis (ICA) [22] is similar to PCA except that the distribution of the components is designed to be non-Gaussian. Maximizing non-Gaussianity promotes statistical independence. Figure 11 presents the different feature extraction properties between PCA and ICA.There are two architectures based on Independent Component Analysis, statistically independent basis images and a factorial code representation, for the face recognition task. The ICA separates the high-order moments of the input in addition to the second-order moments utilized in PCA. Both the architectures lead to a similar performance. The obtained basis vectors based on fast fixed-point algorithm [24] for the ICA factorial code representation are illustrated in Fig. 12. There is no special order imposed on the ICA basis vectors.

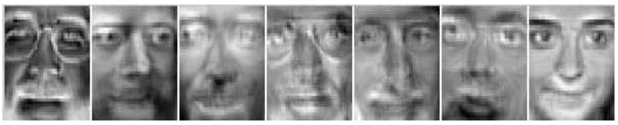


Fig. 4 : ICA basis vectors shown as *p £ p* images; there is no special order for ICA basis vectors

For below example, ICA tends to extract more intrinsic structure of the original data clusters. Independent Component Analysis aims to transform the data as linear combinations of statistically independent data points. Therefore, its goal is to provide an independent rather that uncorrelated image representation. ICA is an alternative to PCA which provides a more powerful data representation [67]. It's a discriminant analysis criterion, which can be used to enhance PCA.The ICA algorithm is performed as follows [25]. Let C_x be the covariance matrix of an image sample X.

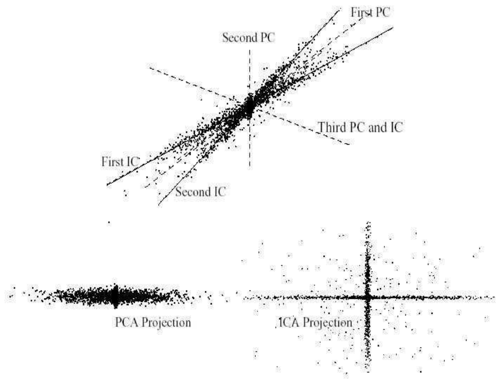


Fig. 5.Top: Example 3D data distribution and the corresponding principal component and independent component axes. Each axis is a direction found by PCA or ICA. Note the PC axes are orthogonal while the IC

axes are not. If only 2 components are allowed, ICA chooses a different subspace than PCA. Bottom left: Distribution of the first PCA coordinate of the data. Bottom right: distribution of the first ICA coordinate of the data [23].

The ICA of X factorizes the covariance matrix c_x into the following form

$$c_{x} = F \Delta F^T$$

Where Δ is diagonal real positive and F transforms the original data into Z (X = FZ). The components of Z will be the most independent possible. To derive the ICA transformation F,

$$X = \phi \wedge^{1/2} U$$

where X and _ are derived solving the following Eigen problem

$$c_x = \phi \wedge \phi^T$$

Then, there are rotation operations which derive independent components minimizing mutual information. Finally, normalization is carried out.

V. PCA PLUS LDA

Swets and Weng proposed PCA plus LDA for face recognition. They applied the PCA for dimensionality reduction of the original image. In this top 15 principal axes were selected and used to derive a 15 dimensional feature vector for every sample. The transformed samples are then used as bases to execute LDA.This approach can be decomposed into two processes, the PCA process followed by the LDA process. They observes a peak recognition rate of more than 90%.Belhumeur and Zhao have proposed system which used similar methodology and named as "Fisherfaces".The face images used in this approach contained face, hair shoulder and background, not solely face. By conducting various experiments it is observed that in the PCA plus LDA approach the non-face portion dominated the entire recognition process. for this approach each projective vector extracted from a face image is 15 dimensional. Experimental results show that influence of middle face portion on the recognition process was much smaller than that of the non-face portion. That is the non-face portion of a face image dominated the recognition process.

VI. COMPARISON BETWEEN PCA , LDA & ICA

- LDA is able to use a smaller size i.e. 15 of subspace to achieve a higher recognition accuracy than LDA i.e. 30.
- LDA has less error rate i.e. 7.3 than PCA i.e.24.4 with close crop.
- LDA has less full face error rate i.e.0.6 than LDA i.e.19.4.
- It is observed that accuracy of PCA and LDA is similar i.e. 70% and projections applied on both are linear.
- For floating point operations PCA gives 10^8 computations while ICA gives 10^9 computations.
- PCA and LDA show the opposite pattern for lower face Occlusion and upper face Occlusion.
- LDA shows better performance to lighting conditions.

VII. PROBLEMS OF FACE RECOGNITION

Face recognition faces some issues inherent to problem definition, environmental conditions and hardware constraints.

- Illumination: The colour that we perceive from a given surface depends not only on the surface's nature but also on the light upon it. There can be relevant illumination variations on images taken under uncontrolled environment. As many feature extraction methods relay on color/intensity variability measures between pixels to obtain relevant data, they show an important dependency on lighting changes. The big problem is that two faces of the same subject but with illumination variations may show more differences between them than compared to another subject. Summing up, illumination is one of the big challenges of automated face recognition systems. all this linear analysis algorithms do not capture satisfactorily illumination variations. This illumination problem can be faced employing different approaches like Heuristic approach, Statistical approach, Light-modeling approach etc.
- Pose: Pose variation and illumination are the two main problems face by face recognition researchers. The uncontrolled environment constraint involves several obstacles for face recognition. Pose variation is another one. There are several approaches used to face pose variations like Multi-image based approaches, Single-model based approaches, Geometric approaches etc.
- Occlusion: The recognition process can rely heavily on the availability of a full input face. Therefore, the absence of some parts of the face may lead to a bad classification. This problem speaks in favor of a piecemeal approach to feature extraction, which doesn't depend on the whole face. There are also objects that can occlude facial features -glasses, hats, beards, certain hair cuts, etc.

- Expression**:** Facial expression is another variability provider. However, it isn't as strong as illumination or pose. Several algorithms don't deal with this problem in a explicit way, but they show a good performance when different facial expressions are present. On the other hand, the addition of expression variability to pose and illumination problems can become a real impediment for accurate face recognition.

VIII. CONCLUSIONS

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained. Causal Productions has used its best efforts to ensure that the templates have the same appearance.

IX. ACKNOWLEDGMENT

The heading of the Acknowledgment section and the References section must not be numbered.

Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

REFERENCES

[1] Kirby, M., and Sirovich, L., "Application of the Karhunen-Loeve procedure for the characterization of human faces", IEEE PAMI, Vol. 12, pp. 103-108, (1990).

[2] Sirovich, L., and Kirby, M., "Low-dimensional procedure for the characterization of human faces", J. Opt. Soc. Am. A, 4, 3, pp. 519-524, (1987).

[3] Terzopoulos, D., and Waters, K., "Analysis of facial images using physical and anatomical models", Proc. 3rd Int. Conf. on Computer Vision,pp. 727-732, (1990).

[4] Manjunath, B. S., Chellappa, R., and Malsburg, C., "A feature based approach to face recognition", Trans. of IEEE, pp. 373-378, (1992).

[5] Harmon, L. D., and Hunt, W. F., "Automatic recognition of human face profiles", Computer Graphics and Image Processing, Vol. 6, pp. 135-156, (1977).

[6] Harmon, L. D., Khan, M. K., Lasch, R., and Ramig, P. F., "Machine identification of human faces", Pattern Recognition, Vol. 13(2), pp. 97-110, (1981).

[7] Kaufman, G. J., and Breeding, K. J, "The automatic recognition of human faces from profile silhouettes", IEEE Trans. Syst. Man Cybern., Vol. 6, pp. 113-120, (1976).

[8] Wu, C. J., and Huang, J. S., "Human face profile recognition by computer", Pattern Recognition, Vol. 23(3/4), pp. 255-259, (1990).

[9] Kerin, M. A., and Stonham, T. J., "Face recognition using a digital neural network with self-organizing capabilities", Proc. 10th Int. Conf. on Pattern Recognition, pp.738-741, (1990).

[10] Nakamura, O., Mathur, S., and Minami, T., "Identification of human faces based on isodensity maps", Pattern Recognition, Vol. 24(3), pp. 263-272, (1991).

[11] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71–86, Mar. 1991

[12] "Why recognition in a statistics based face recognition system should be based on the pure face portion" ,Li-fen chen,Hong Yuan mark liao,Ja-chen Lin,Chin- Chuan han.

[13] "A Comparative study of Face Recognition with Principal Component Analysis and Cross-Correlation Technique", Srinivasulu Asadi, Dr.Ch.D.V.Subba Rao, V.Saikrishna, International Journal of Computer Applications (0975 – 8887) Volume 10– No.8, November 2010.

[14] "Face Recognition Algorithms", Proyecto Fin de Carrera June 16, 2010.

[15] M. Kirby and L. Sirovich, "Application of the Karhunen-Lo´eve procedure for the characterization of human faces," IEEE Trans Pattern Analysis and Machine Intelligence, vol. 12, no. 1, pp. 103–108, Jan. 1990.

[16] "Quo vadis Face Recognition?" Ralph Gross Jianbo Shi Jeff Cohn CMU-RI-TR-01-17, June 2001.

[17] "Face recognition using eigenfaces",Istanbul technical university Institute of science and Technolog ilker atalay January, 1996

[18] "Fast Eigenspace Decomposition of Images of Objects With Variation in Illumination and Pose Randy C. Hoover, Member, IEEE, Anthony A. Maciejewski, Fellow, IEEE, and Rodney G. Roberts,IEEE transactions on systems, man, and cybernetics—part b: cybernetics, vol. 41, no. 2, April 2011 .

[19] M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski, "Face recognition by independent component analysis," IEEE Trans. Neural Networks, vol. 13, no. 6, pp. 1450–1464, 2002 .

[20] B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 6, pp. 780–788, Feb. 2002.

[21] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711–720, Jul. 1997.

◈◈◈

Adaptive Coding Techniques to Improve BER in OFDM System

Rajeshree Raut[1], Priti Subramanium[2] & Kavita Bramahankar[3]

[1]R& D, SRKNEC, Nagpur-13
[2&3]S.S.G.B.C.O.E.T,Bhusawal
E-mail : raut.rajeshree@gmail.com[1], pritikanna559@gmail.com[2], Bramhankar.krutika@gmail.com[3]

Abstract **-** Adaptive modulation and diversity combining represent very important adaptive solutions for the future generations of communication systems. In order to improve the performance and the efficiency of wireless communication systems these two techniques have been recently used jointly in new schemes named joint adaptive modulation and diversity combining .The highest spectral efficiency with the lowest possible combining complexity, given the fading channel conditions and the required error rate performance. Increase the spectral efficiency with a slight increase in the average number of combined path for the low signal to noise ratio (SNR) range while maintaining compliance with the bit error rate (BER).

