

Natarajan Meghanathan
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

Fourth International Conference on Computer Science, Engineering and
Applications (ICCSEA 2014)
Chennai, India, July 26 ~ 27 - 2014



AIRCC

Volume Editors

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

Dhinaharan Nagamalai,
Wireilla Net Solutions PTY LTD,
Sydney, Australia
E-mail: dhinthia@yahoo.com

ISSN : 2231 - 5403
ISBN : 978-1-921987-08-3
DOI : 10.5121/csit.2014.4708 - 10.5121/csit.2014.4730

This work is subject to copyright. All rights are reserved, whether whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the International Copyright Law and permission for use must always be obtained from Academy & Industry Research Collaboration Center. Violations are liable to prosecution under the International Copyright Law.

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

Fourth International Conference on Computer Science, Engineering & Applications (ICCSEA 2014) was held in Chennai, India, during July 26~27, 2014. Third International Conference on Signal, Image Processing and Pattern Recognition (SPPR 2014), Fifth International Conference on VLSI (VLSI 2014), Sixth International Conference on Wireless, Mobile Network & Applications (WiMoA 2014), Third International Conference on Soft Computing, Artificial Intelligence & Applications (SCAI 2014), Seventh International Conference on Network Security & Applications (CNSA 2014) and Sixth International Conference on Web services & Semantic Technology (WeST 2014) were collocated with the ICCSEA-2014. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The ICCSEA-2014, SPPR-2014, VLSI-2014, WiMoA-2014, SCAI-2014, CNSA-2014, WeST-2014 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, ICCSEA-2014, SPPR-2014, VLSI-2014, WiMoA-2014, SCAI-2014, CNSA-2014, WeST-2014 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the ICCSEA-2014, SPPR-2014, VLSI-2014, WiMoA-2014, SCAI-2014, CNSA-2014, WeST-2014.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Natarajan Meghanathan
Dhinaharan Nagamalai

Organization

Program Committee Members

Aayush C	Rensselaer Polytechnic Institute, USA
Abdurrahman Celebi	Beder University, Albania
Ahmed Elfatetry	Alexandria University, Egypt
Aiden B. Lee	Qualcomm Inc, USA
Akira Otsuki	Tokyo Institute of Technology, Japan
Ali Abid D. Al-Zuky	Mustansiriyah University, Iraq
Ali AL-zuky	Mustansiriyah University, Iraq
Ali Fadhil	Oklahoma State University, USA
Alireza Souri	Islamic Azad university, Iran
Amine Achouri	University of Tunis, Tunisia
Amir Abbas Baradaran	PNU and Azad Universities, Iran
Amir Konigsberg	General Motors, Israel
Aref Tahmasb	Shahid bahonar university, Iran
Arifa Ferdous	Varendra University, Bangladesh
Asmaa Shaker Ashoor	University of Babylon, Iraq
Badji Mokhtar	Annaba University, Algeria
Bankas Edem	University for Development Studies, Ghana
Benmohammed M	University of Constantine, Algeria
Breno C. Costa	Proativa Software Ltd, Brazil
Dac-Nhuong Le	Haiphong University, Vietnam
Debashis De	University of Western, Australia
Dhinaharan Nagamalai	Wireilla Net Solutions, Australia
Dongchen Li	Peking University, China
Ederval Pablo Ferreira da Cruz	Federal Institute of Espirito Santo, Brazil
Ehsan Saradar Torshizi	Urmia university, Iran
El Houssainy Rady	Institute of statistical studies and research, Egypt
El Miloud Ar-Reyouchi	Abdelmalek Essaadi University, Morocco
Eng. Mohamed Abdel Karimx	Pharos University, Egypt
Esmaeel Kheir Khah	Islamic Azad university, Iran
Farhad Soleimanian	Hacettepe University, Turkey
Farshchi S. M. R	Tehran University, Tehran
Fatih Korkmaz	Cankiri Karatekin University, Turkey
Fatma Elghannam	Electronics Research Institute, Egypt
G. Totkov	Medical University - Plovdiv, Bulgaria
Gaurav Ojha	Indian Institute of Information Technology and Management, India
Gullanar M Hadi	Salahaddin University, Kurdistan-Iraq
Guo Yue	Ningbo University of Technology, China
Hacene Belhadeif	University of Constantine 2, Algeria
Hayati Mamur	Cankiri Karatekin University, Turkey
Hossein Jadidoleslami	Mangosuthu University of Technology, Iran
Irving V Paputungan	Universitas Islam Indonesia, Indonesia

Isa Maleki
 Iti Mathur
 Jims Marchang
 Joberto S.B. Martins
 Ka Chan
 Karol Matiasko
 Kenneth Mapoka
 Khaled Merit
 Khanbabaie M
 Khaze S.R
 Kirti Patel
 Laya Pamela
 Mahgoub Hany
 Majlinda Fetaji
 Marjan Mahmoodi
 Masoud Ziabari
 Md. Ibrahim Chowdhury
 Melih Kirlidog
 Michal Wozniak
 Mohamed AlAjmi
 Mohamed Ali Mahjoub
 Mohammad Farhan Khan
 Mohammad Omar Alhawarat
 Mohammed Youssif
 Mohd Shahizan Othman
 Muhammad Imran Khan
 Natarajan Meghanathan
 Nguyen Dinh, Thuc
 Nilanjan Dey
 Niloofar Khanghahi
 Nouredine Bouhmala
 Nur'Aini Abdul Rashid
 Ola A.Younis
 Orhan Dagdeviren
 Oussama Ghorbel
 P. E. S. N. Krishna Prasad,

 Patel Kirti
 Peyman Mohammadi
 Phuc V. Nguyen
 Pr Abdelmonaime LACHKAR
 Radhakrishnan P
 Rafah M. Almuttairi
 Ram Gopal
 Ranjan Kumar
 Rao
 Remus Brad
 Riemann, Ute

Islamic Azad University, Iran
 Banasthali University, India
 University of Plymouth, United Kingdom
 Salvador University, Argentina
 La Trobe University, Australia
 Zilinska univerzita v Ziline, Slovakia
 Botswana College of Agriculture, Botswana
 University of Mascara, Alegria
 Islamic Azad University, Iran
 Islamic Azad University, Iran
 Prairie View A & M University, USA
 Islamic Azad University, Iran
 Menoufia University, Egypt
 South East European University, Macedonia
 Islamic Azad University, Iran
 Mehr Aeen University, Iran
 City University, Bangladesh
 Marmara University, Turkey
 Wroclaw University of Technology, Poland
 King Saud University, Saudi Arabia
 National Engineering School of Sousse, Tunisia
 University of Kent, United Kingdom
 Salman Bin Abdulaziz University, Saudi Arabia
 Hewlett-Packard, Egypt
 Universiti Teknologi Malaysia, Malaysia
 University of Toulouse, France
 Jackson State University, USA
 HCMC-University of Science, Vietnam
 JIS College of Engineering, India
 Islamic Azad University, Iran
 Vestfold and Buskerud College, Norway
 Universiti Sains Malaysia, Pulau Pinang
 Philadelphia university, Jordan
 Ege university, Turkey
 Sfax University, Tunisia
 Prasad V. Potluri Siddhartha Institute of
 Technology, India
 Prairie View A & M University, USA
 Islamic Azad University, Iran
 L'institut Polytechnique Saint-Louis, France
 SMBA University Fez, Morocco
 King Khalid University, Saudi Arabia
 University of Babylon, Iraq
 Nokia Solutions and Networks, USA
 Cambridge Institute of Technology, India
 Hewlett-Packard, USA
 Lucian Blaga University of Sibiu, Romania
 SAP AG, Germany

Romildo Martins da Silva Bezerra
S.R. KHAZE
Saad M. Darwish
Saeid Ghazi
Sassi Abdessamed
Sayyed majid mazinani
Seifedine Kadry
Sergio Takeo Kofuji
Seyyed Mohammadreza Farshchi
Seyyed Reza Khaze
Shilpi Bose
Siamak
Sutanu
T.C.Manjunath
Tinatin Mshvidobadze
Vuda Sreenivasarao
Wahiba Ben Abdessalem
Xiao G
Yassine Maleh
Yusmadi Jusoh

Federal Institute of Bahia (IFBA), Brazil
Islamic Azad University, Iran
Alexandria University, Egypt
Islamic Azad University, Iran
Biskra university, Algeria
ImamReza International University, Iran
American University of the Middle East, Kuwait
University of Sao Paulo, Brazil
Tehran University, Tehran
Islamic Azad University, Iran
Netaji Subhash Engineering College, India
Blekinge Institute of Technology, Sweden
University of Mosul, Iraq
Visvesvaraya Technological University, India
Gori University, Georgia
Bahir Dar University, Ethiopia
High Institute of Management of Tunis, Tunisia
Southwest University, China
Hassan 1st university, Morocco
Universiti Putra Malaysia, Malaysia

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)



Networks & Communications Community (NCC)



Digital Signal & Image Processing Community (DSIPC)



Organized By



Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

Fourth International Conference on Computer Science, Engineering and Applications (ICCSEA 2014)

Performance of Private Cache Replacement Policies for Multicore Processors.....	01 - 07
--	----------------

Matthew Lentz and Manoj Franklin

Dynamic Data Management Among Multiple Databases for Optimization of Parallel Computations in Heterogeneous HPC Systems.....	09 - 19
---	----------------

Pawel Rosciszewski

The Effect of Parallel Corpus Quality vs Size in English-Toturkish SMT.....	21 - 30
--	----------------

Eray Yildiz, Ahmed Cüneyd Tantug and Banu Diri

Vector ANT Colony Optimization and Travelling Salesman Problem.....	31 - 39
--	----------------

Chiranjib Patra and Pratyush

Detection and Tracking of Multiple Objects in Cluttered Backgrounds with Occlusion Handling.....	41 - 48
---	----------------

Sukanyathara J and Alphonsa Kuriakose

Heat Stress Risk Prediction by Using Bayesian Net Model with Sensor Network.....	49 - 57
---	----------------

Kanchan M.Taiwade and Prakash S. Mohod

Third International Conference on Signal, Image Processing and Pattern Recognition (SPPR 2014)

Orthogonal Discrete Frequency Coding Space Time Waveform for MIMO Radar Detection in Compound Gaussian Clutter.....	59 - 66
--	----------------

B. Roja Reddy and M. Uttarakumari

Estimation of Recursive Order Number of Photocopied Document Based on Probability Distributions.....	67 - 74
---	----------------

Suman V Patgar, Rani K and Vasudev T

Cost-Effective Stereo Vision System for Mobile Robot Navigation and 3D Map Reconstruction.....	75 - 86
---	----------------

Arjun B Krishnan and Jayaram Kollipara

**Performance Comparison of Online Handwriting Recognition System
for Assamese Language Based on HMM and SVM Modelling.....** 87 - 95
Deepjoy Das, Rituparna Devi, SRM Prasanna, Subhankar Ghosh and Krishna Naik

Fifth International Conference on VLSI (VLSI 2014)

Graphite : A Graph Search Framework..... 97 - 110
Sushanta Pradhan

Low Power VLSI Compressors for Biomedical Applications..... 111 - 120
Thottempudi Pardhu, S.Manusha and K.Sirisha

**Performance Comparison of 4T, 3T and 3T1D DRAM Cell Design on
32 NM Technology.....** 121 - 133
Prateek Asthana and Sangeeta Mangesh

**A Highly Adaptive Operational Amplifier with Recycling Folded Cascode
Topology.....** 135 - 145
Saumya Vij, Anu Gupta and Alok Mittal

Sixth International Conference on Wireless, Mobile Network & Applications (WiMoA 2014)

**Transferring of Information in Wireless Adhoc Sensor Network Using
Shortest Path Algorithm.....** 147 - 155
N. Pushpalatha and B.Anuradha

Duty Cycled Multi Channel MAC for Wireless Sensor Networks..... 157 - 174
M. Ramakrishnan

**An Efficient and More Secure ID Based Mutual Authentication Scheme
Based on ECC for Mobile Devices.....** 175- 186
Shubhangi N. Burde and Hemlata Dakhore

Third International Conference on Soft Computing, Artificial Intelligence & Applications (SCAI 2014)

Human Gait Analysis and Recognition Using Support Vector Machines..... 187 - 195
Deepjoy Das and Sarat Saharia

A Mind Map Query in Information Retrieval : The 'User Query Idea' Concept and Preliminary Results.....	197 - 213
<i>Rihab Ayed, Farah Harrathi, M. Mohsen Gammoudi and Mahran Farhat</i>	

Seventh International Conference on Network Security & Applications (CNSA 2014)

Real-Time Detection of Phishing Tweets.....	215 - 227
<i>Nilesh Sharma, Nishant Sharma, Vishakha Tiwari, Shweta Chahar and Smriti Maheshwari</i>	

New Functions for Secrecy on Real Protocols.....	229 - 250
<i>Jaouhar Fattahi and Mohamed Mejri and Hanane Houmani</i>	

Sixth International Conference on Web services & Semantic Technology (WeST 2014)

Web Service Composition Based on Popularity.....	251 - 263
<i>Selwa Elfirdoussi, Zahi Jarir and Mohamed QUAFAFOU</i>	

A Formalized Model for Semantic Web Service Selection Based on QoS Parameters.....	265 - 283
<i>Divya Sachan, Saurabh Kumar Dixit and Sandeep Kumar</i>	

PERFORMANCE OF PRIVATE CACHE REPLACEMENT POLICIES FOR MULTICORE PROCESSORS

Matthew Lentz and Manoj Franklin

Department of Electrical and Computer Engineering,
University of Maryland, College Park, USA

mlentz743@gmail.com

manoj@eng.umd.edu

ABSTRACT

Multicore processors have become ubiquitous, both in general-purpose and special-purpose applications. With the number of transistors in a chip continuing to increase, the number of cores in a processor is also expected to increase. Cache replacement policy is an important design parameter of a cache hierarchy. As most of the processor designs have become multicore, there is a need to study cache replacement policies for multi-core systems. Previous studies have focused on the shared levels of the multicore cache hierarchy. In this study, we focus on the top level of the hierarchy, which bears the brunt of the memory requests emanating from each processor core. We measure the miss rates of various cache replacement policies, as the number of cores is steadily increased from 1 to 16. The study was done by modifying the publicly available SESC simulator, which models in detail a multicore processor with a multi-level cache hierarchy. Our experimental results show that for the private L1 caches, the LRU (Least Recently Used) replacement policy outperforms all of the other replacement policies. This is in contrast to what was observed in previous studies for the shared L2 cache. The results presented in this paper are useful for hardware designers to optimize their cache designs or the program codes.

KEYWORDS

Multicore, Cache memory, Replacement policies, & Performance evaluation

1. INTRODUCTION

All major high-performance microprocessor vendors are currently selling multicore chips. Future generations of these processors will undoubtedly have an increasing number of cores. Multicore architectures exploit the inherent parallelism present in programs, which is the primary means of increasing processor performance, besides decreasing the clock period and the memory latency. This high performance is based on the assumption that the memory system does not significantly stall the processor cores.

Most modern multi-core processors incorporate multiple levels of the cache hierarchy. Typically, the top level of the hierarchy consists of *private* L1 caches, as the top level needs to support the high bandwidth requirements of all the cores [1][2]. Moreover, the working sets of the threads executing in multiple cores will not cause interference in each other's L1 cache. Also, small private caches allow fast access. The subsequent levels of the cache hierarchy — L2 and L3 — are

generally shared between multiple cores. Because the memory requests arriving at the L2 cache have already been filtered by the private L1 caches, the shared L2 cache can handle the bandwidth requirements. Figure 1 illustrates such a cache hierarchy for a 2-core processor.

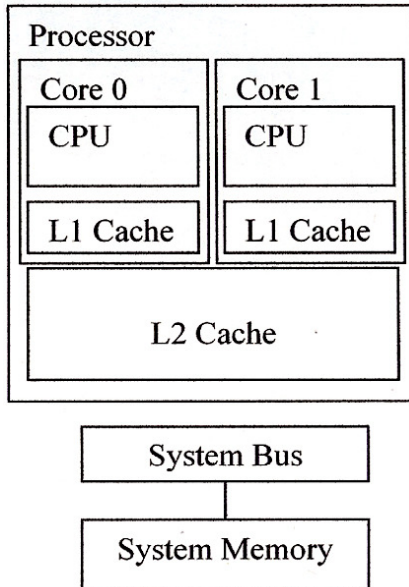


Figure 1. A 2-core processor with private L1 caches and a shared L2 cache

This paper studies the suitability of various cache replacement policies for the private L1 caches, as future processors increase the number of cores in each processor. Given the drastic change in processor hardware design, it is important to ensure that cache replacement policies are chosen such that they scale well with an ever-increasing core count. With increasing numbers of cores, caches will end up with more and more cache blocks that are replicated among the various cores as a direct result of shared variables in a multi-threaded program. The change in how the caches are used may have an impact on which cache replacement policy performs the best with increasing processor core counts.

In a set-associative cache, the cache replacement policy determines which block (in a set) to replace when there is a miss. Some of the widely used replacement policies are LRU (Least Recently Used), FIFO (First-In First-Out), and RAND (Random). Among these, LRU generally has the lowest miss rates for single-core systems, as it most closely correlates with the concept of temporal locality. However, it is also the one that has the most hardware complexity; the cache needs to maintain LRU information for each set and update this information during every memory access.

1.1. Prior Work

Cache replacement policies have been explored extensively for single-core systems over the last three decades. However, in today's multicore environment, there is a need to re-evaluate the policies, especially as the number of cores increases. Several multi-core cache studies have been done recently [3, 4, 5, 6, 7]. However, all of these studies were directed at the shared L2 cache. Because L1 cache hits do not propagate to the L2 cache, the L2 cache observes only a filtered memory access stream. The behavior of the L2 cache is therefore different from that of the private L1 caches. Therefore, there is a need to investigate cache replacement policies for the private L1 caches when running multiple threads at the same time. Although several single-core studies have

been done for the L1 cache in the past, those results may not be directly applicable to a multi-core environment where the per-core cache is smaller than that provided in a typical single-core processor. Moreover, the behavior of a private cache can be affected by how the shared cache levels below it behave. This paper studies the performance of various cache replacement policies for the L1 private caches, as the number of processor cores is varied from 1 to 16.

The rest of this paper is organized as follows. Section 2 describes the experimental framework and methodology we used for this study. Section 3 presents the experimental results and analysis. The paper concludes in Section 4 with a summary of the findings.

2. EVALUATION METHODOLOGY

We use detailed execution-based simulation to evaluate the performance of different cache replacement policies. Execution-based simulation captures subtle effects not possible with trace-based simulation. For example, different cache replacement policies might cause the execution to go along paths that are different from that followed in the trace. For evaluation, we use the publicly available SESC (SuperESCalar) simulator [8][9], a cycle-accurate simulator that models in detail a multicore processor with a multi-level cache hierarchy. SESC models different processor architectures, such as single processors, chip multi-processors, and processors-in-memory. It models a full out-of-order pipeline with branch prediction, caches, buses, and every other component of a modern processor necessary for accurate simulation.