***Keywords* -** *BER, OFDM, WCDMA, DAB, DVB.*

I. INTRODUCTION

Wireless communications is an emerging field, which has seen enormous growth in the last several years. The huge uptake rate of mobile phone technology, Wireless Local Area Networks (WLAN) and the exponential growth of the Internet have resulted in an increased demand for new methods of obtaining high capacity wireless networks. Most WLAN systems currently use the IEEE802.11b standard, which provides a maximum data rate of 11 Mbps. Newer WLAN standards such as IEEE802.11a and HiperLAN2, are based on OFDM technology and provide a much higher data rate of 54 Mbps. However systems of the near future will require WLANs with data rates of greater than 100 Mbps, and so there is a need to further improve the spectral efficiency and data capacity of OFDM systems in WLAN applications.

For cellular mobile applications, we will see in the near future a complete convergence of mobile phone technology, computing, Internet access, and potentially many multimedia applications such as video and high quality audio. In fact, some may argue that this convergence has already largely occurred, with the advent of being able to send and receive data using a notebook computer and a mobile phone.

Although this is possible with current 2G (2nd Generation) Mobile phones, the data rates provided are very low (9.6 kbps – 14.4 kbps) and the cost is high (typically $0.20 - $1.30 AUD per minute), limiting the usefulness of such a service.

The goal of third and fourth generation mobile networks is to provide users with a high data rate, and to provide a wider range of services, such as voice communications, videophones, and high speed Internet access. The higher data rate of future mobile networks will be achieved by increasing the amount of spectrum allocated to the service and by improvements in the spectral efficiency. OFDM is a potential candidate for the physical layer of fourth generation mobile systems. This thesis presents techniques for improving the spectral efficiency of OFDM systems applied in WLAN and mobile networks.

THIRD GENERATION WIRELESS SYSTEMS

Third generation mobile systems such as the Universal Mobile Telecommunications System (UMTS) [1], [2], [3], [4] and CDMA2000 [6] will be introduced over the next 1-5 years (2002 onwards) [5]. These systems are striving to provide higher data rates than current 2G systems such as the Global System for Mobile communications (GSM) and IS-95. Second generation systems are mainly targeted at providing voice services, while 3rd generation systems will shift to more data oriented services such as Internet access. Third generation systems use Wide-band Code Division Multiple Access (WCDMA) as the carrier modulation scheme [10]. This modulation scheme has a high multipath tolerance, flexible data rate, and allows a greater cellular spectral efficiency than 2G systems. Third generation systems will provide a significantly higher data rate (64 kbps – 2 Mbps) [1] than second-generation systems (9.6 – 14.4 kbps). The higher data rate of 3G systems will be able to support a wide range of applications including Internet access, voice communications and mobile videophones. In addition to this, a large number of new applications will emerge to utilise the permanent network connectivity, such as wireless appliances, notebooks with built in mobile phones, remote logging, wireless web cameras, car navigation systems, and so forth. In fact most of these applications will not be limited by the data rate provided by 3G systems, but by the cost of the service.

The demand for use of the radio spectrum is very high, with terrestrial mobile phone systems being just one of many applications vying for suitable bandwidth. These applications require the system to operate reliably in non-line-of-sight environments with a propagation distance of 0.5 - 30 km, and at velocities up to 100 km/hr or higher. This operating environment limits the maximum RF frequency to 5 GHz, as operating above this frequency results in excessive channel path loss, and excessive Doppler spread at high velocity. This limits the spectrum available for mobile applications, making the value of the radio spectrum extremely high. In Europe auctions of 3G licenses of the radio spectrum began in 1999. In the United Kingdom, 90 MHz of bandwidth [8] was auctioned off for £22.5 billion [9]. In Germany the result was similar, with 100 MHz of bandwidth raising $46 billion (US) [7]. This represents a value of around $450 Million (US) per MHz. The length of these license agreements is 20 years [8] and so to obtain a reasonable rate of return of 8% on investment, $105 Million (US) per MHz must be raised per year. It is therefore vitally important that the spectral efficiency of the communication system is maximised, as this is one of the main limitations to providing a low cost high data rate service.

4TH GENERATION SYSTEMS AND BEYOND

Research has just recently begun on the development of 4th generation (4G) mobile communication systems. The commercial rollout of these systems is likely to begin around 2008 - 2012, and will replace 3rd generation technology. Few of the aims of 4G networks have yet been published, however it is likely that they will be to extend the capabilities of 3G networks, allowing a greater range of applications, and improved universal access. Ultimately 4G networks should encompass broadband wireless services, such as High Definition Television (HDTV) (4 - 20 Mbps) and computer network applications (1 - 100 Mbps). This will allow 4G networks to replace many of the functions of WLAN systems. However, to cover this application, cost of service must be reduced significantly from 3G networks. The spectral efficiency of 3G networks is too low to support high data rate services at low cost. As a consequence one of the main focuses of 4G systems will be to significantly improve the spectral efficiency. In addition to high data rates, future systems must support a higher Quality Of Service (QOS) than current cellular systems, which are designed to achieve 90 - 95% coverage [11], i.e. network connection can be obtained over 90 - 95% of the area of the cell. This will become inadequate as more systems become dependent on wireless networking. As a result 4G systems are likely to require a QOS closer to 98 - 99.5%.

ORTHOGONAL FREQUENCY DIVISION MULTIPLEXING

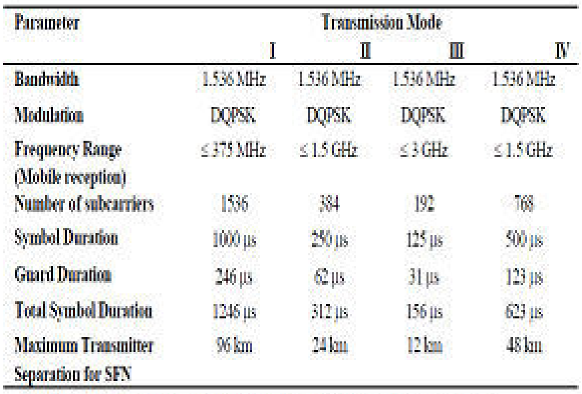
Orthogonal Frequency Division Multiplexing (OFDM) is an alternative wireless modulation technology to CDMA.

OFDM has the potential to surpass the capacity of CDMA systems and provide the wireless access method for 4G systems. OFDM is a modulation scheme that allows digital data to be efficiently and reliably transmitted over a radio channel, even in multipath environments. OFDM transmits data by using a large number of narrow bandwidth carriers. These carriers are regularly spaced in frequency, forming a block of spectrum. The frequency spacing and time synchronisation of the carriers is chosen in such a way that the carriers are orthogonal, meaning that they do not cause interference to each other. This is despite the carriers overlapping each other in the frequency domain. The name ‘OFDM’ is derived from the fact that the digital data is sent using many carriers, each of a different frequency (Frequency Division Multiplexing) and these carriers are orthogonal to each other, hence Orthogonal Frequency Division Multiplexing. The origins of OFDM development started in the late 1950’s with the introduction of Frequency Division Multiplexing (FDM) for data communications. In 1966 Chang patented the structure of OFDM and published the concept of using orthogonal overlapping multi-tone signals for data communications. In 1971 Weinstein introduced the idea of using a Discrete Fourier Transform (DFT) for implementation of the generation and reception of OFDM signals, eliminating the requirement for banks of analog subcarrier oscillators. This presented an opportunity for an easy implementation of OFDM, especially with the use of Fast Fourier Transforms (FFT), which are an efficient implementation of the DFT. This suggested that the easiest implementation of OFDM is with the use of Digital Signal Processing (DSP), which can implement FFT algorithms. It is only recently that the advances in integrated circuit technology have made the implementation of OFDM cost effective. The reliance on DSP prevented the wide spread use of OFDM during the early development of OFDM. It wasn’t until the late

1980's that work began on the development of OFDM for commercial use, with the introduction of the Digital Audio Broadcasting (DAB) system.

II. DIGITAL AUDIO BROADCASTING

DAB was the first commercial use of OFDM technology [19], [20]. Development of DAB started in 1987 and services began in U.K and Sweden in 1995. DAB is a replacement for FM audio broadcasting, by providing high quality digital audio and information services. OFDM was used for DAB due to its multipath tolerance. Broadcast systems operate with potentially very long transmission distances (20 -100 km). As a result, multipath is a major problem as it causes extensive ghosting of the transmission. This ghosting causes Inter-Symbol Interference (ISI), blurring the time domain signal. For single carrier transmissions the effects of ISI are normally mitigated using adaptive equalisation. This process uses adaptive filtering to approximate the impulse response of the radio channel. An inverse channel response filter is then used to recombine the blurred copies of the symbol bits. This process is however complex and slow due to the locking time of the adaptive equaliser. Additionally it becomes increasing difficult to equalise signals that suffer ISI of more than a couple of symbol periods. OFDM overcomes the effects of multipath by breaking the signal into many narrow bandwidth carriers. This results in a low symbol rate reducing the amount of ISI. In addition to this, a guard period is added to the start of each symbol, removing the effects of ISI for multipath signals delayed less than the guard period (see section 2.3 for more detail). The high tolerance to multipath makes OFDM more suited to high data transmissions in terrestrial environments than single carrier transmissions.



Parameter	Transmission Mode			
	I	II	III	IV
Bandwidth	1.536 MHz	1.536 MHz	1.536 MHz	1.536 MHz
Modulation	DQPSK	DQPSK	DQPSK	DQPSK
Frequency Range (Mobile reception)	≤ 375 MHz	≤ 1.5 GHz	≤ 3 GHz	≤ 1.5 GHz
Number of subcarriers	1536	384	192	768
Symbol Duration	1000 μs	250 μs	125 μs	500 μs
Guard Duration	246 μs	62 μs	31 μs	123 μs
Total Symbol Duration	1246 μs	312 μs	156 μs	623 μs
Maximum Transmitter Separation for SFN	96 km	24 km	12 km	48 km

Table 1-1. DAB Transmission parameters for each transmission mode

Table 1-1 shows the system parameters for DAB. DAB has four transmission modes.

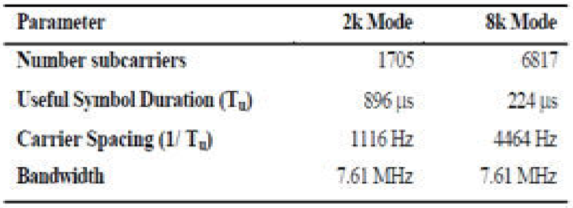
The transmission frequency, receiver velocity and required multipath tolerance all determine the most suitable transmission mode to use. Doppler spread is caused by rapid changes in the channel response due to movement of the receiver through a multipath environment. It results in random frequency modulation of the OFDM subcarriers, leading to signal degradation. The amount of Doppler spread is proportional to the transmission frequency and the velocity of movement. The closer the subcarriers are spaced together, the more susceptible the OFDM signal is to Doppler spread, and so the different transmission modes in DAB allow trade off between the amount of multipath protection (length of the guard period) and the Doppler spread tolerance.

The high multipath tolerance of OFDM allows the use of a Single Frequency Network (SFN), which uses transmission repeaters to provide improved coverage, and spectral efficiency. For traditional FM broadcasting, neighbouring cities must use different RF frequencies even for the same radio station, to prevent multipath causes by rebroadcasting at the same frequency. However, with DAB it is possible for the same signal to be broadcast from every area requiring coverage, eliminating the need for different frequencies to be used in neighbouring areas**.**

The data throughput of DAB varies from 0.6 - 1.8 Mbps depending on the amount of Forward Error Correction (FEC) applied. This data payload allows multiple channels to be broadcast as part of the one transmission ensemble. The number of audio channels is variable depending on the quality of the audio and the amount of FEC used to protect the signal. For telephone quality audio (24 kbps) up to 64 audio channels can be provided, while for CD quality audio (256 kb/s), with maximum protection, three channels are available.

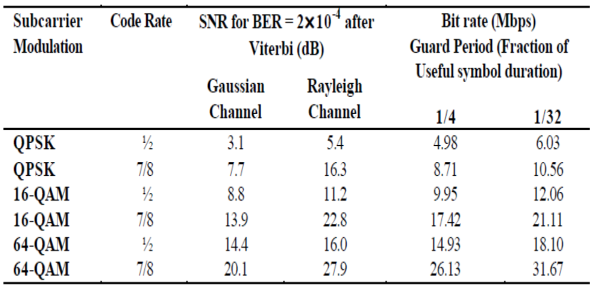
III. DIGITAL VIDEO BROADCASTING

The development of the Digital Video Broadcasting (DVB) standards was started in 1993 [14]. DVB is a transmission scheme based on the MPEG-2 standard, as a method for point to multipoint delivery of high quality compressed digital audio and video. It is an enhanced replacement of the analogue television broadcast standard, as DVB provides a flexible transmission medium for delivery of video, audio and data services [17]. The DVB standards specify the delivery mechanism for a wide range of applications, including satellite TV (DVB-S), cable systems (DVB-C) and terrestrial transmissions (DVB-T) [15]. The physical layer of each of these standards is optimised for the transmission channel being used. Satellite broadcasts use a single carrier transmission, with QPSK modulation, which is optimised for this application as a single carrier allows for large Doppler shifts, and QPSK allows for maximum energy efficiency [16]. This transmission method is however unsuitable for terrestrial transmissions as multipath severely degrades the performance of high-speed single carrier transmissions. For this reason, OFDM was used for the terrestrial transmission standard for DVB. The physical layer of the DVB-T transmission is similar to DAB, in that the OFDM transmission uses a large number of subcarriers to mitigate the effects of multipath. DVB-T allows for two transmission modes depending on the number of subcarriers used [18]. Table 1-2 shows the basic transmission parameters for these two modes. The major difference between DAB and DVB-T is the larger bandwidth used and the use of higher modulation schemes to achieve a higher data throughput. The DVB-T allows for three subcarrier modulation schemes: QPSK, 16-QAM (Quadrature Amplitude Modulation) and 64- QAM; and a range of guard period lengths and coding rates. This allows the robustness of the transmission link to be traded at the expense of link capacity. Table 1-3 shows the data throughput and required SNR for some of the transmission combinations. Lowered to meet DVB-T is a uni-directional link due to its broadcast nature. Thus any choice in data rate verses robustness affects all receivers. If the system goal is to achieve high reliability, the data rate must be the conditions of the worst receiver. This effect limits the usefulness of the flexible nature of the standard. However if these same principles of a flexible transmission rate are used in bi-directional communications, the data rate can be maximised based on the current radio conditions. Additionally for multiuser applications, it can be optimised for individual remote transceivers



Parameter	2k Mode	8k Mode
Number subcarriers	1705	6817
Useful Symbol Duration (T_u)	896 µs	224 µs
Carrier Spacing ($1/T_u$)	1116 Hz	4464 Hz
Bandwidth	7.61 MHz	7.61 MHz

Table 1-2, DVB transmission parameters



Subcarrier Modulation	Code Rate	SNR for BER = 2×10^{-4} after Viterbi (dB)		Bit rate (Mbps) Guard Period (Fraction of Useful symbol duration)	
		Gaussian Channel	Rayleigh Channel	1/4	1/32
QPSK	½	3.1	5.4	4.98	6.03
QPSK	7/8	7.7	16.3	8.71	10.56
16-QAM	½	8.8	11.2	9.95	12.06
16-QAM	7/8	13.9	22.8	17.42	21.11
64-QAM	½	14.4	16.0	14.93	18.10
64-QAM	7/8	20.1	27.9	26.13	31.67

Table 1-3, SNR required and net bit rate for a selection of the coding and modulation combinations for DVB

IV. MULTIUSER OFDM

DAB and DVB systems are only uni-directional from the base station to the users. Not much work has been done on using OFDM for two-way communications or for multiuser applications. These applications include wireless modems, Wireless Local Area Networks (WLAN's), Wireless Local Loop (WLL), mobile phones, and mobile high speed internet. This thesis aims to look at applying OFDM to such applications, and to look at the resulting advantages and problems. This thesis also presents some new techniques that can be used to improve broadcast and multiuser OFDM systems. The performance of adaptive modulation and adaptive user allocation schemes in a multiuser OFDM system. These techniques improve the spectral efficiency and QOS. Fattouche patented a method for implementing a wireless multiuser OFDM system in 1992, predating any published research in this field. This system used half duplex Time Division Multiplexing (TDM) to allow multiuser access, with the base stations and portable units taking turns to transmit. Carrier modulation was fixed and used D8-PSK (Differential 8 Phase Shift Keying). The system was bandwidth limited by using a raised cosine guard period. Fattouche is the founder WiLan Inc., which is one of the few companies currently producing multiuser OFDM modems. Williams and Prodan [82], patented the use of multiuser OFDM in cable applications in 1995. This introduced the use of a hybrid user allocation, using Frequency Division Multiplexing (FDM) and TDM. In this system the users were allocated time and frequency slots depending on the data demand. This patent however, fails to address problem of obtaining and maintaining accurate time and frequency synchronisation between users, which is critical for maintaining orthogonality between users. Cimini, Chuang, Sollenberger outlined an Advanced Cellular Internet Service using multiuser OFDM. The aim of this system was to provide Internet access at a data rate of 1 – 2 Mbps. This system uses time synchronised base stations, which are allocated time slots in a self-organising fashion. These base station time slots are then broken down in to time slots for users. In addition to TDM, users are allocated subcarriers dynamically based on the channel Signal to Interference Ratio (SIR), to allow minimisation of inter-cellular interference.

Wahlqvist described one possible implementation of multiuser OFDM in a wireless environment, outlining a user allocation scheme where users were allocated small blocks of time and frequency. In this scheme, each transmission block consists of a small group of subcarriers, (5 - 10) and a small number of symbols, about 11 in length. The aim of this structure is to allocate time and frequency slots to utilise the high correlation between neighbouring subcarriers, and the small channel variation between a small group of symbols. This allows the block to be characterised with a simple pilot tone structure.

V. CONCLUSION

The detail knowledge of a current key issue in the field of communications named Orthogonal Frequency Division Multiplexing (OFDM). We elaborated on the performance theory of the codes. First I developed an OFDM system model then try to improve the performance by applying forward error correcting codes to our uncoded system. From the study of the system, it can be concluded that we are able to improve the performance of uncoded OFDM by convolutional coding scheme.

REFERENCES

[1] Rajeshree Raut & Kishore Kulat, “SDR Design with Advanced Algorithms for Cognitive Radio”.

[2] Shanon Wichman, “Comparison of speech coding Algorithms ADPCM Vs CELP”.

[3] S. Shirani, “ Multimedia communications Sub-band coding”, McMaster Univ.

[4] Transform Coding”, from Wikipedia, the free

[5] Othman O. Khalifa, Sering Habib Harding and Aisha- Hassan A. Hashim, “ Compression using Wavelets transform”.

[6] Othman O. Khalifa, Sering Habib Harding Aisha-Hassan A. Hashim, “ Compression using wavelet transform”.

[7] MATLAB & Simulink on www.mathwork.com

[8] MATLAB 7.13. MATLAB and Simulink book.

[9] Aki Harma and Matti Karjalainen, Sept. 15, 2000 Feb 11, 2009, “ WARPTB-MATLAB Toolbox for Warped DSP”.

[10] “Software Defined Radio “,from Wikipedia, the free encyclopedia

[11] Jeffrey H. Reed, Charles W. Bostain, ”Presentation, Virginia Tech “, Bradly Dept. of Electrical & Computer Engg.

[12] “Cognitive Radio”, from Wikipedia, the free encyclopedia .

[13] Rajeshree Raut & Kishore Kulat, “ Software Defined Adaptive Codec for Cognitive Radio”,ISSN1109-2742, Vol.No 8, Dec 2009.

[14] Bin Le, Thomas W. Rondeau and Charles W. Bostain “Cognitive Radio Realities”.

[15] Bruce Fette Ph.D. Chief Scientist, General Dynamics C4 systems,” Implementing SDR Technologies with Matlab/Simulink”.

[16] W. T. Webb and R. Steele, "Variable Rate QAM for Mobile Radio," IEEE Trans. Comm., Vol. 43, pp. 2223-2230, July 1995.

[17] A. J. Goldsmith and S.-G. Chua, "Variable-rate Variable-Power MQAM for Fading Channels," IEEE Trans. Comm., Vol. 45, pp. 1218-1230, Oct. 1997.

[18] A. J. Goldsmith, S-G Chua, "Adaptive Coded Modulation for Fading Channels" IEEE Trans. Comm., Vol. 46, NO. 5, May 1998.

[19] P. Robertson and T. Worz, "Bandwidth Efficient Turbo Trellis-Coded Modulation using Punctured Component Codes" IEEE Journal on Selected Areas in Comms, Vol. JSAC-16, pp. 206- 218 Feb. 1998.

[20] S. Le Goff, A. Glavieux, and C. Berrou, "Turbo-Codes and High Spectral Efficiency Modulation," Proc. ICC'94, pp. 645-649, May 1994.

[21] U. Wachsmann and J. Huber, "Power and Bandwidth Efficient Digital Communication using Turbo Codes in Multilevel Codes," European Trans. Telecommun., Vol. 6, no. 5, 1995.

[22] L. Hanzo, W. Webb, and T. Keller, "Singleand Multi-Carrier Quadrature Amplitude Modulation", Chichester: Wiley/IEEE Press, 1999.

[23] Sigen Ye, Rick..Blum, "Adaptive Modulation for Variable-Rate OFDM System with Imperfect Channel Information ", Proc of Int. Conf. on Comm., Vol 6, pp 1861-1865, June 2001.