2.1. Cache Replacement Policies

The SESC simulator incorporates the LRU and RANDOM replacement policies for the L1 caches. We modified the simulator to include additional cache policies that might prove to be more scalable when it comes to multi-threaded programs running on many cores. The implementations were tested with both single threaded and multi-threaded programs to ensure their validity. Specifically, we included the following three replacement policies:

- a. Most Recently Used (MRU)
- b. First In-First Out (FIFO)
- c. First In-First Out with 2nd Chance (FIFO2)

LRU and FIFO are both common cache replacement policy schemes, and as such they were included in our study. MRU was included because it works well for programs that access a given amount of memory in the same way multiple times (such as looping through an array more than once), as it saves older entries that will then provide cache hits upon the following times through. FIFO2 was included because it does provide a higher percentage of cache hits than FIFO, due to the fact that it allows old entries a chance to stay in the queue through the checking of a reference bit that is reset at every replacement. However, this gain naturally comes with a higher implementation cost [2].

2.2. Cache Configuration

The cache configuration simulated includes private L1 instruction and data caches, and a shared L2 cache. Each of the L1 caches was 32kB (for both the instruction and the data cache) with an associativity of 4, while the L2 cache was 512kB with an associativity of 8. Both the L1 and the L2 caches use a 32 byte cache block size.

2.3. Benchmarks Simulated

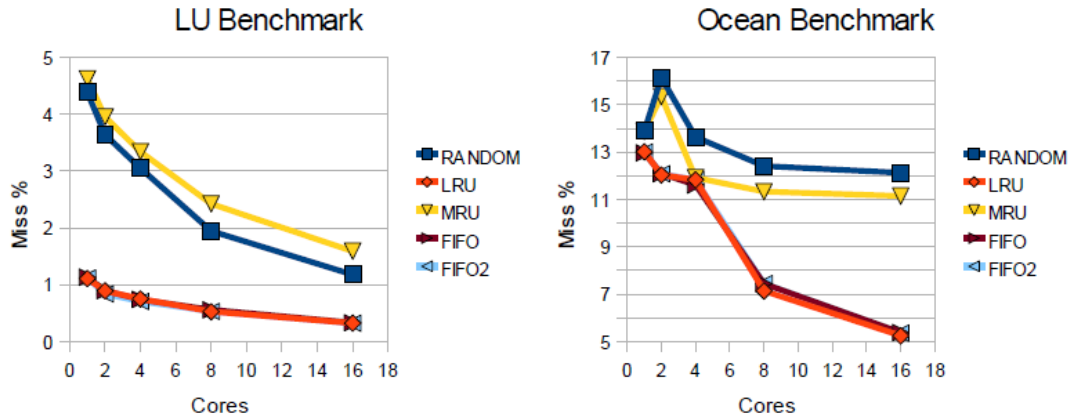
In order to analyze which cache policy performs the best as the number of cores in the processor increases, it is important to have a wide variety of benchmarks that test a number of different cache-usage patterns. The benchmarks must also be multi-threaded and scalable up to 16 threads in order to test a number of processor configurations and ensure that all available core resources are being utilized. The benchmarks that we used consist of some of the SPLASH2 benchmarks published by Stanford: FFT, LU, Radix, Ocean, and Raytrace [10]. These benchmarks are selected based on their relevance to multicore processors and cache memories. We used pre-compiled SPLASH2 benchmarks from a link on the mailing list archive [11]. Each benchmark is run by itself on the multi-core system. Core 0 runs the initialization code of the program, and then spawns the parallel threads, which run on all the cores (including core 0).

3. EXPERIMENTAL RESULTS AND ANALYSIS

For analyzing the data that was gathered from the simulations, we focus primarily on the miss ratios for the private L1 Data caches. The L1 Instruction cache miss percentages were extremely similar between the various benchmarks, regardless of the number of cores for the simulation. The differences were on the order of around just 0.01%. Given the large number of reads and writes to the L1 Instruction caches, the inclusion of the L1 Instruction cache hit/miss data would suppress differences that could be seen in the L1 Data cache hit/miss data.

3.1. Results

Figure 2 presents the miss ratios we obtained for the private L1 data caches. Data is presented as a separate graph for each benchmark. For each graph, the X axis depicts the number of cores and the Y-axis depicts the miss ratio as a percentage. Each graph shows 5 plots, corresponding to the 5 cache replacement policies investigated. Each data point represents a miss ratio calculated as a weighted average across each of the cores present, where the weight for each core is determined by the number of instructions that are executed on the core compared with the total number of instructions executed. This weight really only differs for the spawning processor, due to the extra instructions that it needs to execute in order to spawn the remaining 1-15 threads.



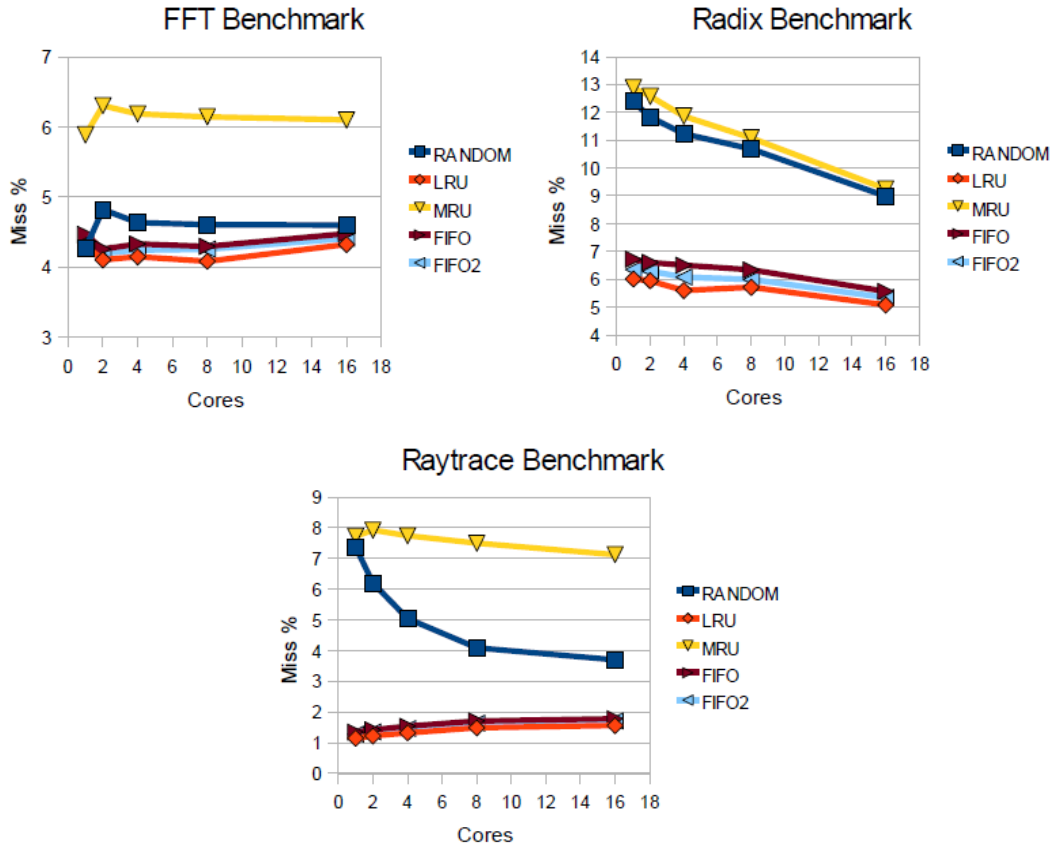


Figure 2. Miss ratios obtained for the private L1 data caches

3.2. Analysis

Overall, the results are very consistent across the benchmarks, with LRU showing the best results in terms of performance scalability, as the number of cores in the processor is increased. This was the expected outcome, given the strong performance that LRU delivers on single-core systems, due to being a close approximation for the OPT (optimum) cache replacement policy. Future work involves considering modifications to the basic LRU scheme to see if additional performance gains are possible in multicore environments.

One trend to note is how closely FIFO and FIFO2 follow LRU in terms of the miss percent as the number of cores on the processor increases. This is due to how closely both FIFO and FIFO2 can approximate an LRU implementation – especially FIFO2 due to the ability to push older, yet still actively referenced, cache lines to the back of the replacement queue: allowing it to perform replacements similar to that of an LRU implementation.

Another trend to note is the relatively poor performance of MRU across all configurations and benchmarks, with the exception of the Ocean benchmark running on a single core, where it came surprisingly close to LRU, FIFO, and FIFO2. Looking at other miss ratios for that benchmark, it can be seen that LRU, FIFO, and FIFO2 end up trending in a vastly different direction than MRU as the number of cores increases. The constant private L1 cache sizes across all cores, regardless of how many cores there are in the processor, is most likely the cause of MRU not scaling very well with increasing core counts for the Ocean benchmark. The single processor data workload

makes an MRU implementation work pretty well in comparison with the other replacement policies, but the lower per-processor data workload with increasing core counts brings about the large difference in performance.

The performance of RANDOM is generally bad, except for FFT in which case its miss ratios are comparable to that of FIFO. In general, the RANDOM policy is not a good choice for the top level of the cache hierarchy in a multicore environment.

4. CONCLUSIONS

In this paper, we explored a wide range of cache replacement policies for the top level of the memory hierarchy in a multi-core system. We varied the number of cores from 1 to 16, in order to study the scalability of each replacement policy. The analysis of the experimental results points clearly to LRU being the most scalable cache replacement policy for private L1 caches, as it consistently performed the best over all the benchmark simulations and for all core counts. This result jibes well with the strong performance that LRU had exhibited in previous-era single-core systems. FIFO and FIFO2 follow the same general trend of LRU in terms of scalability across all of the benchmarks, trailing it by a relatively small amount at any given data point. MRU did not prove to be a very useful cache replacement policy, as it only tends to work well for certain data request patterns, such as looping over the same array multiple times. The RANDOM policy, while sometimes popular in systems in which the overhead of any cache replacement calculation is impossible to deal with, also did not perform that well. Clearly, for the multicore configurations simulated and the benchmarks considered, LRU is the most scalable and FIFO/FIFO2 follow very closely with its performance trend.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone! An important bug fix for the SESC simulator came from Alireza Haghdoost, a graduate student in Computer Engineering at Sharif University of Technology, who provided a fix that solved a segmentation fault that occurred when running SESC for more than two processor cores.

REFERENCES

- [1] Rajeev Balasubramonian, Norman Jouppi, and Naveen Muralimanohar, (2011), “Multi-Core Cache Hierarchies”, Synthesis Lectures in Computer Architecture, Morgan & Claypool Publishers.
- [2] John Hennessy and David Patterson (2007) *Computer Architecture A Quantitative Approach*, 4th ed. Morgan Kaufmann.
- [3] T.S.B. Sudarshan, Rahil Abbas Mir, and S. Vijayalakshmi (2004) “Highly Efficient Implementations for High Associativity Cache Memory”, *Proc. 12th IEEE International Conf. on Advanced Computing and Communications*, Vol. 10, No. 5, pp87-95.
- [4] Mohamed Zahran (2007) “Cache Replacement Policy Revisited,” *Proc. Annual Workshop on Duplicating, Deconstructing, and Debunking (WDDD)* held in conjunction with the International Symposium on Computer Architecture (ISCA).
- [5] Rahul V. Garde, Samantika Subramaniam, and Gabriel H. Loh (2008) “Deconstructing the Inefficacy of Cache Replacement Policies”, *Proc. Annual Workshop on Duplicating, Deconstructing, and Debunking (WDDD)* held in conjunction with the International Symposium on Computer Architecture (ISCA).
- [6] Tripti S. Warriar, B. Anupama, and Madhu Mutyam (2013) “An Application-Aware Cache Replacement Policy for Last-Level Caches”, *Architecture of Computing Systems (ARCS) Lecture Notes in Computer Science* Volume 7767, pp. 207-219.

- [7] S. Muthukumar and P. K. Jawahar (2014) “Sharing and Hit based Prioritizing Replacement Algorithm for Multi-Threaded Applications”, *International Journal of Computer Applications*, Vol. 90, No. 9.
- [8] Jose Renau, SESC. <http://sesc.sourceforge.net>
- [9] Pablo Ortego & Paul Sack (2004) “SESC: SuperEScalar Simulator”.
- [10] Christian Bienia, Sanjeev Kumar, and Kai Li, (2008), “Parsec vs SPLASH-2: A Quantitative Comparison of two multithreaded Benchmark Suites on Chip Multiprocessors”, IEEE International Symposium on Workload Characterization, pp. 47-56.
- [11] "Stanford Parallel Applications for Shared Memory." 07 Sept 2001. Web. 20 Oct 2009. <<http://www-flash.stanford.edu/apps/SPLASH/>>.

AUTHORS

Matthew Lentz is a graduate student in the Ph.D. program in the ECE Department at the University of Maryland at College Park, where he also obtained his B.S. in Computer Engineering. He is broadly interested in the areas of networking and systems research.



Manoj Franklin is a faculty member of the Electrical and Computer Engineering Department at the University of Maryland at College Park. His research interests are in the broad areas of computer architecture and systems.



INTENTIONAL BLANK

DYNAMIC DATA MANAGEMENT AMONG MULTIPLE DATABASES FOR OPTIMIZATION OF PARALLEL COMPUTATIONS IN HETEROGENEOUS HPC SYSTEMS

Paweł Rościszewski

Department of Computer Architecture,
Faculty of Electronics, Telecommunications and Informatics,
Gdańsk University of Technology, Gdańsk, Poland
`pawel.roszczewski@pg.gda.pl`

ABSTRACT

Rapid development of diverse computer architectures and hardware accelerators caused that designing parallel systems faces new problems resulting from their heterogeneity. Our implementation of a parallel system called KernelHive allows to efficiently run applications in a heterogeneous environment consisting of multiple collections of nodes with different types of computing devices. The execution engine of the system is open for optimizer implementations, focusing on various criteria. In this paper, we propose a new optimizer for KernelHive, that utilizes distributed databases and performs data prefetching to optimize the execution time of applications, which process large input data. Employing a versatile data management scheme, which allows combining various distributed data providers, we propose using NoSQL databases for our purposes. We support our solution with results of experiments with our OpenCL implementation of a regular expression matching application.

KEYWORDS

Parallel Computing, High Performance Computing, Heterogeneous Environments, NoSQL, OpenCL

1. INTRODUCTION

The market of electronic hardware is developing in extreme pace, making sophisticated computing devices accessible to households. Research and development departments of hardware manufacturing companies compete in designing new architectures and accelerators. HPC (High Performance Computing) systems no longer can be considered as sets of very expensive devices forming a cluster, physically installed in one room. The HPC field has to deal with increasing heterogeneity of the systems and it should be taken into account that the parallelization is performed on many levels. We should be able to combine concepts as Grid Computing [1], GPGPU [2] and Volunteer Computing [3] into one multi-level parallel design.

Our parallel processing framework, KernelHive [13] is able to perform parallel computations on a set of distributed clusters containing nodes with different types of computing devices. We

presented the KernelHive system and its performance capabilities in [4] and proposed an execution optimizer focusing on energy efficiency in [5]. Within this article, we add data intensity capabilities to the KernelHive system. For this purpose we propose MongoDB [6] database as a backend. For our experiments, we use our solution to the regular expression matching problem [7].

2. PROBLEM FORMULATION

From the parallelization point of view, the spectrum of computational problems in general can be structured as shown in Figure 1. The parallelization process requires dividing the problem into subproblems, solving them independently by parallel processes and finally merging the results. Certain problems require only partitioning the input data into chunks, which are processed independently. Problems of this type are called embarrassingly parallel and on Figure 1 are located in the compute intensive corner. Until this work, the KernelHive system was dealing only with this type of problems (e.g. breaking MD5 hashes).

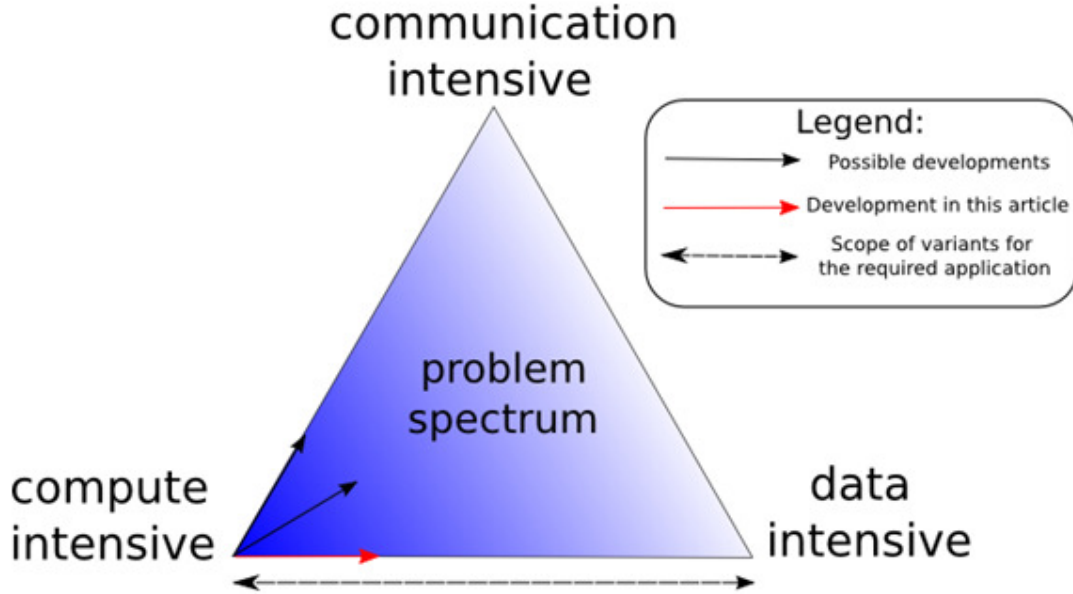


Figure 1. Spectrum of Computational Problems from the Parallelization Viewpoint

The black arrows on Figure 1 show possible directions of development of the KernelHive system. Moving in the direction towards communication intensive vertex of the problem spectrum, we would be dealing with applications that require communication between the processes (for example for frequent updating intermediate results). This direction of development should be addressed in the future.

In this paper we follow the red arrow and add data intensity flavor to the KernelHive system. For this purpose, we propose an architecture of distributed databases and a versatile protocol, that allows using various database systems. In order to show the possibilities of this architecture, using the exchangeable optimizer API in KernelHive and MongoDB for storage, we implement a data prefetching optimizer.