[24] T. Keller, L. Hanzo "Unscheduled Report for the Median Project ", Technical Report, Dept. of ECS, University of Southampton, 1998.

[25] P. Robertson "An overview of bandwidth Efficient Turbo Coding Schemes", Proc. Int. Symposium on Turbo Codes and Related Topics, Brest, France, pp. 103 -110, Sept. 1997.

[26] L. Piazzo," TTCM-OFDM over Wideband Fading Channels" TMR - Marie Curie Grant, the final scientific report.

[27] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate," IEEE Trans. Inform. Theory, Vol. IT-20, pp. 284-287, Mar. 1974.

[28] A. J. Goldsmith and S. Chua, "Adaptive Coded Modulation for Fading Channels," IEEE Trans. on Comm, Vol. 46, no. 5, pp.595-602, May 1998.

[29] A. J. Goldsmith and S.-G. Chua, "Variablerate Variable-power MQAM for Fading Channels," IEEE Trans. on Commun., Vol. 45, pp. 1218-1230, Oct. 1997.

[30] L. Hanzo, T. H. Liew, B. I. Yeap, "Turbo Coding Turbo Equalization and Sparse Time Coding for Transmission over Fading Channels", John Wiley & Sons, Chichester, 2002.

[31] Pinar Ormeci, Xueting Liu, Dennis L. Goeckel, and Richard D. Wesel, "Adaptive Bit-Interleaved Coded Modulation", IEEE Trans. Comm., Vol. 49, No. 9, Sept. 2001.

◈◈◈

Facilitate Traffic Allocation Among Multiple Paths Based on Empirical Jamming Statistics

B. Sreekanth & P. V. Naga Jayudu

Intell Engineering College, Anantapur

Abstract - Multiple-path source routing protocols allow a data source node to distribute the total traffic among available paths.In this article, we consider the problem of jamming-aware source routing in which the source node performs traffic allocation based on empirical jamming statistics at individual network nodes. We formulate this traffic allocation as a lossy network flow optimization problem using portfolio selection theory from financial statistics. We show that in multi-source networks, this centralized optimization problem can be solved using a distributed algorithm based on decomposition in network utility maximization (NUM). We demonstrate the network's ability to estimate the impact of jamming and incorporate these estimates into the traffic allocation problem. Finally, we simulate the achievable throughput using our proposed traffic allocation method in several scenarios.

Keywords - *Jamming, Multiple path routing, Portfolio selection theory, Optimization, Network utility maximization.*

I. INTRODUCTION

Jamming point-to-point transmissions in a wireless mesh network [1] or underwater acoustic network [2] can have debilitating effects on data transport through the network. The effects of jamming at the physical layer resonate through the protocol stack, providing an effective denial-of-service (DoS) attack [3] on end-to-end data communication. The simplest methods to defend a network against jamming attacks comprise physical layer solutions such as spread-spectrum or beamforming, forcing the jammers to expend a greater resource to reach the same goal. However, recent work has demonstrated that intelligent jammers can incorporate cross layer protocol information into jamming attacks, reducing resource expenditure by several orders of magnitude by targeting certain link layer and MAC implementations [4]–[6] as well as link layer error detection and correction protocols [7]. Hence, more sophisticated anti-jamming methods and defensive measures must be incorporated into higher-layer protocols, for example channel surfing [8] or routing around jammed regions of the network [6].

In order to characterize the effect of jamming on throughput,each source must collect information on the impact of the jamming attack in various parts of the network. However, the extent of jamming at each network node depends on a number of unknown parameters, including the strategy used by the individual jammers and the relative location of the jammers with respect to each transmitter-receiver pair. Hence, the impact of jamming is probabilistic from the perspective of the network1, and the characterization of the jamming impact is further complicated by the fact that the jammers' strategies may be dynamic and the jammers themselves may be mobile.

In order to capture the non-deterministic and dynamic effects of the jamming attack, we model the packet error rate at each network node as a random process. At a given time, the randomness in the packet error rate is due to the uncertainty in the jamming parameters, while the time-variability in the packet error rate is due to the jamming dynamics and mobility. Since the effect of jamming at each node is probabilistic, the end-to-end throughput achieved by each source-destination pair will also be non-deterministic and, hence, must be studied using a stochastic framework.

In this article, we thus investigate the ability of network nodes to characterize the jamming impact and the ability of multiple source nodes to compensate for jamming in the allocation of traffic across multiple routing paths. Our contributions to this problem are as follow:

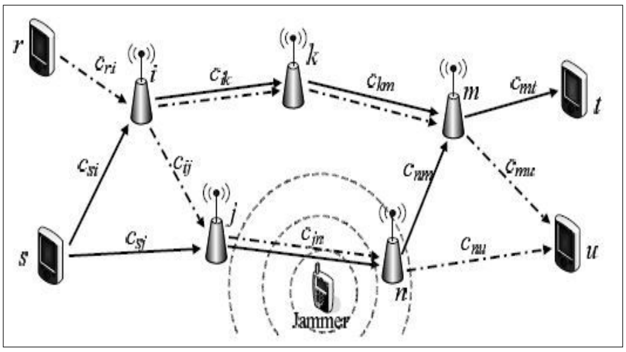


Fig. 1 : An example network with sources S = {r, s} is illustrated. Each unicast link (i, j) ! E is labeled with the corresponding link capacity.

- We formulate the problem of allocating traffic across multiple routing paths in the presence of jamming as a lossy network flow optimization problem. We map the optimization problem to that of asset allocation using portfolio selection theory [12], [13].
- We formulate the centralized traffic allocation problem for multiple source nodes as a convex optimization problem.
- We show that the multi-source multiple-path optimal traffic allocation can be computed at the source nodes using a distributed algorithm based on decomposition in network utility maximization(NUM) [14].
- We propose methods which allow individual network nodes to locally characterize the jamming impact and aggregate this information for the source nodes.
- We demonstrate that the use of portfolio selection theory allows the data sources to balance the expected data throughput with the uncertainty in achievable traffic rates.

The remainder of this article is organized as follows. In Section II, we state the network model and assumptions about the jamming attack. To motivate our formulation, in Section III, we present methods that allow nodes to characterize the local jamming impact. These concepts are required to understand the traffic allocation optimization and the mapping of this problem to Portfolio selection. In Section IV, we formulate the optimal multiple path traffic allocation problem for multisource networks. In Section V, we evaluate the performance of the optimal traffic allocation formulation. We summarize our contributions in Section VI.

II. SYSTEM MODEL & ASSUMPTIONS

The wireless network of interest can be represented by a directed graph G = (N, E). The vertex set N represents the network nodes, and an ordered pair (i, j) of nodes is in the edge set E if and only if node j can receive packets directly from node i. We assume that all communication is unicast over the directed edges in E, i.e. each packet transmitted by node i ! N is intended for a unique node j ! N with (i, j) ! E. The maximum achievable data rate, or capacity, of each unicast link (i, j) ! E in the absence of jamming is denoted by the predetermined constant rate cij in units of packets per second3. Each source node s in a subset S " N generates data for a single destination node ds ! N. We assume that each source node s constructs multiple routing paths to ds using a route request process similar to those of the DSR [9] or AODV [10] protocols.We let Ps = {ps1, . . . , psLs} denote the collection of Ls loop-free routing paths for source s, noting that these paths need not be disjoint as in MP-DSR [11].

Representing each path ps! by a subset of directed link set E, the sub-network of interest to source s is given by the directed subgraph of the graph G.

$$G_s = \left(\mathcal{N}_s = \bigcup_{\ell=1}^{L_s} \{ j : (i,j) \in p_{s\ell} \}, \mathcal{E}_s = \bigcup_{\ell=1}^{L_s} p_{s\ell} \right)$$

Figure 1 illustrates an example network with sources S ={r, s}. The subgraph Gr consists of the two routing paths pr1 = {(r, i), (i, k), (k,m), (m, u)} pr2 = {(r, i), (i, j), (j, n), (n, u)}, and the subgraph Gs consists of the two routing paths ps1 = {(s, i), (i, k), (k,m), (m, t)} ps2 = {(s, j), (j, n), (n,m), (m, t)}.

In this article, we assume that the source nodes in S have no prior knowledge about the jamming attack being performed. That is, we make no assumption about the jammer's goals, method of attack, or mobility patterns. We assume that the number of jammers and their locations are unknown to the network nodes. Instead of relying on direct knowledge of the jammers, we suppose that the network nodes characterize the jamming impact in terms of the empirical packet delivery rate. Network nodes can then relay the relevant information to the source nodes in order to assist in optimal traffic allocation.

Each time a new routing path is requested or an existing routing path is updated, the responding nodes along the path will relay the necessary parameters to the source node as part of the reply message for the routing path. Using the information from the routing reply, each source node s is thus provided with additional information about the jamming impact on the individual nodes.

III. CHARACTERIZING THE IMPACT OF JAMMING

In this section, we propose techniques for the network nodes to estimate and characterize the impact of jamming and for a source node to incorporate these estimates into its traffic allocation. In order for a source node s to incorporate the jamming impact in the traffic allocation problem, the effect of jamming on transmissions over each link (i, j) ! Es must be estimated and relayed to s. However, to capture the jammer mobility and the dynamic effects of the jamming attack, the local estimates need to be continually updated. We begin with an example to illustrate the possible effects of jammer mobility on the traffic allocation problem and motivate the use of continually updated local estimates

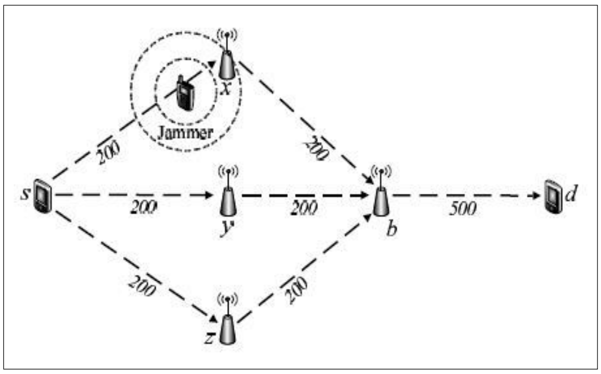


Fig. 2 : An example network that illustrates a single-source network with three routing paths. Each unicast link (i, j) is labeled with the corresponding link capacity cij in units of packets per second. The proximity of the jammer to nodes x and y impedes packet delivery over the corresponding paths, and the jammer mobility affects the allocation of traffic to the three paths as a function of time.

A. Illustrating the Effect of Jammer Mobility on Network Throughput:

Figure 2 illustrates a single-source network with three routing paths

p1 = {(s, x), (x, b), (b, d)}, p2 = {(s, y), (y, b), (b, d)} and p3 = {(s, z), (z, b), (b, d)}.