3. PROPOSED SOLUTION

3.1. Overview of the Existing Architecture

The system architecture so far is shown on Figure 2. Using the graphical interface, the user defines an application in a form of a directed acyclic graph. Graph nodes correspond to computational tasks and are selected from a repository of predefined node types (e.g. processor, partitioner, merger). Each node is provided with a number of computational codes corresponding to its role. The edges of the graph denote the direction of data flow between the tasks.

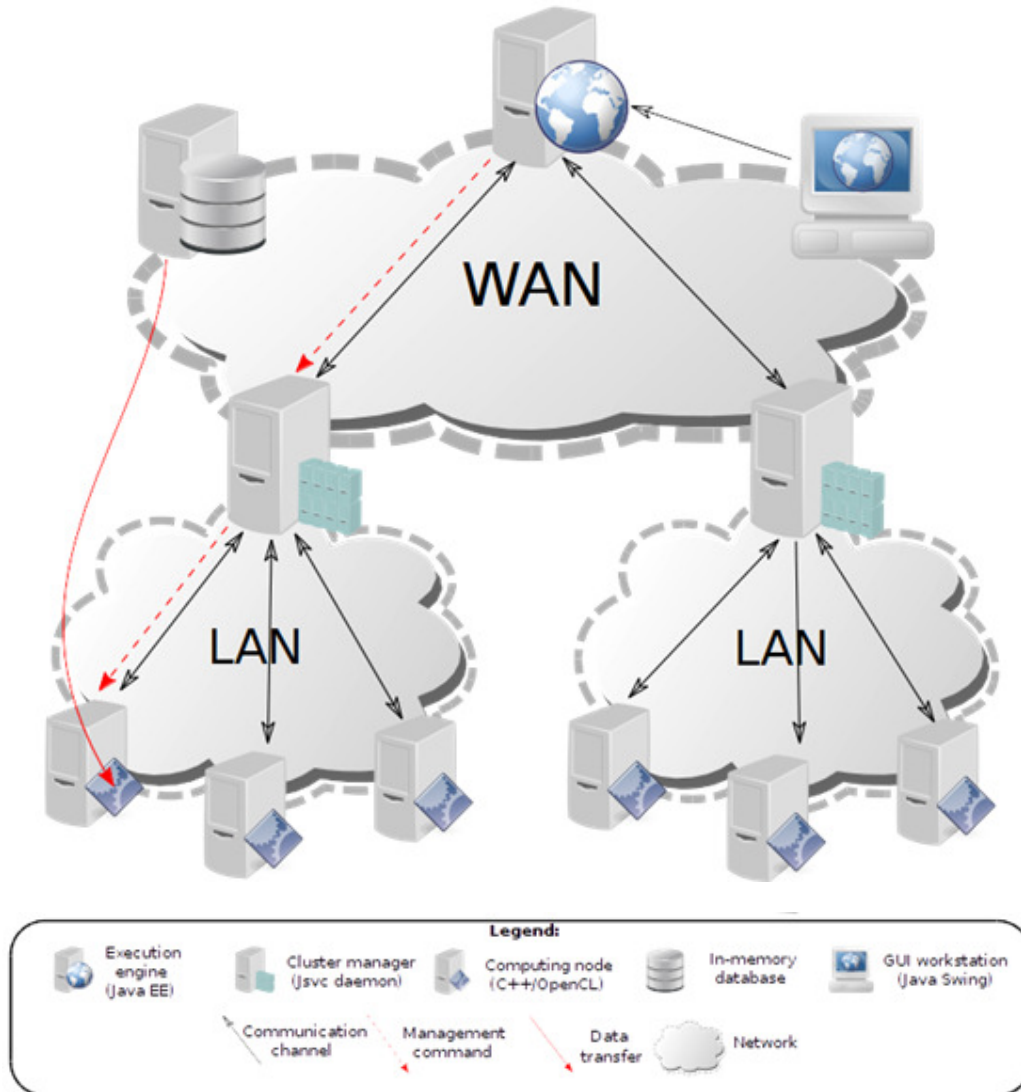


Figure 2. Basic Architecture of the Parallel System

Applications for the system can be defined using our graphical tool called *hive-gui* (Java Swing application), however they are represented in a XML format, allowing other front-ends to use the parallel system. A tested example of such front-end is the Galaxy Simulator [9] which was

extended by a plugin for KernelHive executions. The application XML, along with data addresses are dispatched to execution by a SOAP web service.

Analysis and deployment of the applications is performed by the *Engine*, which is a high-level Java application. All of the subject modules report their state to the engine, keeping a live representation of the whole system in the engine. Thus we can define rules of scheduling the tasks that have a rich view of the available infrastructure and its state.

One of the distinct features of KernelHive is that it is designed to combine multiple distributed clusters. The only requirement towards a cluster is that there should be one machine playing a role of an entry point to the cluster, which has to be visible (in terms of network) by the *Engine*. To address this requirement, the system utilizes the *Cluster* subsystem, working as a Java system daemon.

The cluster manager is a middleman between the engine and computing devices, which are managed by C++ daemons, capable of dynamic compiling and running OpenCL [10] computational code.

3.2. Novelty

In the embarrassingly parallel applications considered so far, the time of sending the input data to the computational node was negligible: the data could be considered as a part of the management command and stored in memory. However, in case of larger data a method of storing data on hard drive has to be employed, should it be a database system or file system. What is more, it should be noted that the bandwidth between the cluster manager and engine and, more importantly, data server, is significantly lower than in the local network between cluster manager and computational nodes. In case of larger input data, we propose an approach, where management commands contain only addresses of data packages. The addresses are defined in a versatile way and consist of:

- hostname – the TCP hostname of the data server
- port – the TCP port of the data server
- ID – identifier of the data package unique within the data server scope

This approach has two main advantages:

- tolerates different technologies for the data servers, which allows to adjust the data server to the characteristics of given application and deployment
- grants the possibility to move the data between the servers during the application runtime and changing the addresses in management commands at low cost

In this paper we show examples of exploring both these advantages. For the first one, we propose using MongoDB key-value store with the GridFS [11] drivers as the technology for data servers. The power of the second advantage is exposed on the example of communication and computation overlapping by prefetching input data to local servers. The modified architecture of the system is depicted by Figure 3.

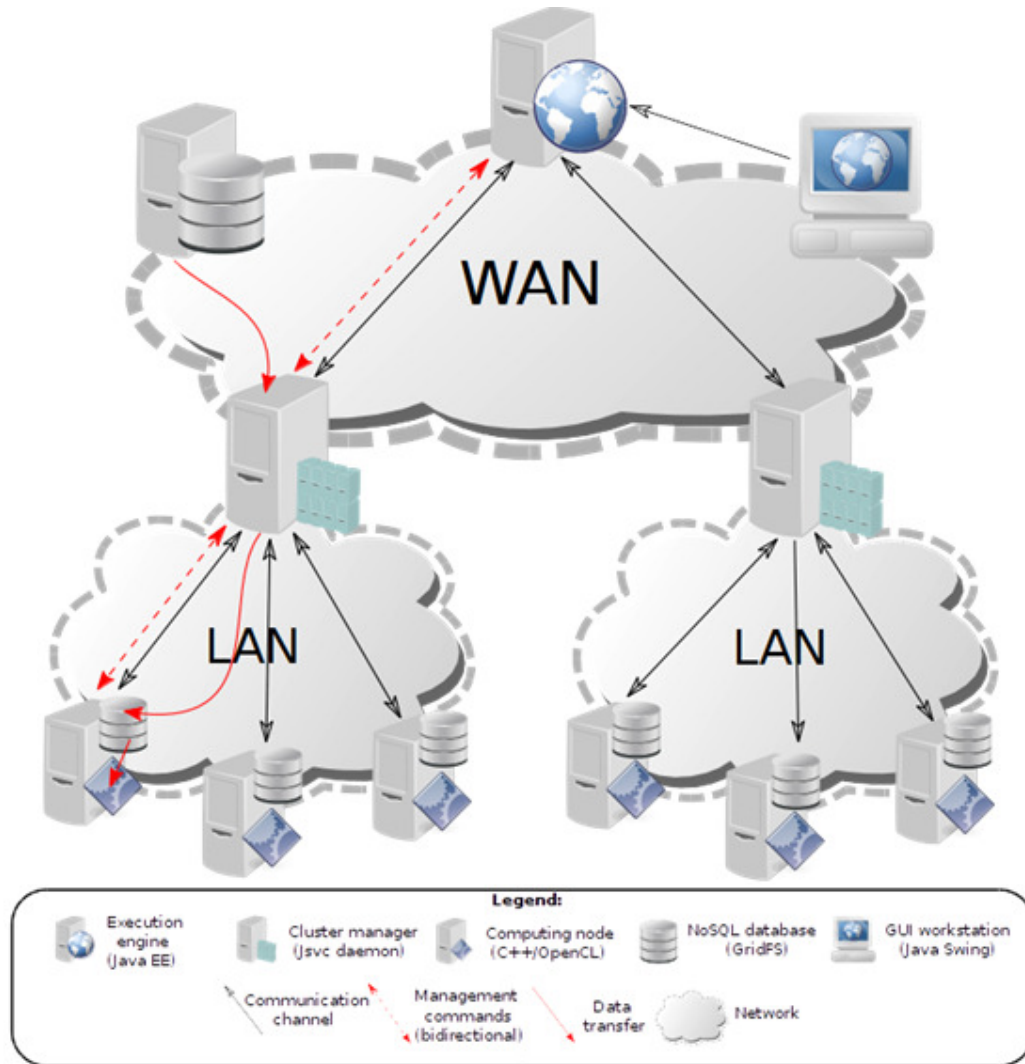


Figure 3. Modified Architecture of the Parallel System

3.3. NoSQL Data Servers

There are numerous technologies designed for storing and accessing big data. The concept of file systems evolved from basic structures for storing data on local hard drives to sophisticated distributed file systems. These solutions are closely connected with the operating system issues, like access control and hierarchical organisation of data. Because of this, they often introduce some constraints on file names, limit number of files in a directory etc. However, file systems are widely used as the backend for HPC systems, which have to be aware of the characteristics of used file systems.

Another important approach towards storing data is relational databases. Database management systems (DBMS) deal with the low level details of storage and hide them from the user. They provide wide functionality of storing, retrieving, filtering data, often with regard to transactions and cascading of operations. The data is modeled in a rigid form of relational tables with columns corresponding to certain object attributes and rows representing some objects.

In case of HPC systems, we rarely require the database to understand the model of our data. Often, we just need to store a big file and keep an address to refer to it later. However, we would like to benefit from the low-level internal transparency offered by the database systems. For this reason, we propose to use a NoSQL database for our reasons.

The NoSQL [12] concept is close to the relational databases, however abandons the rigid representation of data. For our experiments, we chose the most popular NoSQL database at the time, MongoDB. This database system comes with an extension called GridFS. The extension is actually a functionality of the MongoDB drivers that allows automatic dividing the data to chunks, storing them separately, but keeping information about the whole files in metadata.

Another reason for using MongoDB in our heterogeneous system is that it offers mature driver implementations for different programming platforms. We benefited from the implementations in:

- python – for the input data package generator
- C++ - for the program on the computing devices to download the input database
- Java – for the cluster manager to perform the data prefetching

3.4. Data Prefetching

The KernelHive *Engine* defines a *IOptimizer* programming interface, listed in Figure 4.

```
public interface IOptimizer {
    /**
     * Given a submitted workflow and available infrastructure,
     * return a list of scheduled jobs with assigned devices to
     * be deployed by the Engine.
     */
    public List<Job> processWorkflow(Workflow workflow, Collection<
        Cluster> infrastructure);
}
```

Figure 4. *IOptimizer* interface

The input of each Optimizer implementation consists of:

- *Workflow* – class representing the whole application workflow, including individual jobs and relations between them. The optimizer has access to the state of each job (e.g. pending, ready, prefetching, prefetching finished, finished)
- collection of *Clusters* – a set of instances of *Cluster* class, each representing a collection of computational nodes. The optimizer has access to the full infrastructure model, including the computing devices, their characteristics and current state.

The value returned by the optimizer is a list of jobs, that were scheduled for execution. Additionally, the optimizer should change the states of affected jobs and devices.

The interface is general enough to allow its implementations to focus on different criteria and perform diverse tasks. It is also possible to combine several optimizers to achieve a complex goal. We have already implemented scheduling optimizers aimed for dynamic assignment of jobs that

became ready for execution to available devices according to certain criteria (e.g. performance, energy efficiency).

The new *PrefetchingOptimizer* implementation requires an internal optimizer for scheduling. This way, choosing the hardware for computations can be done by an exchangeable component. Such base is extended by a data prefetching mechanism, listed in Figure 5. The optimizer implementation keeps the information about currently performed prefetchings in a map. Each prefetching process is represented by a key-value pair of jobs:

- key – a job that is being processed (data has already been downloaded by the worker)
- value – next job assigned to the same computational device as the key job, however not yet scheduled for execution – only for data downloading

Using a data structure defined this way, the tasks of the optimizer are as follows:

- if a key job has ended and the prefetching for the corresponding next job is over, mark the latter as scheduled for execution
- let the internal optimizer perform the scheduling of jobs that became ready for execution (due to the workflow dependencies) using hardware that became available (because it finished its computations or has been just connected to the system)
- ensure, that for each currently processed job, there is a corresponding job, for which the input data is being prefetched (provided there are some jobs ready for execution)

The optimizers *processWorkflow* method is called by the *Engine* upon every event that changes the aforementioned states of jobs and hardware, including finishing a job, finishing a prefetching, submitting new workflow or connecting new hardware.

After each call of this method, the list of scheduled jobs returned by the optimizer is sent by the *Engine* to appropriate *Cluster* subsystem instances. Then, the jobs are forwarded to the assigned machines, where the *Unit* subsystem listens for jobs to run. Finally, the adequate *Worker* binary is executed. It downloads the necessary input data, application code, builds it and runs the computations.

When the computations are finished, the output data is saved in a previously configured database. A management command is send back through the *Cluster* to the *Engine*, containing the resulting data package ID. In case of final results, the ID is used to download them upon users request. In case of intermediate data, the ID is used by following jobs in the workflow.

```

public class PrefetchingOptimizer implements IOptimizer {

    private Map<EngineJob, EngineJob> prefetchingMap = new HashMap<EngineJob, EngineJob>();
    private IOptimizer baseOptimizer;

    public PrefetchingOptimizer(IOptimizer baseOptimizer) {
        this.baseOptimizer = baseOptimizer;
    }

    @Override
    public List<Job> processWorkflow(Workflow workflow,
        Collection<Cluster> infrastructure) {
        // First schedule jobs that were prefetched
        List<Job> scheduledJobs = schedulePrefetchedJobs();

        // Then schedule jobs to free resources
        scheduledJobs.addAll(baseOptimizer.processWorkflow(workflow, infrastructure));

        // Then start prefetchings
        List<EngineJob> processingJobs = workflow.getJobsByState(Job.JobState.PROCESSING);
        processingJobs.removeAll(prefetchingMap.keySet());

        if(processingJobs.size() > 0) {
            List<EngineJob> readyJobs = workflow.getJobsByState(Job.JobState.READY);
            Iterator<EngineJob> readyIterator = readyJobs.iterator();
            for(EngineJob pj : processingJobs) {
                if(!readyIterator.hasNext())
                    break;

                EngineJob prefetchingJob = readyIterator.next();

                prefetchingJob.device = pj.device;
                prefetchingJob.runPrefetching();

                prefetchingMap.put(pj, prefetchingJob);
            }
        }

        return scheduledJobs;
    }

    private List<Job> schedulePrefetchedJobs() {
        List<Job> scheduledJobs = new ArrayList<Job>();
        List<EngineJob> toRemove = new ArrayList<EngineJob>();
        for(EngineJob processingJob : prefetchingMap.keySet()) {
            if(processingJob.state == JobState.FINISHED) {
                EngineJob prefetchingJob = prefetchingMap.get(processingJob);
                if(prefetchingJob.state == JobState.PREFETCHING_FINISHED) {
                    prefetchingJob.schedule(prefetchingJob.device);
                    scheduledJobs.add(prefetchingJob);
                    toRemove.add(processingJob);
                }
                // Do not schedule new job if we are waiting for prefetching
                else prefetchingJob.device.busy = true;
            }
        }

        for(EngineJob tr : toRemove)
            prefetchingMap.remove(tr);

        return scheduledJobs;
    }
}

```

Figure 5. The new *PrefetchingOptimizer*

4. EXPERIMENTS

The proposed solution was tested in one series of experiments. We measured the execution times of a regular expression matching application with different numbers of input data packages. The data packages are 20MB files of random characters, generated and stored in MongoDB by our generator script. Additionally, each package is prefixed with a header containing the needle and haystack sizes, and the needle itself. In the experiments we searched for the occurrences of the pattern "a*b*c*d".

The prefetching algorithm should enhance the systems performance provided the WAN network shown on Figure 3 brings significant delays and bandwidth limits. To reflect this situation during the experiments, the source database was hosted on a server in France, while the computations took place in our department lab in Poland.

4.1. Experiments on a Single Device

We started with testing the solution on a basic setup with one machine equipped with one Intel Core i5 processor. The execution times in seconds were gathered in Table 1. As it turns out, the results in case of a single device are as expected: for one data package, the difference between execution time with and without prefetching is negligible. The scenario of execution is the same in both cases. The more data packages, the higher the speedup of the prefetching version, reaching 30% in case of 4 input packages. The difference is significant and increasing, because in the prefetching scenarios, data transmission and computations are overlapping.

Table 1. Results of the Single Device Experiment

NPackages	No prefetching	Prefetching
1	28s	29s
2	60s	54s
3	91s	63s
4	123s	86s

4.2. Experiments on a Heterogeneous Infrastructure

After testing the proposed design in action and proving the usefulness of the prefetching optimizer, we tested the same application on a cluster of nodes equipped with different types of devices. The infrastructure for this extended tests is shown on Figure 6, which is actually a screenshot from the hive-gui application, that enables generating the infrastructure charts based on the data from the *Engine*.

In order to compare the results in the new testbed configuration to the previous ones, we had to run the application with package numbers N times higher, where N is the number of computing devices.

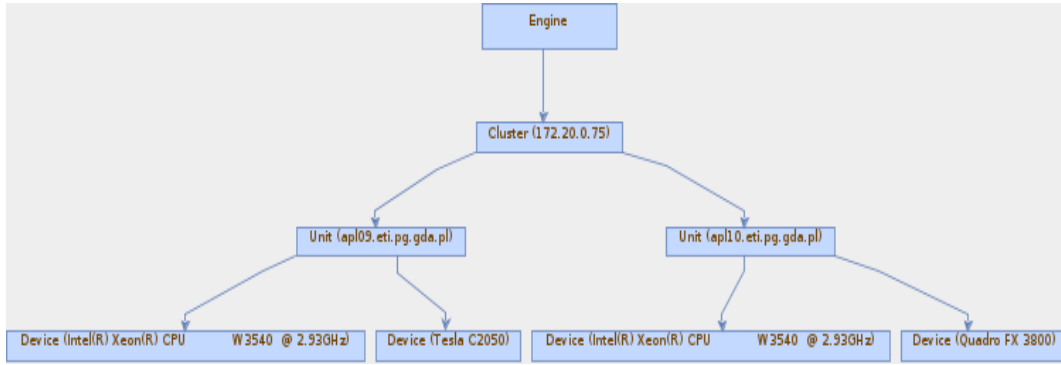


Figure 6. The Heterogeneous Testbed Configuration

The results of the experiment (Table 2) show, that the benefit from prefetching, though significant and increasing, is lower than for one device and in case of 40 packages reaches 11%. Such results could be an effect of sharing the network between multiple prefetching tasks. Still, the optimizer shows promising results in a heterogeneous environment.

Table 2. Results of the Heterogeneous Environment Experiment

NPackages	No prefetching	Prefetching
4	31s	32s
8	77s	66s
12	87s	73s
16	103s	91s
20	122s	104s

5. SUMMARY AND FUTURE WORK

Focusing on the aspect of data management in parallel computing systems brings up a number of issues, especially in case of heterogeneous multi-level systems. In this paper we addressed a subset of those issues by extending our parallel framework KernelHive.

We proposed an architecture with multiple distributed data servers and a versatile data addressing scheme that enables using various data storage technologies and high-level optimizations. On this basis we used GridFS as a data storage engine and presented the implementation of a new optimizer for KernelHive, that enables prefetching data to the computing devices, causing the overlapping of computations and communication.

Our experiments, based on a regular expression matching application showed that the proposed solution is a base for new data management schemes. In the future we could extend this solution by mechanisms of dynamic transferring of intermediate results between the parallel processes with regard to their distribution, possibly among distant clusters.

ACKNOWLEDGEMENTS

The work was performed partially within grant “Modeling efficiency, reliability and power consumption of multilevel parallel HPC systems using CPUs and GPUs” sponsored by and covered by funds from the National Science Center in Poland based on decision no DEC-2012/07/B/ST6/01516.

Special thanks to Tomasz Boiński (<http://tboinski.eti.pg.gda.pl>) for his invaluable support.

REFERENCES

- [1] Grandinetti, L., ed.: Grid computing. Elsevier, Amsterdam [u.a.] (2005)
- [2] Thompson, C.J., Hahn, S., Oskin, M.: Using modern graphics architectures for general-purpose computing: a framework and analysis. In: MICRO, ACM/IEEE (2002) 306–317
- [3] Anderson, D.P.: Volunteer computing: the ultimate cloud. ACM Crossroads 16(3) (2010) 7–10
- [4] Czarnul, P., Lewandowski, R., Rościszewski, P., Schally-Kacprzak, M.: Multi-level parallelization of computations using clusters with gpus (2013) Poster presented at GPU Technology Conference, San Jose, USA.
- [5] Czarnul, P., Rościszewski, P.: Optimization of execution time under power consumption constraints in a heterogeneous parallel system with gpus and cpus. In Chatterjee, M., Cao, J.N., Kothapalli, K., Rajsbaum, S., eds.: ICDCN. Volume 8314 of Lecture Notes in Computer Science., Springer (2014) 66–80
- [6] Chodorow, K., Dirolf, M.: MongoDB - The Definitive Guide: Powerful and Scalable Data Storage. O'Reilly (2010)
- [7] Thompson, K.: Regular expression search algorithm. Communications of the ACM 11(6) (June 1968) 419–422
- [8] Pao, D., Or, N.L., Cheung, R.C.C.: A memory-based nfa regular expression match engine for signature-based intrusion detection. Computer Communications 36(10-11) (2013) 1255–1267
- [9] Kacala, B., Sagadyn, G., Sidorczak, P., Czarnul, P.: Parallel simulation of the galaxy with dark matter using gpus and cpus (2013) Poster presented at GPU Technology Conference, San Jose, USA.
- [10] Khronos OpenCL Working Group: The OpenCL Specification, version 1.0.29. (8 December 2008)
- [11] Bhardwaj, D., Sinha, M.: Gridfs: highly scalable i/o solution for clusters and computational grids. IJCSE 2(5/6) (2006) 287–291
- [12] Edlich, S.: NOSQL Databases (January 2011)
- [13] KernelHive website, viewed on July 2014, <http://kask.eti.pg.gda.pl/en/projekty/kernelhive/>.

INTENTIONAL BLANK

THE EFFECT OF PARALLEL CORPUS QUALITY VS SIZE IN ENGLISH-TO- TURKISH SMT

Eray Yıldız¹Ahmed Cüneyd Tantug²and Banu Diri³

¹Department of Computer Engineering,
Yildiz Technical University, Istanbul, Turkey
yildizeray@hotmail.com.tr

²Computer and Informatics Faculty,
Istanbul Technical University, Istanbul, Turkey
tantug@itu.edu.tr

³Department of Computer Engineering,
Yildiz Technical University, Istanbul, Turkey
banu@ce.yildiz.edu.tr

ABSTRACT

A parallel corpus plays an important role in statistical machine translation (SMT) systems. In this study, our aim is to figure out the effects of parallel corpus size and quality in the SMT. We develop a machine learning based classifier to classify parallel sentence pairs as high-quality or poor-quality. We applied this classifier to a parallel corpus containing 1 million parallel English-Turkish sentence pairs and obtained 600K high-quality parallel sentence pairs. We train multiple SMT systems with various sizes of entire raw parallel corpus and filtered high-quality corpus and evaluate their performance. As expected, our experiments show that the size of parallel corpus is a major factor in translation performance. However, instead of extending corpus with all available “so-called” parallel data, a better translation performance and reduced time-complexity can be achieved with a smaller high-quality corpus using a quality filter.

KEYWORDS

Machine Translation, Machine Learning, Natural Language Processing, Parallel Corpus, Data Selection

1. INTRODUCTION

A corpus which is comprised of aligned sentences that are translations of each other, namely a parallel corpus is an essential training data for statistical machine translation (SMT) [3]. Also a parallel corpus is useful for other natural language processing applications such as cross-language information retrieval, word disambiguation and annotation projection. Building a corpus that includes vast amount of parallel sentences is one of the most time-consuming and important works for a high-performance SMT system [10]. Training the translation model component in SMT requires large parallel corpora for the parameters to be estimated [24]. Therefore, higher translation accuracy can be achieved when machine translation systems are trained on increasing amounts of training data [12]. The quality of an SMT system output extremely depends on the

quality of sentence pairs. A small portion of the available parallel text collections are generated naturally within multi-language organizations and governments (like UN, European Parliament, Canada) for a limited set of languages. The manual compilation of a parallel corpus is too expensive, so most of available parallel corpora are generated automatically. Automatic methods for compiling parallel sentence pairs are imprecise and using such a low-quality training corpus that has many non-parallel sentence pairs would cause low quality translations [24]. The noise in an automatically generated corpus might be due to any difference between the contents of source and target documents, non-literal translations or sentence alignment mistakes. It is infeasible to manually eliminate alignment errors in a large parallel corpus; therefore, an automatic evaluation method is desirable to determine if a parallel sentence pair is accurate or not [14]. Such an automatic evaluation method can depend on multiple factors such as overlapping words, sentence lengths, grammar correctness, fluentness, and word usage correctness.

This study focuses on the effect of using a large parallel text versus using a filtered high-quality corpus. The experiments are carried out for an English-to-Turkish SMT task.

Section 2 gives brief information about related previous studies while section 3 is devoted to the details for our training data. Section 4 introduces our method for filtering parallel sentence pairs and experimental results. The final section includes conclusions and future work.

2. RELATED WORK

The process of filtering out non-parallel sentence pairs is considered as a post-processing step of bilingual data mining process. Gale and Church [7] measure the rate of lengths between two bilingual sentences. The length based approaches work remarkably well on language pairs with high correlation in sentence lengths, such as French and English. On the other hand, the performance of a length based sentence aligner significantly decreases for the language pairs with low correlation on length, such as Chinese and English [28]. Chen and Nie [4] build a system in order to search the web for gathering English-Chinese parallel data and clean their gathered data using sentence length features and language detecting methods. Resnik and Smith [21] also build a system for mining parallel text on the web. This system introduces a translation similarity score for data cleaning. The similarity score is symmetric word-to-word model, which controls whether the translation equivalents of the word in the source sentence occur in the target sentence.

Khadivi and Ney [29] develop a rule based algorithm to filter a parallel corpus by using length constraints and translation likelihood measure for the given sentence pairs. In this work, the sentence pairs in a noisy corpus generated from European Union Web Site are reordered so that the noisy sentence pairs are moved to end of the corpus. They report 47.2 BLEU score [19] when the SMT system is trained with the clean corpus (top 97.5% of corpus), whereas the translation score is 46.8 when the system is trained with the entire corpus.

Yasuda et al. [27] develop a training set selection method for an English-Chinese SMT system using a perplexity score. The perplexity score calculated for each sentence by using language model probabilities and the word counts of the sentences. The perplexity of a parallel sentence pair is considered as the geometric means of the perplexities of source and target sentences. They train a SMT system with an initial in-domain corpus and the selected translation pairs whose perplexities are smaller than a threshold. Their method improves BLEU score by 1.76% in comparison with an initial in-domain corpus usage.

Liu and Zhou [14] propose a machine learning based method for evaluating the alignment quality. Linguistic features and constructs extracted from the sentences are unified together into a feature vector to characterize different dimensions of the alignment quality. They use support vector

machines (SVM) to discriminate high-quality and low-quality parallel sentence pairs. The features in this study are the number of misspelled words, language model score for each English sentence, the number of unlinked words provided by a link grammar parser and the translation equivalence measure. The word alignments are obtained by using a comprehensive bilingual dictionary and the word alignment counts are normalized by the sentence lengths to get equivalence measurement. They train their SVM classifier on a 40K training set that is checked by bilingual annotators. Their classifier is reported to have 0.88 precision and recall rates. Taghipour et al. [24] also proposed a classification method for cleaning parallel data. Many features have been tested and used to build the models such as translation probabilities based on IBM translations models [3], the number of the null aligned words, length based features and features based on language model. They chose maximum entropy model for the classification process and they achieved 98.3% accuracy on a 48K Farsi-English parallel corpus expanded with artificial noise.

In the study of Cui et al. [5], an unsupervised method - namely random walk algorithm - is conducted to iteratively compute the importance score of each sentence pair that indicates its quality. Their method utilizes the phrase translation probabilities based on Maximum Likelihood Estimation. They tested their method on various Chinese-English parallel corpus and eventually, their method improves system performance about 0.5~1.0 BLEU.

Munteanu and Marcu [17] train a maximum entropy classifier in order to extract parallel data from large Chinese, Arabic and English non-parallel newspaper corpora. They start from a parallel corpus of 5K sentence pairs to compute word alignments and extract feature values. Their classifier uses the following features: lengths of sentences as well as the length difference and length ratio, percentage of words on each side that have a translation on the other side and the alignment features obtained from word alignment models. They use dictionaries learned from initial parallel corpus whose sizes are 100K, 1M, 10M, 50M and 95M tokens. The precision and recall values of their classifier are 0.97 and 0.45 respectively with the data from Arabic-English in-domain corpus whereas they have 0.94 precision and 0.67 recall with the Arabic-English out-of-domain corpus.

Hoang et al. [10] present a model for extracting parallel sentence pairs from non-parallel corpora resources in different domains. They combine length-based filtering, cognate condition and content similarity measurement. Their content based similarity is the similarity between the target sentence and the translation of the source sentence obtained by a SMT system. Initially they train their SMT system with an initial out-of-domain corpus (50K) and use this system for content based similarity. Afterwards, they filter an in domain comparable corpora (958K). If the candidate sentence pairs pass the condition of similarity measurement, they expand their train data with these new parallel sentence pairs. They retrain the SMT system and repeat the process of expanding parallel corpus as well as improving the SMT system. They expand 50K initial parallel corpus to 95K after 5 iterations and achieve to increase translation BLEU scores from 8.92 to 24.07.

A somehow related approach is Schwenk's [22] lightly supervised training method which is used to translate large amounts of monolingual texts, to filter them and add them to the translation model training data. Another approach that assigns weights to each sentence in the training bitext is proposed in [16]. Foster et al. [8] describe a new approach to SMT adaptation that weights out-of-domain phrase pairs according to their relevance to the target domain. Axelrod et al. [1] propose a method for extracting sentences from a large general-domain parallel corpus by a domain-relevant data selection method.

3. TRAINING DATA

We utilized an English-Turkish parallel corpus of one million sentences compiled from various sources with varying quality levels. This corpus contains news text [26], literature text [23], subtitles text [25] and web crawled text [28]. Although the sentences in these sources are supposed to be parallel, a quick inspection of these data sets exposes the existence of serious sentence alignment issues. Parallel sentences pairs from each source are pre-processed in order to overcome the alignment problems and to maintain a basic level of alignment quality. A lexicon based sentence aligner; Champollion [15] is used for a crude elimination of the sentence pairs that seem to be misaligned. Table 1 shows the number of sentences in each corpus after this filtering.

Table 1 Training dataset after basic filtering

Dataset	Sentences
News	200K
Subtitles	1,200K
Web	150K
Literature	680K
Total	2,230K

4. EXPERIMENTS

Having parallel data described in the previous section, our first investigation is about the quality of each source. In order to alleviate the impact of the dataset size on the translation quality, a fair comparison setup is ensured by generating equal sized versions of the datasets. These clipped versions of datasets include only top 150K parallel sentences from each dataset. Likewise, 1000 parallel sentences are randomly drawn from the remaining parts of these datasets to generate a balanced test set of 4K sentences.

Each 150K bilingual corpus is used to train a separate SMT system using MOSES toolkit [30]. In Table 2, we present the translation performance of each system.

Table 2 Evaluation of SMT systems trained with 150K datasets for each corpus

Training Data	Training Data Size	BLEU Score
News _{150K}	150K	16.59
Subtitles _{150K}	150K	4.72
Web _{150K}	150K	19.56
Literature _{150K}	150K	9.49

As seen in the table above, there are significant discrepancies in the translation quality. One of the reasons for this result is the varying alignment quality among the datasets. A detailed observation inside the datasets reveals the fact that, even after the first coarse filtering, these “so-called” parallel datasets still contain misaligned pairs or pairs suffering from poor translation quality. Though the performance of a SMT system can be improved by incorporating larger monolingual data for the language model (LM) component, it can be also a good idea to filter out low quality sentence pairs to improve the translation model (TM). Moreover, it is obvious that an effective filtering substantially reduces the time-consuming training phase of SMT even if it does not contribute in translation accuracy. So, a classifier that can discriminate the high quality sentence pairs from low quality ones in the parallel corpus is considered as a beneficial pre-processing step.

4.1 Building The Quality Classifier

For each candidate sentence pair in the corpus, we need a reliable way of deciding whether the two sentences in the pair are the proper translations of each other. We extract some features that are considered as quality indicators from the sentence pairs so that an automatic classifier can be trained. A pair quality depends on not only the correctness of alignment, but also the grammar correctness, fluency and word usage correctness [14].

An ungrammatical sentence may cause degradations in the translation model, so spelling attributes of a sentence play an important role for evaluating a sentence’s quality. Therefore, we opt for using a spell checker [11] to calculate the number of misspellings in English side only, as a representation feature.

Another feature which represents the grammatical correctness is based on a language model. The probability is calculated from the language model for each English sentence and it is used as a fluency indicator feature. Since the probability of a sentence decreases as its length increases, we introduce the sentence length as a feature also. The BerkeleyLM toolkit [20] is used for constructing an English language model created from the Web 1T Corpus and assigning probabilities to sentences.

The relation between the lengths of the sentences in the pair is the most common feature for many parallel text applications such as sentence alignment. Hence, we selected the lengths of the sentences, as well as the length differences and length ratio as features.

The last feature is the percentage of words on each side that have a valid translation on the other side according to a dictionary. An electronic English-Turkish bilingual dictionary which contains Turkish equivalents of 88,824 English words is used in conjunction with a naïve stemmer for Turkish. This stemmer assumes the first 5 letters of a word form as the lemma, which is linguistically incorrect but sufficiently effective for lookup purposes [28].

In order to train and evaluate a classifier, a train and a test set is required respectively. The train set is generated by sentence pairs randomly selected from the whole corpora and then manually labelled as high-quality or poor-quality. 983 instances are manually labelled as high-quality while 160 instances are manually labelled as poor-quality. These 160 instances labelled as poor-quality are not completely useless instances; their low quality might be due to non-literal translations, grammatical or spelling errors. Also, additional 674 misaligned sentence pairs are generated by randomly matching non-parallel sentences from both sides. These artificially generated noise instances are auto-labelled as poor-quality and joined with the other instances in the train set. Table 3 depicts the detailed information about our training data.

We have employed several experiments with a number of different machine learning algorithms to build our classifier. The WEKA tool [9] is used to train Radial Basis Function (RBF) Network, Random Forest (RF), Multilayer Perceptron and Support Vector Machine (SVM) based classifiers. Table 4 shows the classification results in terms of micro averaged precision, recall and F1 values. All tests are run in a 10-fold cross validation evaluating process.

Table 3: Training Data for Classifiers

Sentence Pair Quality	Count
High-Quality	983
Poor-Quality	
Auto-Labeled Artificial Noise	674
Manually Labeled From Corpus	160
Total	1817

Table 4: Cross Validation Results of Classification Algorithms

Classifier	Precision	Recall	F1
RBF Network	0.903	0.964	0.933
Random Forest	0.969	0.960	0.965
Multilayer Perceptron	0.938	0.953	0.946
SVM	0.932	0.963	0.947

For our filtering purposes, the classification algorithm that yields the best precision score sounds reasonable because our aim is to guarantee that the filtered corpus only contains high-quality pairs as much as possible, with the cost of leaving out some good pairs. The results show that RF algorithm is the most efficient model for our task. RF is a tree-based ensemble classifier that consists of a number of decision tree classifiers on various sub-samples of the dataset. [2]. Consequently, we have preferred RF as the classification method to produce a higher quality corpus.

It could be said that the classification of the poor quality instances that are manually labelled is harder than the artificial poor quality instances. In order to see the effectiveness of the RF classifier on these 160 instances, 80 manually labelled instances are excluded from training data and used as test instances. The RF classifier labelled 76 instances correctly which means 0.95 accuracy.