The label on each edge (i, j) is the link capacity cij indicating the maximum number of packets per second (pkts/s) which can be transported over the wireless link. In this example, we assume that the source is generating data at a rate of 300 pkts/s. In the absence of jamming, the source can continuously send 100 pkts/s over each of the three paths,yielding a throughput rate equal to the source generation rate of 300 pkts/s. If a jammer near node x is transmitting at high power, the probability of successful packet reception, referred to as the packet success rate, over the link (s, x) drops to nearly zero, and the traffic flow to node d reduces to 200 pkts/s. If the source node becomes aware of this effect,the allocation of traffic can be changed to 150 pkts/s on each of paths p2 and p3, thus recovering from the jamming attack at node x. The relay of information from the nodes can be done periodically or at the instants when the packet success rates change significantly. These updates must be performed at a rate comparable to the rate of the jammer movement to provide an effective defense against the mobile jamming attack.

Next, suppose the jammer continually changes position between nodes x and y, causing the packet success rates over links (s, x) and (s, y) to oscillate between zero and one. This behavior introduces a high degree of variability into the observed packet success rates, leading to a less certain estimate of the future success rates over the links (s, x) and (s, y).

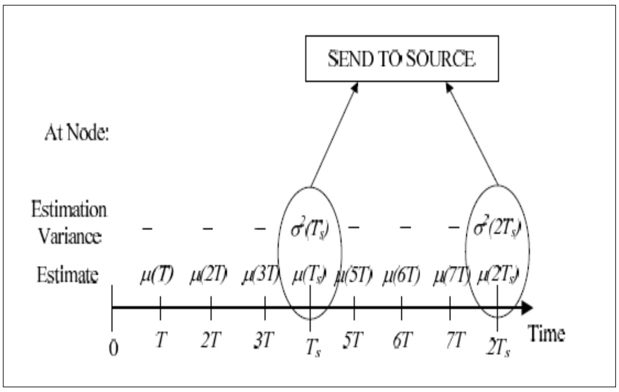


Fig. 3 : The estimation update process is illustrated for a single link. The estimate μij (t) is updated every T seconds, and the estimation variance σ^2_{ij} (t) is computed only every Ts seconds. Both values are relayed to relevant source nodes every Ts seconds.

This solution takes into account the historic variability in the packet success rates due to jamming mobility. In the following section, we build on this example, providing a set of parameters to be estimated by network nodes and methods for the sources to aggregate this information and characterize the available paths on the basis of expected throughput.

B. Estimating End-to-End Packet Success Rates

Given the packet success rate estimates μ_{ij} (t) and σ^2_{ij} (t) for the links (i, j) in a routing path psℓ, the source s needs to estimate the effective end-to-end packet success rate to determine the optimal traffic allocation. Assuming the total time required to transport packets from each source s to the corresponding destination ds is negligible compared to the update relay period Ts, we drop the time index and address the end-to-end packet success rates in terms of the estimates μ_{ij} and σ^2_{ij} . The end-to-end packet success rate ys! for path ps! can be expressed as the product

$$y_{s\ell} = \prod_{(i,j)\in p_{s\ell}} x_{ij}, \qquad (5)$$

which is itself a random variable6 due to the randomness in each xij. We let $s! denote the expected value of ys! and %s!m denote the covariance of ys! and ysm for paths ps!, psm ! Ps. Due to the computational burden associated with in-network inference of correlation between estimated random variables, we let the source node s assume the packet success rates xij as mutually independent, even though they are likely correlated. We maintain this independence assumption throughout this work, yielding a feasible approximation to the complex reality of correlated random variables, and the case of in-network inference of the relevant correlation is left as future work. Under this independence assumption, the mean $s! of ys! given in (5) is equal to the product of estimates µij as

$$\gamma_{s\ell} = \prod_{(i,j)\in p_{s\ell}} \mu_{ij}, \qquad (6)$$

And the covariance $\omega_{s\ell m} = E[y_{s\ell} y_{sm}] - E[y_{s\ell}]E[y_{sm}]$ is similarly given by

$$\omega_{s\ell m} = \prod_{(i,j)\in p_{s\ell}\oplus p_{sm}} \mu_{ij} \prod_{(i,j)\in p_{s\ell}\cap p_{sm}} \left(\sigma_{ij}^2 + \mu_{ij}^2\right) - \gamma_{s\ell}\gamma_{sm}. \qquad (7)$$

In (7), (denotes the exclusive-OR set operator such that an element is in A (B if it is in either A or B but not both.The covariance formula in (7) reflects the fact that the end to- end packet success rates ys! and ysm of paths ps! and psm with shared links are correlated even when the rates xij are independent. We note that the variance %2 s! of the end-to-end rate ys! can be computed using (7) with & = m.

Letting 's! denote the traffic rate allocated to path ps! by the source node s, the problem of interest is thus for each source s to determine the optimal Ls×1 rate allocation vector "s subject to network flow capacity constraints using the available statistics !s and !s of the end-to-end packet success rates under jamming.

IV. A. Traffic Allocation Constraints

In order to define a set of constraints for the multiple-path traffic allocation problem, we must consider the source data rate constraints, the link capacity constraints, and the reduction of traffic flow due to jamming at intermediate nodes.

Due to jamming at nodes along the path, the traffic rate is potentially reduced at each receiving node as packets are lost. The capacity constraint on the total traffic traversing a link (i, j) thus imposes the stochastic constraint

$$\sum_{s\in\mathcal{S}} \sum_{\ell:(i,j)\in p_{s\ell}} \phi_{s\ell} y_{s\ell}^{(i)} \le c_{ij} \qquad (8)$$

on the feasible allocation vectors "s. The element w((i, j), ps!) in row (i, j) and column ps! of Ws is thus given by

$$w\left((i,j),p_{s\ell}\right) = \begin{cases} \min\left\{1, \gamma_{s\ell}^{(i)} + \delta\omega_{s\ell}^{(i)}\right\}, & \text{if } (i,j)\in p_{s\ell} \\ 0, & \text{otherwise.} \end{cases} \qquad (9)$$

Letting c denote the |E| × 1 vector of link capacities cij for (i, j) ! E, the link capacity constraint in (8) including expected packet loss due to jamming can be expressed by the vector inequality

$$\sum_{s\in\mathcal{S}} \boldsymbol{W}_s \phi_s \le \boldsymbol{c}, \qquad (10)$$

which is a linear constraint in the variable "s. We note that this statistical constraint formulation generalizes the standard network flow capacity constraint corresponding to the case of xij = 1 for all (i, j) ! E in which the incidence matrix Ws is deterministic and binary.

B. Optimal Traffic Allocation Using Portfolio Selection Theory

In order to determine the optimal allocation of traffic to the paths in Ps, each source s chooses a utility function Us("s) that evaluates the total data rate, or throughput, successfully delivered to the destination node ds. In defining our utility function Us("s), we present an analogy between traffic allocation to routing paths and allocation of funds to correlated assets in finance.

In Markowitz's portfolio selection theory [12], [13], an investor is interested in allocating funds to a set of financial assets that have uncertain future performance. We describe the desired analogy by mapping this allocation of funds to financial assets to the allocation of traffic to routing paths.We relate the expected investment return on the financial portfolio to the estimated end-to-end success rates !s and the investment risk of the portfolio to the estimated success rate covariance matrix !s. We note that the correlation between related assets in the financial portfolio corresponds to the correlation between non-disjoint routing paths. The analogy between financial portfolio selection and the allocation of traffic to routing paths is summarized below.

Portfolio Selection	Traffic Allocation
Funds to be invested	Source data rate R_s
Financial assets	Routing paths $\mathcal{P}_s$
Expected Asset return	Expected Packet success rate $\gamma_{s\ell}$
Investment portfolio	Traffic allocation ϕ_s
Portfolio return	Mean throughput $\gamma_s^T \phi_s$
Portfolio risk	Estimation variance $\phi_s^T \Omega_s \phi_s$

As in Markowitz's theory, we define a constant risk-aversion factor ks * 0 for source s ! S to indicate the preference for source s to allocate resources to less risky paths with lower throughput variance. This risk-aversion constant weighs the trade-off between expected throughput and estimation variance. For a given traffic rate allocation vector "s, the expected total throughput for source s is equal to the vector inner product !Ts "s. Based on the above analogymaking use of portfolio selection theory, we define the utility function Us("s) at source s as the weighted sum

$$U_s(\phi_s) = \gamma_s^T \phi_s - k_s \phi_s^T \Omega_s \phi_s \tag{11}$$

Setting the risk-aversion factor ks to zero indicates that the source s is willing to put up with any amount of uncertainty in the estimate !s of the end-to-end success rates to maximize the expected throughput. Combining the utility function in (11) with the set of constraints defined in Section IV-A yields the following jamming aware traffic allocation optimization problem which aims to find the globally optimal traffic allocation over the set S of sources.

Optimal Jamming-Aware Traffic Allocation

$$\begin{aligned} \phi^* = \; & \arg\max_{\{\phi_s\}} \sum_{s\in\mathcal{S}} \gamma_s^T \phi_s - k_s \phi_s^T \Omega_s \phi_s \\ \text{s.t.} \; & \sum_{s\in\mathcal{S}} W_s \phi_s \le c \\ & 1^T \phi_s \le R_s \text{ for all } s \in \mathcal{S}, \\ & 0 \le \phi_s \text{ for all } s \in \mathcal{S}. \end{aligned} \tag{12}$$

Since the use of centralized protocols for source routing may be undesirable due to excessive communication overhead in large-scale wireless networks, we seek a distributed formulation for the optimal traffic allocation problem in (12).

C. Optimal Distributed Traffic Allocation using NUM

In the distributed formulation of the algorithm, each source s determines its own traffic allocation "s, ideally with minimal message passing between sources. By inspection, we see that the optimal jamming-aware flow allocation problem in (12) is similar to the network utility maximization (NUM) formulation of the basic maximum network flow problem [14]. We thus develop a distributed traffic allocation algorithm using Lagrangian dual decomposition techniques [14] for NUM.

The dual decomposition technique is derived by decoupling the capacity constraint in (10) and introducing the link prices)ij corresponding to each link (i, j). Letting # denote the |E| × 1 vector of link prices)ij, the Lagrangian L(", #) of the optimization problem in (12) is given by

$$L(\phi,\lambda) = \sum_{s\in\mathcal{S}} \gamma_s^T \phi_s - k_s \phi_s^T \Omega_s \phi_s + \lambda^T \left(c - \sum_{s\in\mathcal{S}} W_s \phi_s \right) \tag{13}$$

The distributed optimization problem is solved iteratively using the Lagrangian dual method as follows. For a given set of link prices #n at iteration n, each source s solves the local optimization

$$\phi_{s,n}^* = \arg\max_{\phi_s \in \Phi_s} \left(\gamma_s^T - \lambda_n^T W_s \right) \phi_s - k_s \phi_s^T \Omega_s \phi_s. \tag{14}$$

problem The link prices #n+1 are then updated using a gradient descent iteration as

$$\lambda_{n+1} = \left(\lambda_n - a \left(c - \sum_{s\in\mathcal{S}} W_s \phi_{s,n}^* \right) \right)^+ \tag{15}$$

where a > 0 is a constant step size and (v)+ = max(0, v) is the element-wise projection into the non-negative orthant. In order to perform the local update in (15), sources must exchange information about the result of the local optimization step. Since updating the link prices # depends only on the expected link usage, sources must only exchange the |E| × 1 link usage vectors $u_{s,n} = W_s \phi_{s,n}^*$ to ensure that the link prices are consistently updated across all sources. The iterative optimization step can be repeated until the allocation vectors "s converge8 for all sources s ! S, i.e. when $\|\phi_{s,n}^* - \phi_{s,n-1}^*\| \le \epsilon$ for all s with a given * > 0. The above approach yields the following distributed algorithm for optimal jamming-aware flow allocation.

Distributed Jamming-Aware Traffic Allocation

```
Initialize $n = 1$ with initial link prices $\lambda_1$.
1.  Each source $s$ independently computes
        $\phi_{s,n}^* = \arg\max_{\phi_s \in \Phi_s} \left( \gamma_s^T - \lambda_n^T W_s \right) \phi_s - k_s \phi_s^T \Omega_s \phi_s.$
2.  Sources exchange the link usage vectors
        $u_{s,n} = W_s \phi_{s,n}^*.$
3.  Each source locally updates link prices as
        $\lambda_{n+1} = \left( \lambda_n - a \left( c - \sum_{s\in\mathcal{S}} u_{s,n} \right) \right)^+.$
4.  If $\|\phi_{s,n}^* - \phi_{s,n-1}^*\| > \epsilon$ for any $s$,
        increment $n$ and go to step 1.
```

V. PERFORMANCE EVALUATION

In this section, we simulate various aspects of the proposed techniques for estimation of jamming impact and jamming aware traffic allocation. We first describe the simulation setup, including descriptions of the assumed models for routing path construction, jammer mobility, packet success rates, and estimate updates. We then simulate the process of computing the estimation statistics $\mu_{ij}(t)$ and $\sigma^2_{ii}(t)$ for a single link (i, j). Next, we illustrate the effects of the estimation process on the throughput optimization, both in terms of optimization objective functions and the resulting simulated throughput. Finally, we simulate a small-scale network similar to that in Figure 2 while varying network and protocol parameters in order to observe performance trends.

VI. CONCLUSION

In this article, we studied the problem of traffic allocation in multiple-path routing algorithms in the presence of jammers whose effect can only be characterized statistically.We formulated multiple-path traffic allocation in multi-source networks as a lossy network flow optimization problem using an objective function based on portfolio selection theory from finance. We showed that this centralized optimization problem can be solved using a distributed algorithm based on decomposition in network utility maximization (NUM). We have thus shown that multiple path source routing algorithms can optimize the throughput performance by effectively incorporating the empirical jamming impact into the allocation of traffic to the set of paths.

◈◈◈

An Analysis of Rigid Image Alignment Computer Vision Algorithms

Rajeshree Joshi & Robert Cook

College of IT, Georgia Southern University, Statesboro, Georgia, USA
E-mail : rajoshi@georgiahealth.edu, bobcook@georgiasouthern.edu

Abstract - Image registration is one of the techniques used in the computer vision field to transform different sets of data into one coordinate system to align images. Registration is important in order to be able to compare or integrate the data obtained from multiple measurements. Rigid image alignment is a type of image registration technique used to align two two-dimensional images into a common coordinate system based on two transformation parameters, translation and rotation.

In our research study, we are analyzing the accuracy of registering images using two rigid image alignment algorithms, namely the Principal Axes algorithm and the Fast Fourier Transform (FFT) based phase correlation algorithm. The software for registering images using these two methods is written in MATLAB R2011a. We also compared our results with alignments achieved for the same images using an existing Statistical Parametric Mapping (SPM8) package for registration. Our registration software is based on work with images acquired from a Magnetic Resonance Imaging (MRI) scanner and especially for images taken of a quality assurance (QA) phantom. A QA phantom is used to test the quality of images acquired by measuring different QA parameters on images acquired over a period of time. By comparing future phantom images with the first image in the series, we can perform a series of Quality Assurance steps to measure any degradation in the MRI device. The QA results can then be used to apply inverse transformations to new customer images to improve their quality. The first step in the QA process is image registration, which is the topic of this paper.

Similarity measures such as Mean Square Error (MSE), translation and rotation errors are computed to derive at the accuracy extent of the registration algorithms. Our analysis shows that the Principal Axes method can successfully register 17 of the 22 non-aligned test images, the FFT method registered 21 test images successfully whereas SPM8 with default settings showed correct alignments for only 9 images in our case study as per our requirement. The Principal Axes algorithm performed better image alignment when the two images were displaced by a larger distance, and the FFT based algorithm performed better for larger rotation angle differences among images. Hence, we conclude that our algorithms have the potential for inclusion in the new QA process.

Keywords - *Rigid image alignment; Computer vision algorithms; Image registration; Principal axes; Fast fourier transform; FFT; Mean square error; MSE.*

I. INTRODUCTION

Computer vision embeds the core technology of automated image analysis which is used in many fields. Medical computer vision or medical image processing is one of the most prominent application fields of computer vision [12]. Rigid body registration methods are effectively used for registering human brain images from MRI [1, 2]. Multiple images captured from different viewpoints or at different times, get distorted with respect to each other. Image registration or image alignment is the process of determining the optimal transformation matrix that results in the images being in spatial alignment [1]. The images need to be geometrically aligned for better observation [6]. This procedure of mapping points from one image to corresponding points in another image is called Image Registration. It is a spatial transform [6]. A rigid-body transformation in two dimensions is defined by two parameters, translation and rotation [2].

Principal Axes method is a spatial, feature-based monomodal rigid image registration method for aligning two images. The Principal Axes algorithm acts upon the features of the images, such as edges, corners, or circular patterns as its feature space. The search space consists of global translations and rotations. The search strategy is finding the closed formed solution based on the eigenvalue decomposition of a certain covariance matrix [1]. The similarity metric is the variance of the projection of the feature's location vector onto the principal axis [1]. The principal axes are the orthogonal axes about which the moments of inertia are minimized. If two objects are identical except for a translation and a rotation, then they can be registered by coinciding their principal axes [6]. The algorithm is suitable for registering images shaped like an ellipse or ellipsoid. For purposes of image registration, the critical features of an ellipse are its center of mass, and principal orientations, i.e., major and minor axes [1]. The principal axes algorithm is easy to implement, and efficient but it does have the shortcoming that it is sensitive to missing data, if any.

The Fast Fourier Transform (FFT) based registration is a frequency-domain type automatic rigid registration method. The feature space it uses consists of all the pixels in the image, and its search space covers all global translations and rotations [1]. The search strategy is the closed form Fourier-based method, and the similarity metric is correlation, and its variants, e.g., phase only correlation [1]. The FFT-based automatic registration method relies on the Fourier shift theorem which guarantees that the phase of a specially defined "ratio" is equal to the phase difference between the images [7]. The Fourier-Mellin transform [7] has been implemented in our algorithm to register images that are misaligned due to translation and rotation. The Fourier–Mellin registration method is based on the principle of phase correlation and the properties of Fourier analysis [9]. The phase correlation finds the translation between two images. The Fourier–Mellin transform extends phase correlation to handle images transformed by both translation and rotation [9].A Fourier transform is applied to images to recover translation [9]. Then a log-polar transformation is applied to the magnitude spectrum and the rotation angle is recovered by using phase correlation in the log-polar space [9]. By operating on the magnitude spectrum of an image, the translational differences between the images are avoided since the magnitude spectrum of an image and its translated counterpart are identical and only their phase spectrum are different [9]. The log-polar transformation causes rotation to be manifested as translation, whereby phase correlation can be applied to recover the rotation angle between the pair of input images [9].

II. RESEARCH PLAN

The research work involved implementing the following steps:

a. Acquiring multiple QA phantom images at different times by positioning the phantom in the MRI scanner. There are a total of 23 images used in our study; the first one is used as the reference image and the remaining 22 are target images to be aligned with the reference image. The target images were created in ImageJ by applying different transformation parameters on the reference image to test our algorithms

b. Converting the raw images to NIFTI file format using ImageJ software to be used for alignment testing using the existing SPM8 package. Use the tool 'Realign' with default settings to align the images. Check the alignment with 'Check Registration' tool.

c. Designing rigid image registration software in MATLAB R2011a. The user is given the choice of using either Principal Axes method or FFT based phase correlation method for registration.

d. Finding the alignment similarity between the two images by showing the user the difference between the two images. Commonly used similarity measure Mean Square Error (MSE) is also calculated. Translation and rotation errors are calculated for registrations achieved using the Principal Axes method.

e. Saving the transformed aligned image as a raw image for further processing by the user.

f. Analyzing the results and comparing the algorithm efficiency using the similarity measures.

III. EXPERIMENTAL SETUP

Our application demands registering images acquired from a quality-assurance (QA) phantom using an MRI scanner. The phantom is designed to test the quality of images acquired from the scanner over a time span. Image quality can be tested based on different QA parameters measured for the images. Before this processing can be done, all the acquired images obtained from the scanner at different times need to be aligned with a reference image to facilitate correct measurements for QA parameters. The images to be aligned are from the same phantom but acquired at different times. Figure 1 shows the experimental setup for acquiring images using the MRI scanner and the QA phantom. Bruker BioSpin 7T MRI scanner is used for acquiring images by placing the QA phantom on the scanner bed. The central disk in the phantom is used for resolution, contrast, and distortion measurements. Bruker ParaVision 5.1 software is used to instruct the scanner to acquire the images.

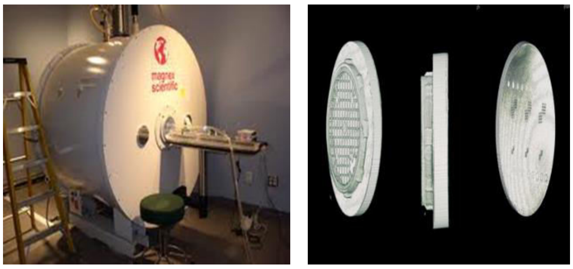


FIG. 1 : EXPERIMENTAL SETUP

The phantom images acquired using the scanner at different times are geometrically distorted by rotation and translation. This is due to different positioning of the phantom in the scanner each time. The circular patterns shown in the images need to be at similar coordinates for all the images for further meaningful processing. The alignment of images then is crucial in deriving correct QA parameter values to test image quality and therefore the proper functioning of the scanner. Misalignments will result in wrong measurements for QA parameters providing us with incorrect results on the scan quality of images and the functioning of the scanner. Since the scanner magnet may deteriorate with time in the long run, it is possible that the images acquired can get distorted with more noise introduced in the future images. To tackle this problem, we have simulated two of our target images by introducing Gaussian noise enough to pixelate the images using the ImageJ (NIH, MD) [14] software.