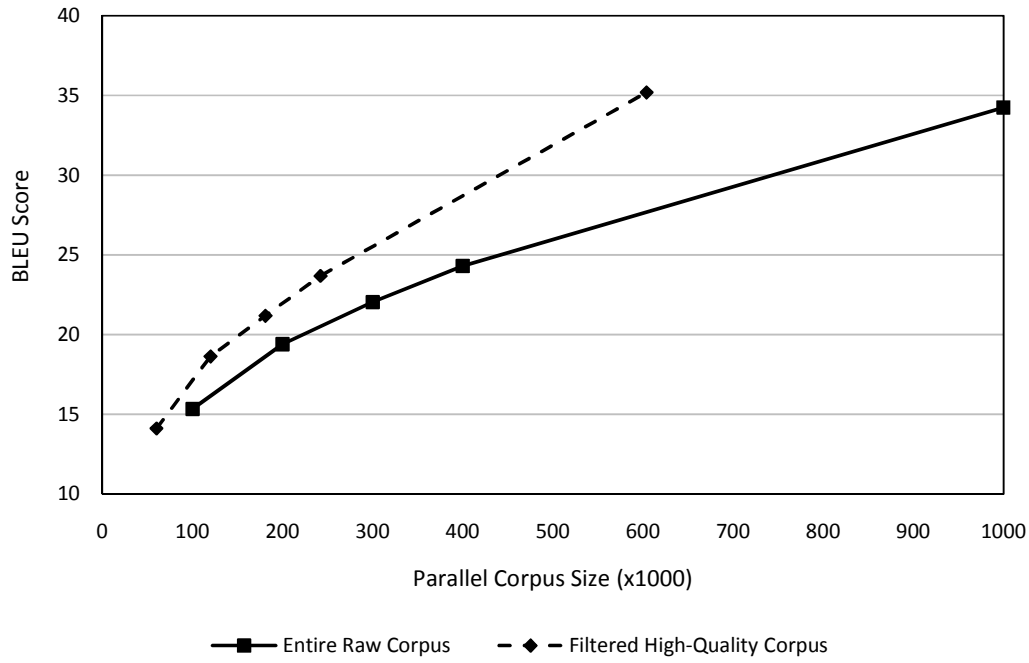


Figure 1 Performance of the SMT system with respect to parallel corpus size and parallel corpus quality

4.2 Experimental Results

Several experiments are conducted with our entire raw parallel corpora and filtered high-quality parallel corpus. To see the impact of the parallel corpus size, we run tests with different corpus sizes up to 1M sentences. The text in the Turkish side of corpus is used for language model training and 10% of each corpus is used as test set. The translation BLEU scores of the SMT outputs are plotted in Figure 1.

As expected, the argument that states the more data, the higher SMT accuracy is proven to be valid. When the number of parallel sentences is increased from 100K to 1000K, the BLEU score increases from 15.33 to 34.24 for the entire raw corpus. The tendency in the plot shows that higher BLEU scores can be achieved with the introduction of more parallel data than 1000K.

On the other hand, Figure 1 reveals another important fact that better translation performance can be acquired by less but high-quality training data. By using only 60% of the raw training data, a BLEU score of 35.19 (which means 2.77% relative improvement) can be achieved with a quality classifier based filtering. The reduction in the training data size also leads to reduction of time complexities of training phase of an SMT system which necessitates a complex and time-consuming process of days or weeks long.

5. CONCLUSIONS

This paper discusses the issue of the quality of the bilingual corpus in SMT. We propose the use of a Random Forest classifier to classify sentence pairs as ‘high-quality’ and ‘poor-quality’ with proposed features based on translation equivalence and grammatical correctness. The experiments show that our filtering method can be used for extracting unsuitable pairs from a noisy parallel

corpus and the remaining pairs can still be effective in achieving results as high as the results of the entire raw corpus. Our results also indicate that there is still need for more parallel data.

The availability of parallel resources in English-Turkish pair is much more limited than the availability of the language pairs that have reliable and big resources such as EuroParl Corpus [13]. Therefore, adding more parallel data is desired, however, the unreliable resources such as Wikipedia and movie subtitles should not be added directly as the training data for a SMT system.

Our filtering method is useful in effective incorporation of these resources in SMT process. Although we presented our results on English-to-Turkish SMT task specifically, the notions in this study can be easily extended for any language pair.

We plan to use this filtering method in our efforts to build up a large-coverage and high-quality English-Turkish parallel corpus.

ACKNOWLEDGEMENTS

We would like to thank to all members of Istanbul Technical University, Natural Language Processing Research Group and our all labmates especially Ezgi, İsmail, Sami. We are also grateful to Natural Language Processing Workgroup in Yıldız Technical University.

REFERENCES

- [1] Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. (2011) "Domain adaptation via pseudo in-domain data selection." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- [2] Breiman, Leo. (2001) "Random forests." *Machine learning* 45.1 : 5-32.
- [3] Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer and Paul S. Roossin. (1990) "A statistical approach to machine translation", *Computational Linguistics*, 16 (2):79-85.
- [4] Chen, Jiang and Jian-Yun Nie. (2000) "Parallel Web text mining for cross-language information retrieval", In *Recherche d'Informations Assistée par Ordinateur (RIA/O)*, pages 62-77, Paris.
- [5] Cui, Lei, Dongdong Zhang, Shujie Liu, Mu Li and Ming Zhou. (2013) Bilingual data cleaning for smt using graph-based random walk. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 340-345)*.
- [6] Esplá-Gomis, Miquel. (2009) "Bitextor, a free/open-source software to harvest translation memories from multilingual websites." *Proceedings of MT Summit XII*, Ottawa, Canada. Association for Machine Translation in the Americas.
- [7] Gale, William A. and Kenneth W. Church. (1993) "A program for aligning sentences in bilingual corpora", *Comput. Linguist.*, 19:75-102
- [8] Foster, George, Cyril Goutte, and Roland Kuhn. (2010) "Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation", In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Cambridge, US-MA, pp 451-459
- [9] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten. (2009) "The weka data mining software: An update", *ACM SIGKDD explorations newsletter* 11.1 : 10-18.
- [10] Hoang, Cuong, Nguyen Phuong Thai, and Ho Tu Bao. (2012) "Exploiting non-parallel corpora for statistical machine translation", In *Proceedings of The 9th IEEE-RIVF International Conference on Computing and Communication Technologies*, pages 97 - 102. IEEE Computer Society.
- [11] Idzelis Mindaugas. (2005) "Jazzy: The java open source spell checker", <http://jazzy.sourceforge.net/>
- [12] Koehn, Philipp. (2002) "Europarl: A multilingual corpus for evaluation of machine translation", Information Sciences Institute, University of Southern California.
- [13] Koehn, Philipp. (2005) "EuroParl: A Parallel Corpus for Statistical Machine Translation", *Machine Translation Summit 2005*. Phuket, Thailand.

- [14] Liu, Xiaohua, and Ming Zhou. (2010) "Evaluating the quality of web-mined bilingual sentences using multiple linguistic features", Asian Language Processing (IALP), 2010 International Conference on. IEEE, 2010
- [15] Ma, Xiaoyi. (2006) "Champollion: A Robust Parallel Text Sentence Aligner", LREC 2006: The Fifth International Conference on Language Resources and Evaluation
- [16] Matsoukas, Spyros, Antti-Veikko I. Rosti, and Bing Zhang. (2009) "Discriminative Corpus Weight Estimation for Machine Translation", In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore : 708--717
- [17] Munteanu, Dragos Stefan, and Daniel Marcu. (2005) "Improving machine translation performance by exploiting non-parallel corpora", Computational Linguistics 31.4 : 477-504.
- [18] Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. (1993) "The mathematics of statistical machine translation: Parameter estimation", Computational linguistics, vol. 19, 1993 : 263– 311.
- [19] Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. (2002) "BLEU: A Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics :311-318
- [20] Pauls, Adam, and Dan Klein. (2011) "Faster and smaller n-gram language models", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics., Portland, Oregon.
- [21] Resnik, Philip, and Noah A. Smith. (2003) "The web as a parallel corpus", Computational Linguistics 29.3 (2003): 349-380.
- [22] Schwenk, Holger. (2008) "Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation", In Proc. of the International Workshop on Spoken Language Translation
- [23] Taşçı, Şerafettin, A. Mustafa Güngör, and Tunga Güngör. (2006) "Compiling a Turkish-English Bilingual Corpus and Developing an Algorithm for Sentence Alignment", International Scientific Conference Computer Science
- [24] Taghipour, Kaveh, Nasim Afhami, Shahram Khadivi and Saeed Shiry. (2010) "A discriminative approach to filter out noisy sentence pairs from bilingual corpora", Telecommunications (IST), 5th International Symposium on 2010 : 537-541.
- [25] Tiedemann, Jörg. (2009) "News from opus - a collection of multilingual parallel corpora with tools and inter-faces", In Recent Advances in Natural Language Processing, volume V, Amsterdam/Philadelphia
- [26] Tyers, Francis M., and Murat Serdar Alperen. (2010) "SETimes: a parallel corpus of Balkan languages", In: Proceedings of the multiLR workshop at the language resources and evaluation conference, LREC2010, Malta, pp 49–53
- [27] Yasuda, Keiji, Ruiqiang Zhang, Hirofumi Yamamoto and Eiichiro Sumita. (2008) "Method of Selecting Training Data to Build a Compact and Efficient Translation Model", In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Hyderabad, India
- [28] Yıldız, Eray, Tantuğ, A.Cüneyd. (2012) "Evaluation of Sentence Alignment Methods for English-Turkish Par-allel Texts", LREC 2012: The International Conference on Language Resources and Evaluation. Istanbul
- [29] Khadivi, Shahram, and Hermann Ney. (2005) "Automatic filtering of bilingual corpora for statistical machine translation."Natural Language Processing and Information Systems. Springer Berlin Heidelberg : 263-274.
- [30] Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, ... and Evan Herbst . (2007) "Moses: Open source toolkit for statistical machine translation", In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions : 177-180.

AUTHORS

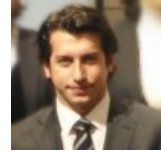
Eray YILDIZ

Graduate Student in Yildiz Technical University, Computer Engineering Department, Istanbul, Turkey B.Sc. : Computer Engineering, Kocaeli University, 2011



Ahmed Cüneyd TANTUĞ

Assistant Professor, Istanbul Technical University Computer Engineering Department, Turkey.
Ph.D. : Computer Engineering, Istanbul Technical University, 2007
M.Sc. : Computer Engineering, Istanbul Technical University, 2002
B.Sc. : Control & Computer Engineering, Istanbul Technical University, 2000



Banu DİRİ

Associate Professor, Yildiz Technical University, Computer Engineering Department , Istanbul, Turkey
Ph.D. : Computer Engineering, Yildiz Technical University, 1999
M.Sc. : Computer Sciences Engineering, Yildiz University, 1990
B.Sc. : Computer Sciences Engineering, Yildiz University, 1987



VECTOR ANT COLONY OPTIMIZATION AND TRAVELLING SALESMAN PROBLEM

Chiranjib Patra¹, Pratyush²

¹Department of Information Technology
Calcutta Institute of Engineering and Management, Kolkata-40
chiranjibpatra@gmail.com

²Department of Information Technology
Jadavpur University Sec-3, Block-LB/8, Salt-Lake, Kolkata- 700098
pratyush.cse1987@gmail.com

ABSTRACT

This paper introduces Vector Ant Colony Optimization (VACO), a distributed algorithm that is applied to solve the traveling salesman problem (TSP). In Any Colony System (ACS), a set of cooperating agents called ants cooperate to find good solutions of TSPs. Ants cooperate using an indirect form of communication mediated by pheromone they deposit on the edges of the TSP graph while building solutions. The proposed system (VACO) based on basic ACO algorithm with well distribution strategy in which the entire search area is initially divided into 2^n number of hyper-cubic quadrants where n is the dimension of search space for updating the heuristic parameter in ACO to improve the performance in solving TSP. From our experiments, the proposed algorithm has better performance than standard bench mark algorithms.

KEYWORDS

Ant colony optimization, traveling salesman problem, pheromone, global minima, VACO.

1. INTRODUCTION

In recent years, many research works have been devoted to ant colony optimization (ACO)[1,2,3,9] techniques in different areas. It is a relatively novel meta-heuristic technique and has been successfully used in many applications especially problems in combinatorial optimization. ACO algorithm models the behavior of real ant colonies in establishing the shortest path between food sources and nests. Ants can communicate with one another through chemicals called pheromones in their immediate environment. The ants release pheromone on the ground while walking from their nest to food and then go back to the nest. The ants move according to the amount of pheromones, the richer the pheromone trail on a path is, the more likely it would be followed by other ants. So a shorter path has a higher amount of pheromone in probability, ants will tend to choose a shorter path. Through this mechanism, ants will eventually find the shortest path. Artificial ants imitate the behavior of real ants, but can solve much more complicated problem than real ants can. [14,15].

ACO has been widely applied to solving various combinatorial optimization problems such as Traveling Salesman Problem (TSP)[4], Job-shop Scheduling Problem (JSP), Vehicle Routing Problem (VRP), Quadratic Assignment Problem (QAP), etc. Although ACO has a powerful capacity to find out solutions to combinatorial optimization problems, it has the problems of stagnation and premature convergence and the convergence speed of ACO is very slow. Those problems will be more obvious when the problem size increases. Therefore, several extensions and improvements versions of the original ACO algorithm were introduced over the years.

We have proposed new optimization algorithm (vector ant colony optimization) based on ACO for solving travelling salesman problem both for discrete and continuous domains. The search in this optimization technique uses number of ants, dimension, number of iteration, upper pheromone value, and lower pheromone value and the search process of the optimization is directed towards the region of hypercube in a multidimensional space where the amount of pheromone deposited is maximum after predefined number of iterations. The entire search area is initially divided into 2^n number of hyper-cubic quadrants where n is the dimension of search space. Each ant traverse the path equals to number of iteration time. We short the coordinates according to the distance from the source. The node which is the nearest from the source has maximum dimension (i.e. number of nodes in the simulation -1) and the node which has largest distance from the source has lowest dimension (i.e. zero). Here, dimension of a point denotes that a node have a path to how many nodes, Or how many unexplored edges it has. The VACO system uses pheromones updates to find the shortest path from source to destination. After a number of iteration we find global minima for source to destination. A global minimum is found with the help of artificial ants. When we call the algorithm a number of artificial ants uses best ant technique to find the path and based on which we found the global minimum of that path.

The main objective of our work is applying VACO in Travelling Salesman Problem and compares it with other bench mark algorithms. The paper is organized as follows 1. Introduction. 2. ACO Background 3.Travelling Salesman Problem 4.Vector Ant Colony Optimization 5. Using VACO for TSP tour construction. 6. Comparison of performance for VACO TSP implementation versus other algorithm 7. Conclusion

2. ACO BACKGROUND

Ant System was first introduced and applied to TSP by Marco Dorigo [5, 9]. Initially, each ant is randomly put on a city. During the construction of a feasible solution, ants select the following city to be visited through a *probabilistic decision rule*. When an ant k states in city i and constructs the partial solution, the probability moving to the next city j neighboring on city i is given by $\tau_{ij} = (1 - \rho) * \tau_{ij} + \Delta * \tau_{ij}$ ($h \in N^k$) where, τ_{ij} is the intensity of trails between edge (i and j) and η_{ij} is the heuristic visibility of edge (i, j), and $\eta_{ij} = 1/d_{ij}$. N^k is a set of city which remains to be visited when the ant is at city i . α and β are two adjustable positive parameters that control the relative weights of the pheromone trail and of the heuristic visibility.

After each ant completes its tour, the pheromone amount on each path will be adjusted with equation

$$\tau_{ij} = (1 - p) * \tau_{ij} + \Delta_{\tau_{ij}}$$

$(1-p)$ is the pheromone decay parameter ($0 < p < 1$) where it represents the trail evaporation when the Ant chooses a city and decide to move.

$\Delta_{\tau_{ij}}$ is defined as

$$\Delta_{\tau_{ij}} = \begin{cases} F(k), & \text{if edge } (i, j) \text{ is part of the solution constructed by ant } k, \\ 0 & \text{otherwise,} \end{cases}$$

$F(k) = 1/L_k$ where L_k is the cost of k^{th} ant tour.

3. TRAVELING SALESMAN PROBLEM (TSP)

The traveling salesman problem (TSP)[4] is the problem of finding a shortest closed tour which visits all the cities in a given set. In this article we take the data set of China, Greece, Burma and Argentina and we assume the TSP graph is completely connected.

TSP asks for the shortest roundtrip of minimal total cost visiting each given city (nod) exactly once. TSP is an NP- hard[11,12,13,16] problem and it is so easy to describe and so difficult to solve. The definition of a TSP is: given N cities, if a salesman starting from his home city is to visit each city exactly once and then return home, find the order of a tour such that the total distances (costs) traveled minimum. Cost can be distance, time, money, energy, etc. or a combination of two or more factor. In this paper, we assume that the distance between two cities is their Euclidean distance. Namely, each distance between cities i and j is $d(i, j) = d(j, i) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. Given a tour T , TSP is to find a tour which minimizes the objective function $S: S = \sum d(i, j)$.

4. VECTOR ANT COLONY OPTIMIZATION

We have proposed an efficient ant colony optimization function namely Vector ant colony optimizations (VACO) technique for optimizing mathematical functions. The search process of the optimization approach is directed towards the region of hypercube in a multidimensional space where the amount of pheromone deposited is maximum after predefined number of iterations. The entire search area is initially divided into 2^n number of hyper cubic quadrants where n is the dimension of search space. Then the pheromone level of each quadrant is measured. Now the search jumps to the new region of max pheromone level and restarts the search process in the new region. However the search area of new region is reduced compared to the previous search area. Thus the search advances and jumps to anew search space with reduced search area in several stages until the algorithm is terminated. The space of the new search region is smaller than the previous hyper-cubic search area. The reduction of search space is done along all dimensions. The pace is reduced in multiple stages with progress of the search process. If the search space is reduced slowly, then the possibility to come out of local optima and the convergence possibility to the global optimum are increased. On the other hand if the search space is reduced faster then there is a possibility to miss the global optimum since the process has no back tracking capabilities.

4a. Function optimization using VACO

The global optimization problem can generally be formulated as a pair (S, f) where S is the subset of R^n is a bounded set on R^n and $f: S \rightarrow R$ is an n – dimensional real valued function. The objective of the problem is to find a point x_{opt} belongs to S on R^n such that $f(x_{opt})$ is a global optimum on S . We have to find x_{opt} belongs to S according to the following equation, for min or max problems respectively:

$$\forall x \text{ that is subset of } S: f(x_{opt}) \leq f(x) \dots \dots \dots (1)$$

$$\forall x \text{ that is subset of } S: f(x_{opt}) \geq f(x) \dots \dots \dots (2)$$

Where f may not be continuous, but bounded.

Initially VACO algorithm starts searching to find the optimum in the entire search space. The search space is divided into a number of quadrants depending on the problems dimensionality in the multi-dimension space where each quadrant will form a hypercube. The search space partitioning is necessary to measure the pheromone level in each partition. If the problem dimension is denoted by n then the number of quadrants will be calculated as follows.

$$q = 2^n \dots \dots \dots (3)$$

The VACO method runs for a certain number of iterations say I_k and measures the pheromone level in each quadrant after the completion of I_k iterations. The pheromone level is measured to direct the search process towards the area with maximum amount of pheromone. We have considered the amount of pheromone deposited in each iteration as p defined by the following expression $p = 1/n$.

VACO technique runs iteratively in multiple stages and we find the quadrant in each iteration in which the best value of that iteration lies. Then the pheromone level of the corresponding quadrant is increased. The amount p_j in the j^{th} quadrant where $j \in \{1, 2, 3, \dots\}$ is increased by $1/n$ in each iteration. Once the optimization method completes I_k iteration the amount of pheromone p_j deposited in each quadrant is calculated. If the amount of deposited pheromone p_m in the m^{th} ($1 \leq m \leq 2^n$) quadrant is maximum, the search then moves towards the m^{th} quadrant. The search space is re-defined around. The ant population is regenerated except the elite one. The VACO approach restarts in the new search area and continues for I_k times before its transferred to another new space considering the highest pheromone level. The VACO finally terminates on the completions of I_{max} iterations.

4b. Algorithmic representation of VACO technique

- STEP 1: Initialize the population of ants and other parameters and iteration $I=1$;
- STEP 2: Create solutions for all ants and partition the search space into 2^n quadrants.
- STEP 3: Find the quadrant where the best solution lies in each iteration. Increase the pheromone level of the corresponding quadrant by $1/n$ following STEP 4.
- STEP 4: Identify the quadrant q_m with highest amount of pheromone deposited after the completion of I_k iterations.
- STEP 5: Refine the search space surrounding the area of the quadrant q_m and regenerate the population of ant following the elitist model. Move the search to the new search space

which is smaller in size than the previous search space and restart the search space process.

STEP 6: Increment $I = I + I_k$ if $I \leq I_{\max}$ then go to STEP 2.

STEP 7: Stop.

4c. Example of Search Space Reduction used in VACO method

Here we take 2-D search space and the repetitive reduction is described. In FIGURE –IA it's shown for the reduction of search space every after I_k (here we take 10) in three steps. Initially the highest number of solutions is in 2nd quadrant hence the intensity of pheromone in 2nd quadrant is $6 \times (1/2)$ and where as $2 \times (1/2), 1/2, 1/2$ for 1st, 3rd and 4th respectively, so VACO jumps to 2nd quadrant as it contains highest intensity of pheromone. Once this is determined, now let's reduce the search space by $u\%$ on both X and Y axis, thus generating X' and Y' respectively Figure II we thus again partition the space into four quadrants, again we find the 1st quadrant contain large number of solutions and then VACO jumps to 1st quadrant unlike previous step which in accordance with the concurrence with an elitist model of ACO. From Figure-IC again a new space is designated and a set of solutions exists in the respective four quadrants viz. . $x_{21} \dots x_{30}$. Now the pheromone concentration measurement is not necessary as the subdivision of the search space will not be economic as all the values are very close to one another. A chosen value close to the average of the found solutions is taken, which becomes the near-global optimum solution of the problem at the end of the final stage or at end of pre-specific number of iteration.

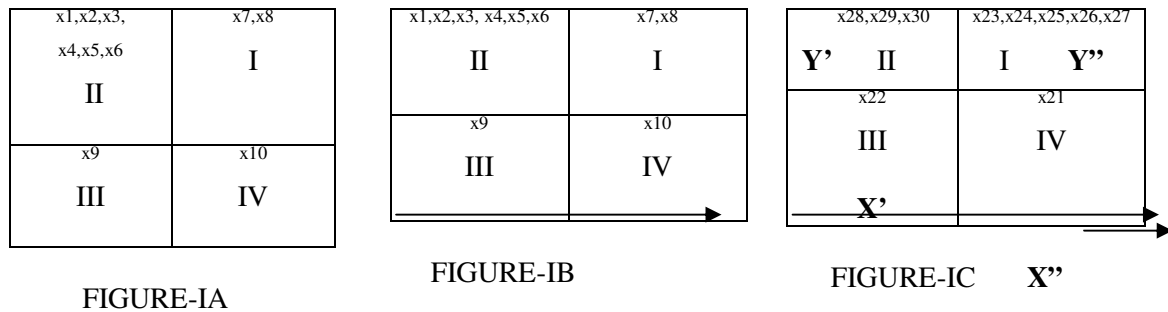


Fig-1: The implementation VACO on hypercube

5. USING VACO FOR TSP TOUR CONSTRUCTION

Considering the above assumptions we deduce the algorithm for TSP construction
Algorithmic representation:

- STEP 1: Initialize the co-ordinates.
- STEP 2: Initialize the starting position.
- STEP 3: Calculate the distance between starting node and all neighboring node.
- STEP 4: Sort the coordinates point on the basis of distance from source.
- STEP 5: Initialize the dimension for each sorted coordinates.
- STEP 6: Call the vector ant colony optimization for each sorted coordinates.
- STEP 7: Find out the coordinate which has lowest amount of global minima.
- STEP 8: New source is the coordinates which has the lowest minima from the source.

STEP 9: New travelling salesman problem space is STEP 1 coordinates point minus STEP 7 coordinate point.

STEP 10: Go to STEP 3.

STEP11: Stop.

In the given Fig 2, it is shown that we start from a point (x_i, y_i) and move towards that point that point has minimum value of global minima. By doing this technique iteratively we build a tour for TSP. The cost of this tour is the summation of all global minimum value between two connected points

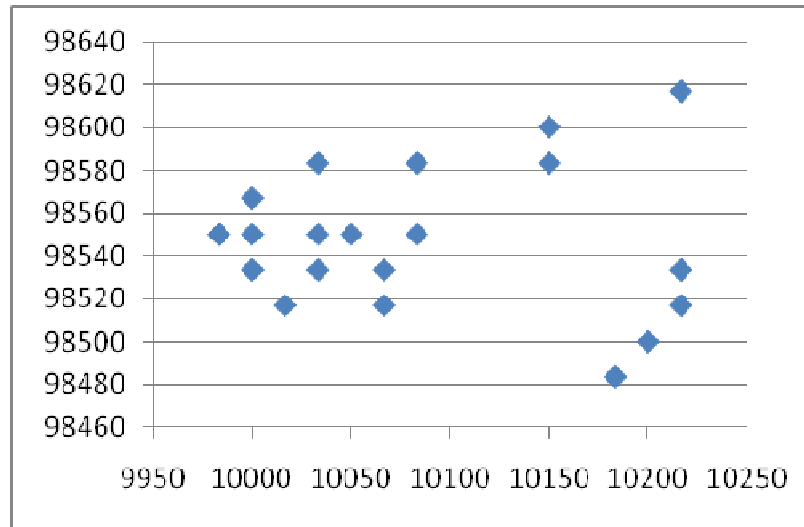


Fig-2A

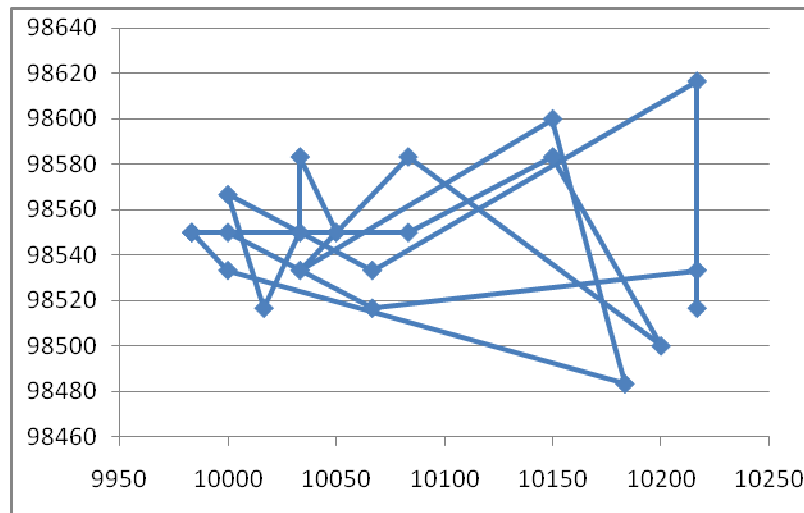


Fig-2B

Fig 2: Graphical Representation of 20 test nodes in space

The performance of VACO algorithm on a set of point is evaluated to find out how tour construction is efficient than other algorithms. In this setup we have considered a set of population size of ants as 50 till the end of the proposed optimization algorithm. We take 20 nodes for illustrating tour construction for which pheromone upper and lower limits are respectively +100 and -100 respectively. We select the path which has lowest value of global minima, we add that path to our resultant solution and dimension is indicated in the table. We did this procedure iteratively to complete the tour.

TSP Solution using Ant Colony Optimization Algorithm:

TSP function	Search Space	Dimension	Optimum value
	[9983.3333,10216.6667] [98483.3333,98616.6667]	19	793.3227
	[10000, 10216.6667] [98483.3333,98616.6667]	18	784.6709
	[10000, 10216.6667] [98483.3333,98616.6667]	17	789.0375
	[10000, 10216.6667] [98483.3333,98616.6667]	16	831.794
	[10000, 10216.6667] [98483.3333,98616.6667]	15	847.5783
	[10000, 10216.6667] [98483.3333,98616.6667]	14	889.6033
	[10000, 10216.6667] [98483.3333,98616.6667]	13	853.8038
	[10000, 10216.6667] [98483.3333,98616.6667]	12	811.2816
	[10000, 10216.6667] [98483.3333,98600]	11	886.8193
	[10000, 10216.6667] [98483.3333,98600]	10	892.7845
	[10000,10200] [98483.3333,98600]	9	907.407
	[10000,10200] [98483.3333,98600]	8	876.8203
	[10000,10200] [98483.3333,98600]	7	881.545
	[10000,10200] [98483.3333,98600]	6	962.8069
	[10000,10200] [98483.3333,98600]	5	889.4022
	[10000,10183.3333] [98483.3333,98600]	4	918.7506
	[10000,10183.3333] [98483.3333,98600]	3	852.6672
	[10000,10183.3333] [98483.3333,98600]	2	988.7577
	[10000, 10183.3333] [98483.3333,98533.3333]	1	986.0945
	[9983.3333,10000] [98533.3333,98550]	1	992.3018
Sum of global minima for all path			17637.25

6. COMPARISON OF PERFORMANCE FOR VACO TSP IMPLEMENTATION VERSUS OTHER ALGORITHM

The performance of VACO technique on a TSP functions was evaluated to find out how VACO is efficient than other algorithm.

In this setup we have considered a set of population size of ants as 50 till the end of the proposed optimization algorithm. For each algorithm VACO was run 50 times and an average of the 50 optimum results is tabulated in table 1 for VACO method. We have run VACO technique for I_k iterations to measure the amount of pheromone deposited in each quadrant for directing the

search in the redefined search space for the next stage. The new search space is generated by reducing $U\%$ of the length in all directions of the previous search space. For all algorithms tested for the TSP function, we have considered the value of I_k to be 5 and the value of u is set to 25. If the value of I_k is increased towards the higher side or more of the range, the search process slows down i.e. the convergence will require more number of iteration. On the other hand the value of u set to more than 25, the VACO may not converge to global or near global optimum solutions. Due fast reduction of the search space after I_k iteration. In this experiment we have tried to maintain the values I_k and u within the specified range so that VACO converges fast.

Function	Algorithm	MFE	G_{min}	SD
TSP function	OGA/Q	167,836	7.56×10^{-1}	1.1×10^{-1}
	M-L	13700	2934.78	133.674
	LEA	168,910	5.5×10^{-1}	1.08×10^{-1}
	VACO	15,550	0	1.77×10^{-13}

MFE denotes the average number of function evaluation to reach the desired value

G_{min} indicates the mean of the best value found in last generation for 50 runs

SD denotes the Standard Deviation

Table -1 Comparative study of VACO with other benchmark algorithm

7. CONCLUSION

In this work we studied the performance of Travelling Salesman Problem (TSP) using Vector Ant Colony Optimization. Simulation results revealed that TSP solution using our VACO method is capable of providing remarkably improved performance compared to other solution of TSP using other Optimization techniques. Similarly unlike traditional ACO the following statement is valid as based on the experiments, it can be seen that the quality of solutions depends on the number of ants. The lower number of ants allows the individual to change the path much faster. The higher number of ants in population causes the higher accumulation of pheromone on edges, and thus an individual keeps the path with higher concentration of pheromone with a high probability.

REFERENCES

- [1] Ramlakhan Singh Jadon, Unmukh Datta "Modified Ant Colony Optimization Algorithm with Uniform Mutation using Self Adaptive Approach for Travelling Salesman Problem" 4th ICCCNT 2013
- [2] A. Colorni, M. Dorigo, V. Maniezzo, "Distributed optimization by ant colonies". Proceedings of European Conference on Artificial Life, Paris, France, pp. 134-142.1991.
- [3] Krishna H. Hingrajiya, Ravindra Kumar Gupta, Gajendra Singh Chandel, "An Ant Colony Optimization Algorithm for Solving" International Journal of Scientific and Research Publications, Volume 2, Issue 8, August 2012.
- [4] Zar Chi Su SuHlaing, May Aye Khine "An Ant Colony Optimization Algorithm for Solving Traveling Salesman Problem" 2011 International Conference on Information Communication and Management.

- [5] Marco Dorigo, Luca Maria Gambardella, "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem" IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 1, NO. 1, APRIL 1997.
- [6] Zulaiha Ali Othman, Helmi Md Rais, and Abdul Razak Hamdan "DACS3: Embedding Individual Ant Behavior in Ant Colony System" International Journal of Computer, Information Science and Engineering Vol:2 No:8, 2008. [7] Marco Dorigo, Mauro Birattari, and Thomas Stutzle "Ant Colony Optimization: Artificial Ants as a Computational Intelligence Technique". 2001
- [8] Alberto Coloni, Marco Dorigo, Vittorio Maniezzo Dipartimento di Elettronica, Politecnico di Milano Piazza Leonardo da Vinci 32, 20133 Milano, Italy "Distributed Optimization by Ant Colonies"- EUROPEAN CONFERENCE ON ARTIFICIAL LIFE, PARIS, FRANCE, ELSEVIER PUBLISHING, 134-142
- [9] Thomas Stutzle, Manuel Lopez-Ibanez, Paola Pellegrini, Michael Maur, Marco Montes de Oca, Mauro Birattari, and Marco Dorigo "Chapter 8 Parameter Adaptation in Ant Colony Optimization"
- [10] Bifan Li, Lipo Wang, and Wu Song "Ant Colony Optimization for the Traveling Salesman Problem Based on Ants with Memory" 2008 IEEE.
- [11] David L. Applegate, Robert E. Bixby, Vasek Chvatal, William J. Cook. The Traveling Salesman Problem: A Computational Study. Princeton University Press, USA, 2007.
- [12] Elmedina Fejzagic, Adna Oputic "Performance Comparison of Sequential and Parallel Execution of the Ant Colony Optimization Algorithm for Solving the Traveling Salesman Problem" MIPRO 2013, MAY 20-24, 2013.
- [13] Dorigo M., Maniezzo V., Coloni A.: The Ant System: An autocatalytic optimizing process. Tech. Rep. 91-016 Revised, Dipartimento di Elettronica, Politecnico di Milano, Italy (1991).
- [14] Dorigo M., Maniezzo V., Coloni A.: Ant System: Optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics- Part B 26(1):29-41 (1996).
- [15] H. Md. Rais, Z. A. Othman, A.R. Hamdan, Improvement DACS3 Searching Performance using Local Search, Conference on Data Mining and Optimization, IEEE, 27-28 October 2009.
- [16] Ahuja, A. & Pahwa, A. (2005). Using ant colony optimization for loss minimization in distribution networks, Proceedings of the 37th Annual North American Power Symposium, pp. 470-474.

AUTHORS

Chiranjib Patra received his undergraduate and post graduate degrees from Calcutta University and Jadavpur University. Currently he is the candidate of PhD at the department of Information Technology, Jadavpur University. He is currently working as the Assistant Professor and Head of the Department, Information Technology at Calcutta Institute of Engineering and Management.

His interests are in application of evolutionary computing in wireless sensor networks and scale free networks.



Pratyush received his undergraduate degree from MMMEC, Gorakhpur and currently he is the candidate of M.E. (Software Engineering) at the Department of Information Technology, Jadavpur University.

His interests are in evolutionary computing and algorithms



INTENTIONAL BLANK

DETECTION AND TRACKING OF MULTIPLE OBJECTS IN CLUTTERED BACKGROUNDS WITH OCCLUSION HANDLING

Sukanyathara J and Alphonsa Kuriakose

Department of Computer Science & Engineering,
Viswajyothi College of Engineering& Technology,
MG University, Kerala, India
sukanyathara.j@gmail.com
alphonsakuriakose2014@gmail.com

ABSTRACT

Segmentation and tracking are two important aspects in visual surveillance systems. Many barriers such as cluttered background, camera movements, and occlusion make the robust detection and tracking a difficult problem, especially in case of multiple moving objects. Object detection in the presence of camera noise and with variable or unfavourable luminance conditions is still an active area of research. This paper proposes a framework which can effectively detect the moving objects and track them despite of occlusion and a priori knowledge of objects in the scene. The segmentation step uses a robust threshold decision algorithm which uses a multi-background model. The video object tracking is able to track multiple objects along with their trajectories based on Continuous Energy Minimization. In this work, an effective formulation of multi-target tracking as minimization of a continuous energy is combined with multi-background registration. Apart from the recent approaches, it focus on making use of an energy that corresponds to a more complete representation of the problem, rather than one that is amenable to global optimization. Besides the image evidence, the energy function considers physical constraints, such as target dynamics, mutual exclusion, and track persistence. The proposed tracking framework is able to track multiple objects despite of occlusions under dynamic background conditions.

KEYWORDS

Surveillance, segmentation, multi-background registration, threshold decision, energy minimization, tracking, computer vision.

1. INTRODUCTION

Segmentation and tracking plays an important role in Visual surveillance systems. Video tracking is the process of locating a moving object (or multiple objects) over time using a camera. Video tracking can be a time consuming process due to the amount of data that is contained in video. Adding further to the complexity is the possible need to use object recognition techniques for tracking, a challenging problem in its own right.

Video object segmentation, detection and tracking processes are the basic, starting steps for more complex processes, such as video context analysis and multimedia indexing. Object tracking in videos can be defined as the process of segmenting an object of interest from a sequence of video scenes. This process should keep track of its motion, orientation, occlusion and etc. in order to extract useful context information, which will be used on higher-level processes.