For registering two images, the user is required to choose two files. The first file is called the reference file to which the misaligned or target file will be aligned to. The file format supported is the raw binary format. The images derived from the QA phantom are unsigned 16-bit images in big-endian format. The file size is 384x384 ie. the images are square images. Each image is made up of three slices, collectively representing the volume of the phantom. The first slice shows only the boundary of the phantom. Second slice shows the varying intensity circular patterns from the phantom disk. And the third slice shows a circular disk with no patterns. Figure 2 shows the complete phantom image.

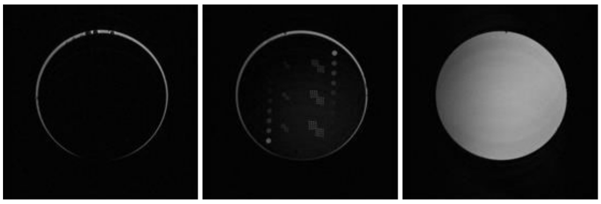
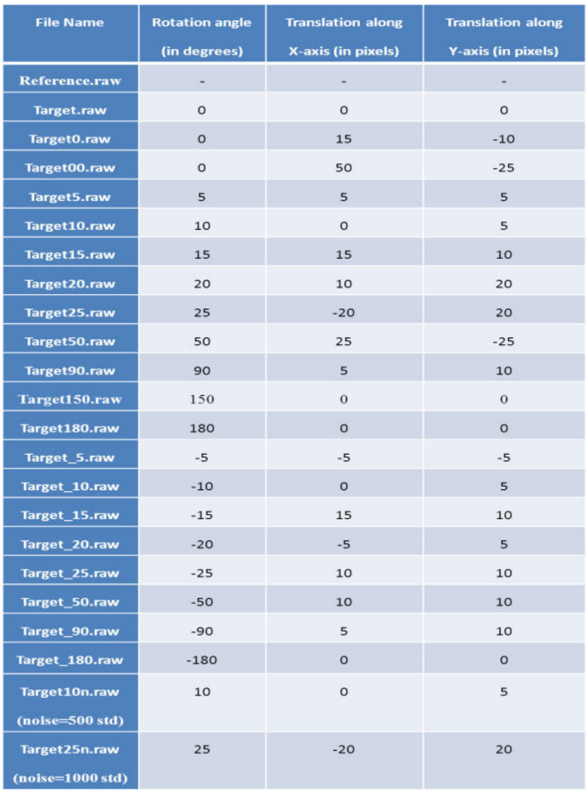


FIG. 2. QA PHANTOM IMAGE

Table I shows the file names and their initial transformation parameters.

Table I : Data file names and the initial transformations



File Name	Rotation angle (in degrees)	Translation along X-axis (in pixels)	Translation along Y-axis (in pixels)
Reference.raw	-	-	-
Target.raw	0	0	0
Target0.raw	0	15	-10
Target00.raw	0	50	-25
Target5.raw	5	5	5
Target10.raw	10	0	5
Target15.raw	15	15	10
Target20.raw	20	10	20
Target25.raw	25	-20	20
Target50.raw	50	25	-25
Target90.raw	90	5	10
Target150.raw	150	0	0
Target180.raw	180	0	0
Target_5.raw	-5	-5	-5
Target_10.raw	-10	0	5
Target_15.raw	-15	15	10
Target_20.raw	-20	-5	5
Target_25.raw	-25	10	10
Target_50.raw	-50	10	10
Target_90.raw	-90	5	10
Target_180.raw	-180	0	0
Target10n.raw (noise=500 std)	10	0	5
Target25n.raw (noise=1000 std)	25	-20	20

The target images are transformed using ImageJ to provide varying displacements and rotations in images for testing our algorithms. The Rotate and Translate tools help perform these transformations on the target files. Gaussian noise is introduced in two of the target images to pixelate them to test for noise sensitivity of our algorithms. The Noise-Add Specific Noise tool in ImageJ was used to introduce Gaussian noise of 500 and 1000 standard deviation in target files target10n and target25n respectively.

IV. PRINCIPAL AXES METHOD

A. Algorithm

Steps involved for Principal Axes based registration are:

i) Perform feature detection by thresholding the images to only detect the circular patterns on the phantom image.

   a) Find the maximum intensity values in both images.

   b) Plot a histogram of these intensities into 10 bins.

   c) Look for the intensity for second highest number of pixels in the bins. The highest number of pixels belong to the image background and second highest to the patterns in the image.

   d) Avoid any background pixels by fixing the cutoff for feature selection to one tenth of the above number.

   e) Set all the pixel intensities below this cutoff value to zeros in both images. Let I1F and I2F be the two images with features selected.

   f) Calculate the centers of both the images with features.

ii) Calculate rotation parameter.

   a) Calculate the center of mass or centroid (x', y') for I1F and I2F using the formulae

      $X' = \sum x, y * x * I(x, y) / \sum x, y * I(x, y)$ and

      $Y' = \sum x, y * y * I(x, y) / \sum x, y * I(x, y)$

      where I(x, y) is the intensity at location (x, y).

   b) Find difference between the two centroid positions.

   c) Find the eigenvectors of the reference image and target image via an eigenvalue decomposition of the covariance matrices. Covariance matrix C can be written as C = (c11 c12 c21 c22),

      where c11 = ?x, y (x-x')2 * I(x, y),

c22 = ?x, y (y-y')2 * I(x, y),

c12 = ?x, y (x-x') * (y-y') * I(x, y) and

c21 = c12.

d) For each image, determine the angle *angl*, the maximum eigenvector makes with the horizontal x-axis using the formula angl = atan2(V(2,1),V(1,1))*180/pi

where V is the maximum eigenvector.

e) Find the difference between the two angles, θ.

f) If θ<= -90.0 and θ> -180.0 then θ= θ+ 180.0

elseif θ<= 180.0 then θ= θ+ 360.0

iii) Perform Rotation transformation on target image.

a) Construct the rotation matrix as

R = [cos(-θ) sin(-θ); -sin(-θ) cos(-θ)];

b) Multiply the rotation matrix with the (x, y) coordinates of the target image to get the new rotated image.

iv) Calculate translation parameters along x-axis and y-axis.

a) Find the center of mass of rotated target image.

b) Calculate the difference in locations of the two centroids.

v) Perform translation transformation.

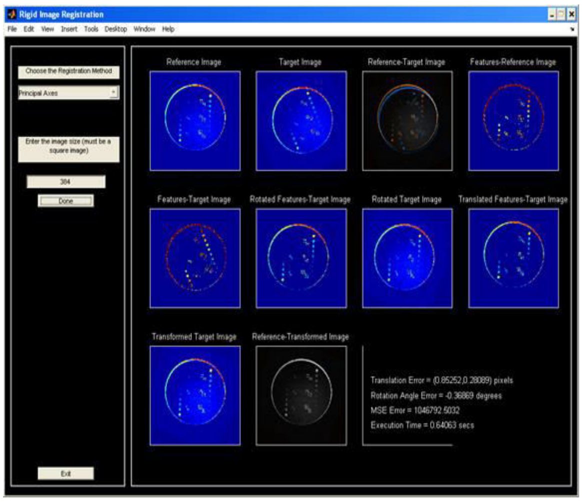
a) Align the two centers of mass of the two images.

b) Construct the translated target image. This is the final transformed aligned target image.

B. Processing

Figure 3 shows the processing window after executing the Principal Axes algorithm. The images displayed are explained as below:

1. The first image shows the reference image.
2. Second image is the target image.
3. Third image shows the initial alignment difference between the two images. The reference image is shown in red and the target image in blue.
4. The fourth image shows the features selected for the reference image for alignment.
5. The fifth image shows the features selected for the target image for alignment.
6. The sixth image is the rotated features of target image for alignment.
7. The seventh image is the rotated target image.
8. The eighth image is translated features of target image for alignment.
9. The ninth window is the transformed or aligned target image.
10. The tenth window shows the difference between the reference image and the aligned target image. If the alignment is correct, the image difference shows up as white and the regions of image that couldn't be aligned show up in the respective red or blue color for misalignment.
11. Finally, the program displays the similarity measure MSE values. Translation error and rotation error measure are also displayed. Algorithm execution speed is shown in seconds.


Rigid Image Registration
File Edit View Insert Tools Desktop Window Help
Choose the Registration Method
Enter the image size (must be a square image)
Done
Exit
Reference Image
Target Image
Reference-Target Image
Features-Reference Image
Features-Target Image
Rotated Features-Target Image
Rotated Target Image
Translated Features-Target Image
Transformed Target Image
Reference-Transformed Image

Principal Axes

384

```
Translation Error = (0.85252,0.28089) pixels
Rotation Angle Error = -0.36869 degrees
MSE Error = 1046792.5032
Execution Time = 0.64063 secs
```

Fig. 3 : Software Processing Window - Principal Axes

C. Experimental Results

Twenty-two of the target images are tested for alignment with the first image in the sequence acquired from the QA phantom, here referred to as the reference image. The Principal Axes algorithm successfully registered seventeen target images with minimal translation and rotation errors.

The algorithm successfully registered the following target files: target, target0, target00, target5, target10, target15, target20, target25, target50, target90, target_5, target_10, target_15, target_20, target_25, target_50, target_90. The algorithm failed to align files target150, target180, and target_180. These images are the ones which are misaligned by very large angles (greater than 90 degrees) compared to reference image. In addition, the algorithm also failed to align the noisy images target10n and target25n. This shows the method's sensitivity to presence of noise in images during registration process.

V. FAST FOURIER TRANSFORM METHOD

A. *Algorithm*

Steps involved for FFT based registration are:

Let I1 and I2 be the reference and target images respectively.

i) Calculate rotation parameter.

   a) Apply FFT function to images I1 and I2, to get F1, F2.

   b) Transform the Cartesian coordinate system points (x, y) into log-polar coordinates (log (p), θ) using formulae

      Log (p) = log (sqrt(x2 + y2)) and θ = atan(y/x).

   c) Find the new intensity values at the corresponding log polar coordinates using bilinear interpolation.

   d) Apply FFT to log-polar images to get Flp1 and Flp2.

   e) Compute the ratio

      R1 = (Flp1 * conj(Flp2)) / (abs(Flp1 * conj(Flp2)))

      where conj is the complex conjugate and abs is the absolute value.

   f) Compute the inverse FFT of R1 as IR1.

   g) Find the location (xo, yo) in log-polar coordinates for maximum value of abs (IR1).

   h) Calculate the rotation angle θ using formulae

      xo = mod (loc, cols); yo = loc / rows;

      Rotation angle θ in radians is the y-displacement in the log-polar coordinate.

ii) Perform rotation transformation on the target image.

   a) Construct the rotation matrix as

      R = [cos(-θ) sin(-θ); -sin(-θ) cos(-θ)];

   b) Multiply the rotation matrix with the (x, y) coordinates of the target image to get the new rotated image I3.

iii) Calculate translation parameters along x-axis and y-axis.

   a) Apply FFT function to image I3, resulting in F3.

   b) Compute the ratio

      R1 = (F1 * conj(F3)) / (abs(F1 * conj(F3))) .

   c) Compute the inverse FFT of R1 as IR1.

   d) Find the location for maximum value of abs(IR1).

   e) Find the translation point (xo, yo) using formulae

      xo = mod (loc, cols); yo = loc / rows;

iv) Perform translation transformation.