When the camera is fixed and the number of targets is small, objects can easily be tracked using simple methods. Computer vision-based methods often provide the only non-invasive solution. Their applications can be divided into three different groups: Surveillance, control and analysis. Under various environmental assumptions, several video object segmentation algorithms have been proposed. [6] - [8] proposes several simple and efficient video object segmentation algorithms. However, the proposed algorithms cannot address dynamic backgrounds because only one background layer is employed in their background model. Some algorithms are complex and require large amount of memory. Vosters *et al.* [9] proposed a more complex algorithm, consisting of an Eigen background and statistical illumination model, which can address sudden changes of illumination, but it has very high computational requirement.

To enable the long-term tracking, there are a number of problems which need to be addressed. The key problem is the detection of the object when it reappears in the camera's field of view. This problem is aggravated by the fact that the object may change its appearance thus making the appearance from the initial frame irrelevant.

Tracking algorithms estimate the object motion. Trackers require only initialization, are fast and produce smooth trajectories. On the other hand, they accumulate error during run-time (drift) and typically fail if the object disappears from the camera view. Research in tracking aims at developing increasingly robust trackers that track "longer". The post-failure behavior is not directly addressed. Detection based algorithms estimate the object location in every frame independently. Detectors do not drift and do not fail if the object disappears from the camera view. However, they require an offline training stage and therefore cannot be applied to unknown objects.

This paper intends to:

1. Propose a new method which combines multi-background registration based object detection to detect objects under dynamic backgrounds and tracking based on continuous energy minimization.
 2. To obtain better results despite of occlusions in complex backgrounds.
- The rest of the paper is organized as follows: The proposed system model is explained in section 3, 4, 5 and 6. In Section 7, conclusion of the work is given.

2. PROPOSED SYSTEM MODEL

In order to solve the problem of detection and tracking in cluttered backgrounds, a robust method which makes use of a Multi-background registration based object detection and Energy minimization based tracking is proposed in this paper. It is an enhanced method over the previous ones, and it is able to detect the area of interest in dynamic background tracks multiple moving objects along with their trajectories. Separate trajectories are assigned to the objects and those trajectories are not destroyed even if the object undergoes inter-object occlusion.

The segmentation method is memory efficient and it is able to detect objects under background clutter. The entire process consists of three major parts namely, Multi-background registration based segmentation, Threshold decision and Multiple-object tracking.

For detecting the moving objects, a background model is found out using multi-background registration and the foreground objects are detected using the built background model. The background model uses multiple background images which suits it for using in dynamic backgrounds. Apart the other methods, this is able to track multiple objects under dynamic backgrounds along with their trajectories.

The proposed system is an enhancement over the stationary background and single object tracking systems and include three main components:

1. An efficient threshold determination for segmentation.
2. Object detection.
3. Tracking multiple objects based on Continuous Energy Minimization.

3. THRESHOLD DECISION

To better deal with dynamic background conditions, an efficient threshold decision is inevitable. This paper makes use of Gaussianity test and Noise level estimation for efficient threshold decision.

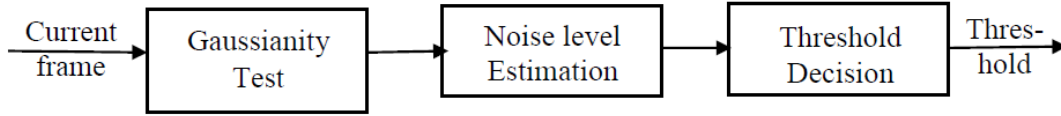


Fig. 1. Threshold decision

The Gaussianity test is applied to each block to determine if the minimal background differences in the block are Gaussian distributed or not. The camera noise is assumed to be Gaussian distributed.

3.1 Gaussianity Test

Divide the frame into a number of non-overlapping blocks of size $M_b * N_b$. Apply Gaussianity test to each block to determine if the minimal background differences in the block are Gaussian distributed or not. The Gaussianity test can be shown as the following equations:

$$I_r = \frac{1}{M_b * N_b} \sum_{m=1}^{M_b} \sum_{n=1}^{N_b} [BDmin(m, n)]^r \quad (1)$$

$$H(I_1, I_2, I_3, I_4) = I_3 + I_4 - 3I_1(I_2 - I_1^2) - 3I_2^2 - I_1^3 - 2I_1^4$$

- 1) Gaussian: $|H(I_1, I_2, I_3, I_4)| < G_{th}$
- 2) Non-Gaussian: $|H(I_1, I_2, I_3, I_4)| \geq G_{th}$

Where the smaller the H value, the closer the distribution of BD_min is to the Gaussian distribution, and G_th is the threshold value for binarizing the decision. If the minimal background differences in a block are Gaussian distributed, the block belongs to the background region because the (minimal) difference between the current frame and the background images is only caused by noise.

3.2 Noise Level Estimation and Threshold Decision

The optimal threshold $BDth^k$ is found out using the following equation:

$$BDth = \max\{|BDmin(i, j)| | (i, j) \in \text{Background blocks}\} \quad (2)$$

The background blocks are indicated by the Gaussianity test described in the previous section. Note that $BDth$ and $BDmin(i, j)$ are all random variables.

4. VIDEO OBJECT SEGMENTATION WITH MULTI-BACKGROUND REGISTRATION

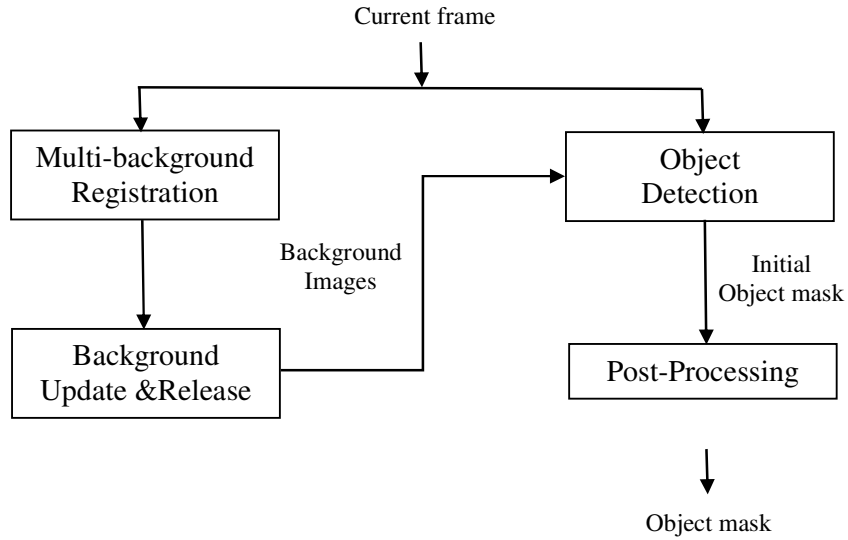


Fig. 2. Multi-background registration

The segmentation method is based on an online multilayer background modeling technique known as Multi-background registration (MBReg). The key concept in this algorithm is the fact that it models the background with N layers of background images instead of a single background layer. For each pixel position, the corresponding pixel in each layer of the background image represents one possible background pixel value.

As shown in Fig. 2. , the background model is established and maintained in the MBReg and background update and release blocks. In the MBReg block, each input pixel of the current frame $CurFrm(i, j, t)$, where (i, j) is the pixel position and t is the time index, is compared with the corresponding background pixels in the multi-background image $BImg(i, j, t - 1, k)$, where $k \in [1, N]$, and a matching flag, $match(i, j, k)$, is recorded by the following equation:

$$match(i, j, t, k) = \begin{cases} 1 & ; \text{if } BD(i, j, t, k) \leq BDth(i, j, t, k) \\ 0 & ; \text{otherwise} \end{cases} \quad (3)$$

The background difference $BD(i, j, t, k)$ can be calculated using the equation:

$$BD(i, j, t, k) = |CurFrm(i, j, t) - BImg(i, j, t - 1, k)| \quad (4)$$

In the background update and release block an ‘unmatched background’ counter $CntSno(i, j, k)$ and a weighting coefficient $Wgt(i, j, t, k)$ are maintained to record the duration when a background pixel is unmatched to the input pixel and the confidence of each background pixel where BDF is the background decaying factor.

$$CntSno(i, j, t, k) = \begin{cases} 0 & ; \text{if } match(i, j, t, k) = 1 \\ CntSno(i, t - 1, k) + 1 & ; \text{otherwise} \end{cases} \quad (5)$$

$$Wgt(i, j, t, k) = \begin{cases} Wgt(i, j, t - 1, k) + 1 & ; \text{if } match(i, j, t, k) = 1 \\ Wgt(i, j, t - 1, k) - 1 & ; \text{if } CntSno(i, j, t, k) > BDF(i, j, t, k) \\ Wgt(i, j, t - 1, k) & ; \text{otherwise} \end{cases} \quad (6)$$

Using the unmatched background counter and weighting coefficient, the background model can be updated or released with the following equations:

$$BImg(i, j, t, k) = \begin{cases} UpdBckgnd(i, j, t, k) & ; \text{if } match(i, j, t, k) = 1 \\ 0(\text{release}) & ; \text{if } k \in \text{Built background layers} \\ & \text{and } Wgt(i, j, t, k) < RELth(i, j, t, k) \\ 0(\text{release}) & ; Wgt(i, j, t, k) = 0 \end{cases} \quad (7)$$

5. TRACKING MULTIPLE OBJECTS

Tracking of multiple objects is seen as a function of continuous energy minimization here. Other than a number of recent approaches, it focus on designing an energy function that represents the problem as faithfully as possible. It uses a suitable optimization scheme to find strong local minima of the proposed energy. The scheme extends the conjugate gradient method with periodic trans-dimensional jumps. These moves allow the search to escape weak minima and explore a much larger portion of the variable-dimensional search space, while still always reducing the energy.

The aim of this method is to find an optimal solution for multi-target tracking over an entire video sequence. In otherwords, each target needs to be assigned a unique trajectory for the duration of the video, which matches the target’s motion as closely as possible. To this end, a global energy function is defined which depends on all targets at all frames within a temporal window, and thus represents the existence, motion and interaction of all objects of interest in the scene. Tracking is performed in world coordinates, i.e. the image evidence is projected onto the ground plane. Additionally, the evidence is weighted with a height prior to reduce false detections.

The state vector X consists of groundplane coordinates of all targets at all times. The (x, y) location of target i at frame t is denoted x_i^t and N indicates the total number of frames and targets

respectively. In this formulation the position of each target is always defined and considered when computing the energy, even in case of occlusion.

5.1 Energy Function

The energy function is made up of five terms: an observation term based on image data; three physically motivated priors for object dynamics, collision avoidance and object persistence; and a regularizer which tries to keep the number of trajectories low:

$$E(x) = E_{obs} + \alpha E_{dyn} + \beta E_{exc} + \gamma E_{per} + \delta E_{reg} \quad (8)$$

5.1.1 Observation Model

This makes use of the object detection step. Here, pedestrians are detected and interpreted as a kind of "intelligent smoothing", which takes into account the other energy terms rather than blindly smooth the nodes of the trajectory curve. It does however go beyond smoothing, for example it helps to prevent identity switches between crossing targets (since it favors straight paths).

5.1.2 Dynamic Model

It uses a constant velocity model:

$$E_{dyn}(x) = \sum_{t=1}^{F-2} \sum_i^N \|v_i^t - v_i^{t+1}\|^2 \quad (9)$$

Where $v_i^t = x_i^t - x_i^{t-1}$ is the current velocity vector of target i .

Dynamic model can be interpreted as a kind of "intelligent smoothing", which takes into account the other energy terms rather than blindly smooth the nodes of the trajectory curve.

5.1.3 Mutual Exclusion

The most obvious physical constraint is that two objects cannot occupy the same space simultaneously. This constraint is included to the energy function by defining a continuous exclusion term where s_g is the scale factor.

$$E_{exc}(x) = \sum_{t=1}^F \sum_{i \neq j} \frac{s_g}{\|x_i^t - x_j^t\|^2} \quad (10)$$

5.1.4 Target Persistence

Another constraint one would in most cases like to integrate into the energy function is the fact that targets cannot appear or disappear within the tracking area (but nevertheless can enter or leave the area). However, only a soft constraint is imposed, since otherwise one would have to explicitly model entry/exit locations (e.g. doors) and long term occlusion. Hence the sigmoid penalty:

$$E_{per}(x) = \sum_{t=1}^N \sum_{t \in \{1, F\}} \frac{1}{1 + \exp(1 - q \cdot b(x_i^t))} \quad (11)$$

where $b(x_i^t)$ and $b(x_j^t)$ are distances of the start, respectively end points of trajectory i to the border of the frame.

5.1.5 Regularization

The regularization drives the minimization towards a simpler explanation of the data, i.e. a model with fewer targets and longer trajectories:

$$E_{reg}(x) = N + \sum_{t=1}^N \frac{1}{F(i)} \quad (12)$$

where $F(i)$ is the temporal length of trajectory i in frames. The regularization balances the model's complexity against its fitting error, and discourages over-fitting, fragmentation of trajectories, and spurious identity changes.

Similar to any non-convex optimization, the result depends on the initial value from which the iteration is started. Empirically, even a trivial initialization with no targets work reasonably well, however it will take many iterations to converge.

The initialization uses the output of an arbitrary simpler tracker as a more qualified initial value. For initialization, per-target extended Kalman filters (EKFs) is used, where the data association is performed in a greedy manner using a maximum overlap criterion, to quickly generate a variety of starting values. The system can keep track of all the objects along with their trajectories.

6. CONCLUSIONS

In this paper we have proposed a novel method for detection and tracking of objects in complex backgrounds. Compared to the previous methods proposed for detection and tracking such as particle filter and extended Kalman filter, the method based on detection in complex backgrounds and multiple object tracking with continuous energy minimization robustly and efficiently detect and track multiple objects in complex environments for video surveillance. Moreover the method is robust to occlusions and it tracks without a priori knowledge of the number of targets, which is a difficult problem to tackle. Tracking is considered as a function of energy minimization which is more suitable for real world applications. The proposed detection method also overcomes the limitations of frame differencing based methods of segmentation.

REFERENCES

- [1] Shao-Yi Chien, Wei-Kai Chan, Yu-Hsiang Tseng, and Hong-Yuh Chen, "Video Object Segmentation and Tracking Framework with Improved Threshold Decision and Diffusion Distance", IEEE Trans. Circuits and Systems for Video Technology, vol. 23, no. 6, June 2013.
- [2] Shao-Yi Chien, Shyh-Yih Ma, and Liang-Gee Chen, "Efficient Moving Object Segmentation Algorithm Using Background Registration Technique", IEEE Trans. Circuits And Systems For Video Technology, Vol. 12, No. 7, July 2002.
- [3] W.-K. Chan and S.-Y. Chien, "Real-time memory-efficient video object segmentation in dynamic background with multi-background registration technique", IEEE Workshop Multimedia Signal Processing, 2007, pp. 219–22.
- [4] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Tracking video objects in cluttered background, IEEE Trans. Circuits Syst. Video Technology, vol. 15, no. 4, Apr. 2005.
- [5] Benjamin Langmann, Seyed E. Ghobadi, Klaus Hartmann, Otmar Loffeld, "Multi-Modal Background Subtraction Using Gaussian Mixture Models", IAPRS, Vol. XXXVIII, Part 3A – Saint-Mande, France, September 1-3, 2010.
- [6] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Realtime tracking of the human body," IEEE Trans. Pattern Anal. Machine Intell., vol. 19, no. 7, pp. 780–785, Jul. 1997.

- [7] S.-Y. Chien, S.-Y. Ma, and L.-G. Chen, "Efficient moving object segmentation algorithm using background registration technique," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 7, pp. 577–586, Jul.2002.
- [8] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 10, pp. 1337–1342, Oct.2003.
- [9] L. Vosters, C. Shan, and T. Gritti, "Background subtraction under sudden illumination changes," in *Proc. IEEE Int. Conf. Advanced Video Signal Based Surveillance*, Aug. 2010.
- [10] Anton Milan, Konrad Schindler, Stefan Roth, "Detection- and Trajectory-Level Exclusion in Multiple Object Tracking", in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, June 2013.

AUTHORS

Sukanyathara J received her B Tech degree in Computer Science & Engineering from M G University, Kerala, India in 2012. She is currently working towards the MTech degree in Computer Science and Engineering from MG University, Kerala, India. Her research interests include Digital image processing, Video Object Segmentation, Multi-target Tracking and Object Detection in real-world complex background scenarios.



Alphonsa Kuriakose received her M Tech degree in Computer Science & Engineering from M G University, Kerala, India in 2012. She is currently working as Assistant Professor in Computer Science and Engineering at Viswa jyothi College of Engineering and Technology, Kerala, India. Her research interests include Computer Networks, Information security, and Digital Image Processing.



HEAT STRESS RISK PREDICTION BY USING BAYESIAN NET MODEL WITH SENSOR NETWORK

Kanchan M. Taiwade¹ and Prof. Prakash S. Mohod²

¹Department of Computer Engineering, Nagpur University, Nagpur, India

kanchan.suryawanshi1111@gmail.com

²Head of Dept of CSE, GHRIETW, Nagpur, India

psmohod@gmail.com

ABSTRACT

With advancement in use of automation system, it is also desired to be able to know about the susceptible risk in advance for taking the preventive measures either automatically or manually. Disaster management is such an area where operatives wearing the suits and performing the activities are prone to the risk of heat stress which may cause mental impairments along with other serious effects leading to death. Such type of risk occurs in human body by not being able to compensate the heat generated into the surrounding air. The paper presents the concept of mechanism which can be used to prevent such situation by activating an alert to the operative or invoke cooling mechanism automatically before onset of the risk. The Bayesian Network Model is used to predict the onset of the risk. The model is based on the probabilities gives flexibility and simplicity in modeling the system. The system was trained with appropriate data and then compared with the real time parameters to check whether possibility of risk or not. Only those body parameters are considered which directly or indirectly participate in indicating heat stress or its onset.

KEYWORDS

Learning model, parameter selection, preventive mechanism, training data.

1. INTRODUCTION

The operatives of the disaster management crew for example a fire-fighting operative have to wear protective suits made up of specific material sometimes are of highly insulating material. These materials do not lead the heat to be passed outside of the suit leading to increased temperature inside the suit, also the operative performs some activities that also produces increase in the body heat, such situation can put the operative on the risk of heat stress because of the increased storage of heat, since body can not compensate it naturally because of the suit. By looking at the scenario, a system can be developed, in fact various efforts have been made in developing such system some of them were carrying an alert mechanism and some tried to accompany the fans along with the suit. The idea is to monitor the body parameters that actively participate in increasing the probability of onset of heat stress such as mean skin temperature, increased heart rate, decreased pulse rate, increased accelerations of the body; specifically limbs and arms, increased CO₂ level etc. any abnormality in any of the data which can lead to heat stress, if observed, the operative will be informed by invoking an alert mechanism. The problem

Natarajan Meghanathan et al. (Eds) : ICCSEA, SPPR, VLSI, WiMoA, SCAL, CNSA, WeST - 2014

pp. 49-57, 2014. © CS & IT-CSCP 2014

DOI : 10.5121/csit.2014.4713

with the alert mechanism is that even if the operative is informed by the alert mechanism, the operative may forget to actuate the preventive mechanism instead it is more desirable that the cooling system such as fan to be actuated automatically before the onset of the risk. For this purpose it is needed to monitor the body parameters on the regular time basis and if the parameter reading comes abnormal the fan is actuated automatically.