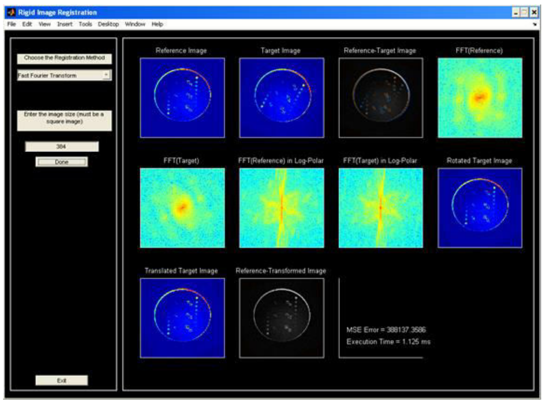
   a) Align image I3 by adding the displacements to the original points in the rotated target image.

   b) Construct the translated target image. This is the final transformed aligned target image.

B. *Processing*

Figure 4 shows the processing window after executing the FFT method.

1. The first subplot shows the reference image.
2. Second subplot is the target image.
3. Third subplot shows the initial alignment difference between the two images. The reference image is shown in red and the target image in blue.
4. The fourth and fifth subplots show the frequency domains of reference and target images respectively.
5. The sixth and seventh subplots show the frequency domains of reference and target images respectively represented in log-polar coordinate system.
6. Eighth image is the rotated version of the target image.
7. Ninth image is the translated target image.
8. Tenth window shows the difference between the reference image and the aligned target image.
9. Finally, the program displays the similarity measure MSE values. Algorithm execution speed is shown in seconds.


Rigid Image Registration
File Edit View Insert Tools Desktop Window Help
Choose the Registration Method
Enter the image size (must be a square image)
Done
Exit
Reference Image
Target Image
Reference-Target Image
FFT(Reference)
FFT(Target)
FFT(Reference) in Log-Polar
FFT(Target) in Log-Polar
Rotated Target Image
Translated Target Image
Reference-Transformed Image

Fast Fourier Transform

384

MSE Error = 388137.3586
Execution Time = 1.125 ms

Fig. 4 : Software Processing Window - FFT

C. *Experimental Results*

Twenty-two of the target images are tested for alignment with the first image in the sequence acquired from the QA phantom, here referred to as the reference image. The Fast Fourier Transform algorithm successfully registered twenty-one target images with minimal translation and rotation errors between the two images.

The algorithm successfully registered the following target files: target, target0, target5, target10, target15, target20, target25, target50, target90, target150, target180, target_5, target_10, target_15, target_20, target_25, target_50, target_90, target_180, target10n, target25n. The algorithm successfully registered target images target10n and target25n in the presence of noise showing its insensitivity to the presence of noise in images during registration process. The algorithm failed to align image target00 which was largely displaced.

VI. SIMILARITY MEASURES

Image similarities are broadly used in medical imaging. An image similarity measure quantifies the degree of similarity between intensity patterns in two images [11] [12]. The choice of an image similarity measure depends on the modality of the images to be registered.

Common examples of image similarity measures include cross-correlation, mutual information, sum of squared intensity differences, sum of absolute differences, mean square error and ratio image uniformity [11].

We have used the error metrics measure Mean Square Error (MSE) for measuring similarity between the reference image and the transformed target image. The MSE is the cumulative squared error between the transformed image and the reference image. The mathematical formula for measuring MSE is

$$\mathrm{MSE} = 1/MN * \sum^{M}{}_{y=1}\sum^{N}{}_{x=1}[I(x, y) - I'(x, y)]^2$$

where I(x, y) is the reference image, I'(x, y) is the transformed image, and M,N are the dimensions of the images. A lower value for MSE means lower similarity error and higher similarity between the two images.

For the Principal Axes method, we have also measured translation and rotation differences between the reference image and the transformed target image to get the translation error and the rotation angle error. This gives us a realistic picture of alignment mismatch extent as can also be validated by looking at the difference between the images. Measuring the translation and rotation errors for confirming the accuracy of alignment by using FFT method is computationally expensive as it requires additional computation of six Fast Fourier Transforms. Hence, MSE is used for measuring the similarity.

The difference between the reference image and the transformed image is shown to the user by subtracting the two images and displaying the difference image in the processing window. A perfect alignment results in difference image shown in gray scale and any misalignments if exist are shown either in red or blue color in the image. This gives us a realistic picture of alignment mismatch extent as can be validated by looking at the difference image.

MSE values showed a value of zero for exact alignment between images for the Principal Axes method. For similar alignments achieved, the MSE values ranged approximately from 1.5e4 to 2.0e6 and for misalignments, the MSE values approximated in the range of 5.9e6 to 6.4e6.

For the FFT method, the MSE values ranged approximately from 1.1e3 to 9.9e5 and for misalignments, the MSE values are shown to be higher in the approximate range of 1.3e6 to 2.5e6. Figure 5 shows the similarity measure distribution for each target image registered using both methods.

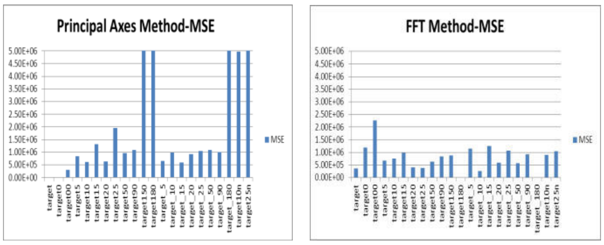


Fig. 5 : Similarity Measure MSE computed.

Translation and rotation errors are calculated for registering images using the Principal Axes method. Figure 6 shows their distribution for each target image. Translation errors measured along x-axis and y-axis are less than 1 pixel for all the successful registrations and rotation angle error is less than 1.5 degrees.

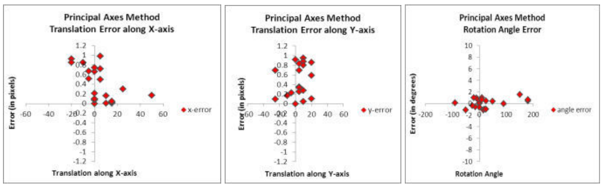


Fig. 6 : Translation and Rotation Errors for Principal Axes method.

VII. CONCLUSIONS AND FUTURE WORK

Based on the Mean Square Error similarity measure analysis, the FFT method performed better in aligning the phantom images with MSE values ten times lesser than those for the Principal Axes method showing more similarity between the reference image and the transformed target image for successful registrations.

Unlike the Principal Axes method, the FFT method successfully registered the noisy images.

Translation and rotation errors calculated for all successful registrations using the Principal Axes method gave minimum errors of up to 1 pixel and 1.5 degrees respectively showing better similarities achieved.

The Principal Axes method successfully registered 17 of the 22 non-aligned test images, the FFT method registered 21 test images whereas using the 'Realign' tool in the existing SPM8 package with default settings showed correct alignments for only 9 images as per our requirement.

If one of the algorithms gives coarse similarity between the images after registration, then the user may realign this output image using the other algorithm to get better alignment results.

Once the phantom images are aligned using our software, the future work in our QA project involves using these aligned images to measure and compute different QA parameters such as Signal to Noise Ratio, Contrast to Noise Ratio, Ghosting Fraction, Resolution, Magnetic Homogeneity and so on. The comparative study of these QA parameters over time will give us an idea about how image quality is getting affected due to functioning of the scanner magnet.

VIII. ACKNOWLEDGMENT

This study is supported by the Graduate Research Thesis Grant awarded to RJ from the College of Graduate Studies, Georgia Southern University.

REFERENCES

[1] Peter J. Kostelec and Senthil Periaswamy, Image Registration for MRI, Modern Signal Processing, MSRI Publications, Volume 46, 2003.

[2] J. Ashburner and K. Friston, Rigid body registration, The Wellcome Dept. of Imaging Neuroscience, 12 Queen Square, London WC1N 3BG, UK.

[3] N. M. Alpert, J. F. Bradshaw, D. Kennedy, and J. A. Correia, The Principal Axes Transformation - A Method for Image Registration, J NucIMed 1990;31:1717-1722.

[4] Robin Kramer, Automatic MRI Image Registration Using Phase Correlation, Department of Computer Science, University of Wisconsin, Madison, United States.

[5] B. Antoine Maintz , Max A. Viergever, An Overview of Medical Image Registration Methods, Symposium of the Belgian Hospital Physicists Association, 1996.

[6] Medha V. Wyawahare, Dr. Pradeep M. Patil, and Hemant K. Abhyankar, Image Registration Techniques: An overview, International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 2, No.3, September 2009.

[7] Hongjie Xie, Nigel Hicks, George R. Keller, Haitao Huang, Vladik Kreinovich, An IDL/ENVI Implementation of the FFT Based Algorithm for Automatic Image Registration, Computers and Geosciences, 2003, Vol. 29, No. 8, pp. 1045-1055.

[8] B. Srinivasa Reddy and B. N. Chatterji, An FFT-Based Technique for Translation, Rotation, and Scale-Invariant Image Registration, Proc. of IEEE Transactions On Image Processing, Vol. 5, No. 8, August 1996.

[9] George Wolberg, Siavash Zokai, Robust Image Registration Using Log-Polar Transform, Proc. of IEEE Intl. Conf. on Image Processing, Sep. 2000.

[10] Cynthia Rodriguez, Understanding FFT- based algorithm to calculate image displacements with IDL programming language, University of Texas at San Antonio, 2007.

[11] A. Ardeshir Goshtasby. 2-D and 3-D Image Registration for Medical, Remote Sensing, and Industrial Applications, Wiley Press, 2005.

[12] http://en.wikipedia.org/wiki/

[13] Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ. Statistical parametric maps in functional imaging: a general linear approach. Human Brain Mapping. 1994; 2(4):189–210.

[14] Abràmoff, M.D., Magalhães, P.J. and Ram, S.J. Image Processing with ImageJ. Biophotonics International, 11(7):36—42, 2004.

[15] The MathWorks, Inc., MA, USA.

[16] http://homepages.inf.ed.ac.uk/rbf/HIPR2/ fourier.htm

◈◈◈