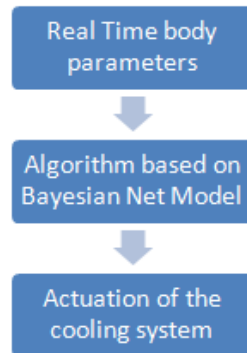


Fig1 : Execution phases

For checking if there is probability of onset of the risk or not a model need to be considered so that decision making will be based on this model. The model described in this paper is a Bayesian Net Model. Using this model for decision making by the system is quite simple as compared to other model since the Bayesian Network model does not require to have all and exact knowledge as input. The author has tried to develop such system for educational purpose.

2. METHODOLOGY

Since the real time parameters are to be monitored on time basis the system needed an external hardware consisting of sensors , analog-to-digital converters, relays etc to provide the real time input to the system. The basic focus is on the Bayesian algorithm with which modeling the dependencies between the parameters had been very easy because of the flexibility provided by this model. Basically the predictor was trained with the real time data; specifically the special cases; since not all the keen knowledge need to be provided. This is because the conditional dependencies can be modeled well using this model.

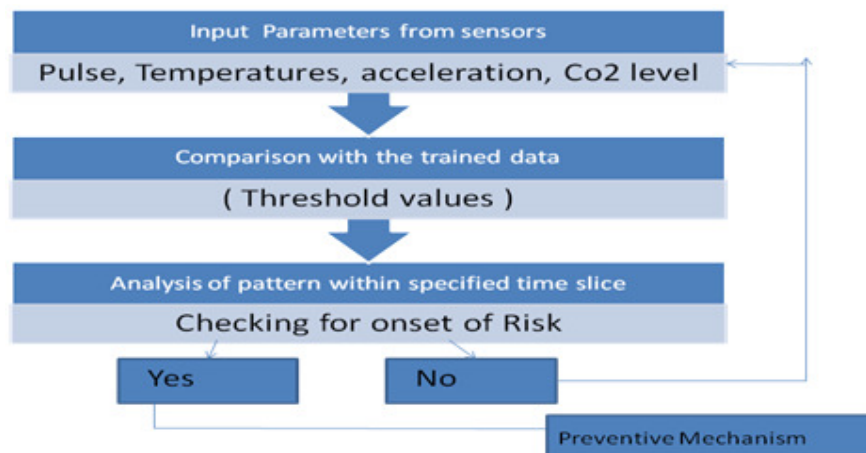


Fig. 2: Flow Diagram

The flow diagram has been shown in fig2. This figure gives an pictorial view of the actual execution of the project. It elaborates that the real-time physiological parameters such as inner and outer temperature, pulse, CO₂ level and accelerations will be monitored and analyzed with the help of training data for the desired time slice then the probability will be evaluated to see for the onset of the risk. With the increased value of probability, the preventive mechanism will be taken accordingly.

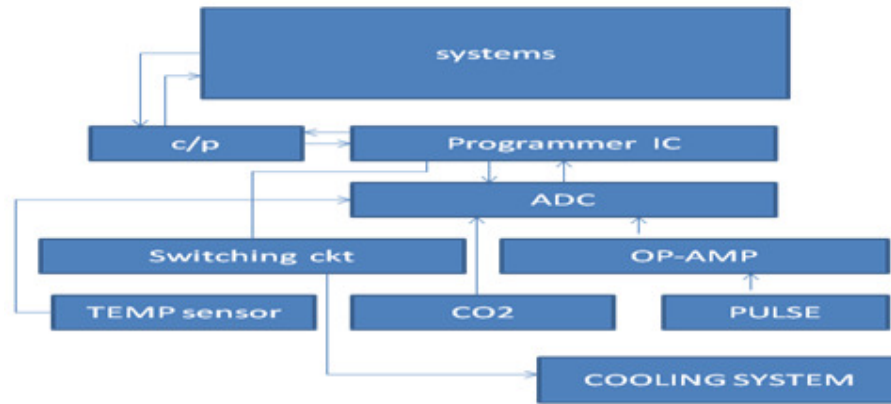


Fig3 : Experimental Setup

The parameter here considered are: inner temperature (inside the suit), outer temperature (outside of the suit), number of accelerations, pulse rate and CO₂ level. These parameters can be well used in order to predict if there is onset of the susceptible risk or not. The following block diagram shows the brief architectural model for gathering the real time parameter data as input the algorithm. Fig. 3 shows experimental setup that used for gathering the input data from the sensors. In the first module two parameters such as inner temperature and outer temperature were gathered and tested if a alert in the form of buzzer can be invoked if the threshold value that has been set extends. This was executed successfully. Fig. 4 shows testing of the first module. The testing of the first module was done successfully. The relays were used for switching between the temperatures: inner temperature and outer temperature. This is done to avoid number of analog to digital converter ICs. The circuit can be enhanced by many perspectives but since the focus is on the application of Bayesian Network Model author limited the involvement in hardware design since the aim with the hardware is to ultimately collect the sensor data and provide it to the system where the algorithm is residing for further processing.

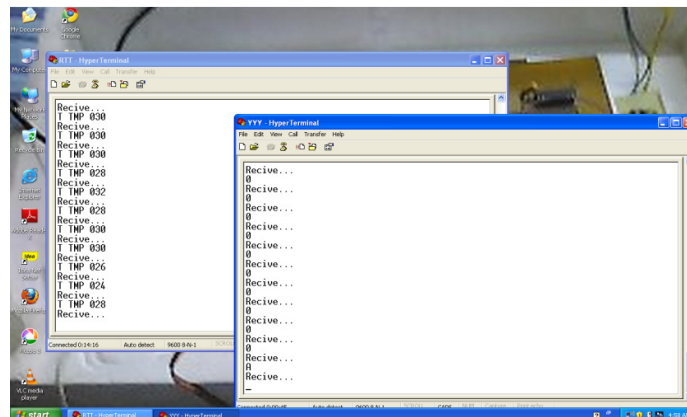


Fig.4: Testing of first module

The screenshot displays a Windows 7 desktop environment. A Notepad application window is open, titled 'log - Notepad', showing a log file with 40 entries. The log entries are as follows:

Timestamp	Direction	Device	Status
4/1/2014 12:54:38 PM	INT0 OUT	28 PLO 0 ACC 436	CO2 436
4/1/2014 12:54:38 PM	INT0 OUT	30 PLO 0 ACC 436	CO2 440
4/1/2014 12:54:38 PM	INT0 OUT	24 PLO 0 ACC 436	CO2 438
4/1/2014 12:54:39 PM	INT0 OUT	26 PLO 0 ACC 436	CO2 436
4/1/2014 12:54:40 PM	INT0 OUT	24 PLO 0 ACC 436	CO2 434
4/1/2014 12:54:40 PM	INT0 OUT	26 PLO 0 ACC 436	CO2 436
4/1/2014 12:54:41 PM	INT0 OUT	26 PLO 0 ACC 436	CO2 436
4/1/2014 12:54:41 PM	INT0 OUT	24 PLO 0 ACC 436	CO2 436
4/1/2014 12:54:42 PM	INT0 OUT	26 PLO 0 ACC 436	CO2 436
4/1/2014 12:54:42 PM	INT28 OUT	0 PLO 0 ACC 436	CO2 436
4/1/2014 12:54:43 PM	INT28 OUT	0 PLO 0 ACC 434	CO2 436
4/1/2014 12:54:43 PM	INT22 OUT	0 PLO 0 ACC 436	CO2 436
4/1/2014 12:54:44 PM	INT28 OUT	0 PLO 0 ACC 438	CO2 436
4/1/2014 12:54:44 PM	INT0 OUT	0 PLO 0 ACC 432	CO2 436
4/1/2014 12:54:45 PM	INT0 OUT	0 PLO 72 ACC 434	CO2 436
4/1/2014 12:54:45 PM	INT0 OUT	0 PLO 72 ACC 432	CO2 436
4/1/2014 12:54:46 PM	INT28 OUT	0 PLO 72 ACC 440	CO2 436
4/1/2014 12:54:46 PM	INT28 OUT	0 PLO 60 ACC 438	CO2 436
4/1/2014 12:54:47 PM	INT24 OUT	0 PLO 60 ACC 436	CO2 436
4/1/2014 12:54:47 PM	INT0 OUT	26 PLO 60 ACC 436	CO2 436
4/1/2014 12:54:48 PM	INT0 OUT	0 PLO 12 ACC 436	CO2 434
4/1/2014 12:54:48 PM	INT0 OUT	24 PLO 12 ACC 436	CO2 434
4/1/2014 12:54:49 PM	INT0 OUT	26 PLO 12 ACC 436	CO2 432
4/1/2014 12:54:49 PM	INT0 OUT	0 PLO 36 ACC 436	CO2 438
4/1/2014 12:54:50 PM	INT0 OUT	0 PLO 36 ACC 436	CO2 434
4/1/2014 12:54:50 PM	INT0 OUT	28 PLO 36 ACC 436	CO2 434
4/1/2014 12:54:51 PM	INT0 OUT	30 PLO 48 ACC 436	CO2 438
4/1/2014 12:54:51 PM	INT0 OUT	26 PLO 48 ACC 436	CO2 436
4/1/2014 12:54:52 PM	INT0 OUT	0 PLO 48 ACC 436	CO2 438
4/1/2014 12:54:52 PM	INT24 OUT	0 PLO 48 ACC 436	CO2 438
4/1/2014 12:54:53 PM	INT26 OUT	0 PLO 48 ACC 432	CO2 438
4/1/2014 12:54:53 PM	INT0 OUT	0 PLO 48 ACC 436	CO2 438
4/1/2014 12:54:54 PM	INT0 OUT	0 PLO 72 ACC 436	CO2 438
4/1/2014 12:54:54 PM	INT24 OUT	0 PLO 72 ACC 440	CO2 438
4/1/2014 12:54:55 PM	INT26 OUT	0 PLO 72 ACC 438	CO2 438
4/1/2014 12:54:55 PM	INT0 OUT	0 PLO 60 ACC 436	CO2 438
4/1/2014 12:54:56 PM	INT0 OUT	0 PLO 60 ACC 436	CO2 438
4/1/2014 12:54:56 PM	INT0 OUT	0 PLO 60 ACC 436	CO2 438

The taskbar at the bottom shows the Start button, taskbar icons for various applications, and the system clock showing 11:37 AM on 4/1/2014.

Fig. 5 : Testing result with inner/outer temperature, pulse, CO₂ level and accelerations

In fig5. We can see that the readings are showing the value zero at some places this is because of the use relays and switching; though it has been managed that the switching time is as minimum as possible. The relays have been used to keep the hardware minimum yielding simplicity, also to reduce the cost. The use of relays allowed author to use single analog-to-digital convertor for the required no. of parameter sensors.

3. USER INTERFACE

The user interface of the system has been designed as simple as possible. It hardly includes complexity of the internal system so that operating can be made easy to the new user of the system.

The real-time parameter readings collected from the external hardware are also shown in the user interface to provide the user with information about what is happening with the suit-wearers body. The graphs are added to get the quick view of the changes in the parameters. For adding the new data to get more and more accurate results the options are provided in the user interface. At the beginning it was avoided to add a large data base regarding the training data in order to minimize the comparisons so as to minimize the execution time.

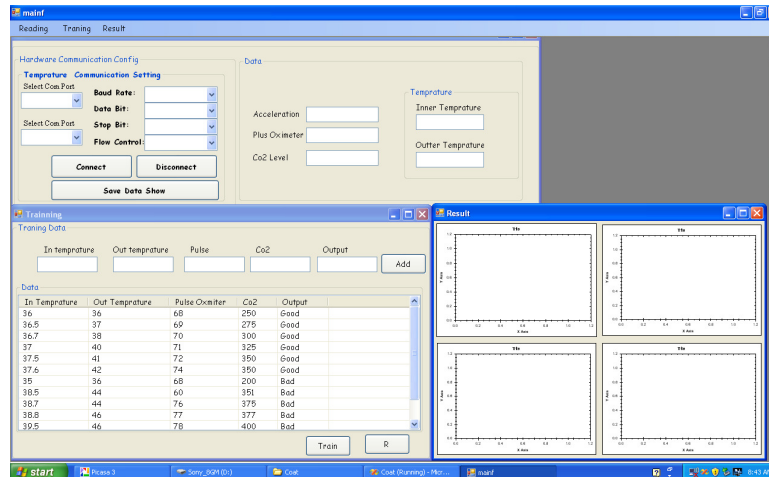


Fig 4: User Interface

4. IMPLEMENTATION

The project is designed in such a way that it can be viewed as an embedded one. The operative can carry the kit with him along with wired connection to the system (laptop) kept in the shoulder bag of the operative. The system can be used as a whole or can used as a single unit; only a CO₂ monitoring system along with other parameters or as a whole with CO₂ monitoring as well as with the suit specifically for temperature monitoring. The following fig. 5 gives of both the system. This gives the user to use the system according to the requirement of various environments.

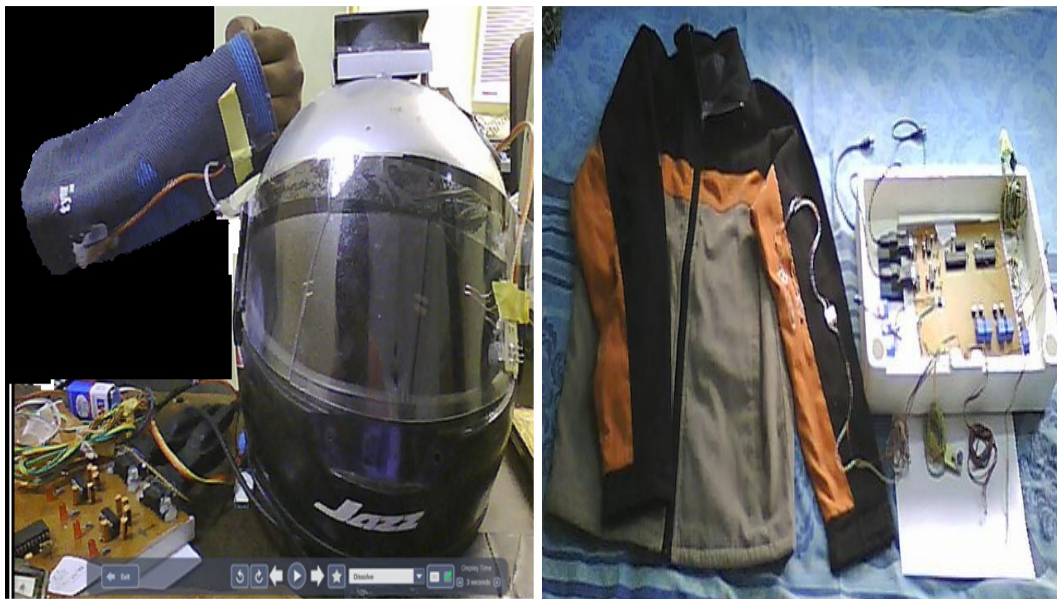


Fig 5. System can be used as a single unit or as a whole along with temperature monitoring.

4.1 Matching Technique

$T_{sk,t+1}$: Be the event when $T_{sk,t+1} > 36^{\circ}\text{C}$

Then Aim :

$$P(T_{sk,t+1}|S) = P(IT) P(OT) P(PI)P(Ac)P(CO_2)$$

Normal : { It, Ot, P, Ac, CO_2 }

Where each of the parameter with probability value :

1 : Good

0 : Bad

As shown the condition Normal is treated as a tuple consisting of the probabilistic values of the parameters sensed. The parameters acquires the value as good or bad depending on their threshold values. The threshold values are taken both ways as minimum and maximum. If a particular parameter value exist between the minimum and maximum value then it will be treated as a good value i.e as a 1. If the parameter value either is below the minimum value or beyond the maximum value then it will be treated as an abnormal i.e. bad and numerically 0.

Such various instances can be monitored at a regular time slice. A specific number of such patterns can be analyzed and evaluated to see for the probability that the temperature will reach the threshold value. Using the Bayesian Net [1] model for such type of prediction makes the modeling of the system simplified also it prevents from making complex calculations for evaluation and prediction as in the case of use of Kalman filtering approach [6].

4.2 Selection of the Parameters

The parameters in the system for monitoring are selected in such way that they are well-correlated and significantly contribute in the elevation of the body temperature. The skin temperature above 35°C reflects the inner temperature as mentioned in paper by Elena Gaura [1]. In this paper the author is not calculating the mean skin temperature to get the inner temperature as in [1], because pulse rate monitor has been used which specifies that if there is a increase in the pulse rate with this scenario then it is due to raise in inner temperature so increased pulse rate is a sign of increased inner temperature hence no need to calculate the inner temperature. The outer temperature parameter has been selected because the raise in outer temperature beyond the compensable rate increases the body inner temperature and the skin temperature. When body tries to cope with the raised temperature heart rate increases and to the breathing rate. If the helmet is wore on then CO_2 volume inside the helmet will also increase and will require exhausting the CO_2 to make breathing easy. Hence a fan is mounted on the helmet to exhaust the CO_2 and keep the helmet environment normal. While wearing a protective suit [1] if an operative performs some physical activities, as he has to during mission, such as climbing, crawling, walking this also increased body temperature hence an acceleration sensor is also used to monitor the accelerations. If there is abnormality in more than two to three parameters for specific time period then the user will be prompted through the buzzer or the preventive mechanism will be made on automatically. In some specific cases such as fire-fighting scenario [1] if the external temperature is extremely high, in such situation the system can be equipped with some extra components such dry ice pack with circulatory pipes inside the suit as mentioned in paper [1]. On the same basis if the external environment outside the helmet is consisting of low oxygen volume then in such cases the

operative can be provided with extra oxygen cylinder depending on which environment the system is going to be used.

5. RESULT

The system is evaluated with the real time parameters i.e. the readings were gathered with normal ranges as well as abnormal ranges while wearing an insulating material made suit as shown in fig. 5. The results obtained are shown as below in fig 6. With normal ranges of parameters obtained no action was taken out. As soon as the pattern of abnormal ranges were observed within the particular time slice the preventive mechanism, here it is fan, fan mounted on helmet, gets actuated automatically by the system as shown in fig. 7.

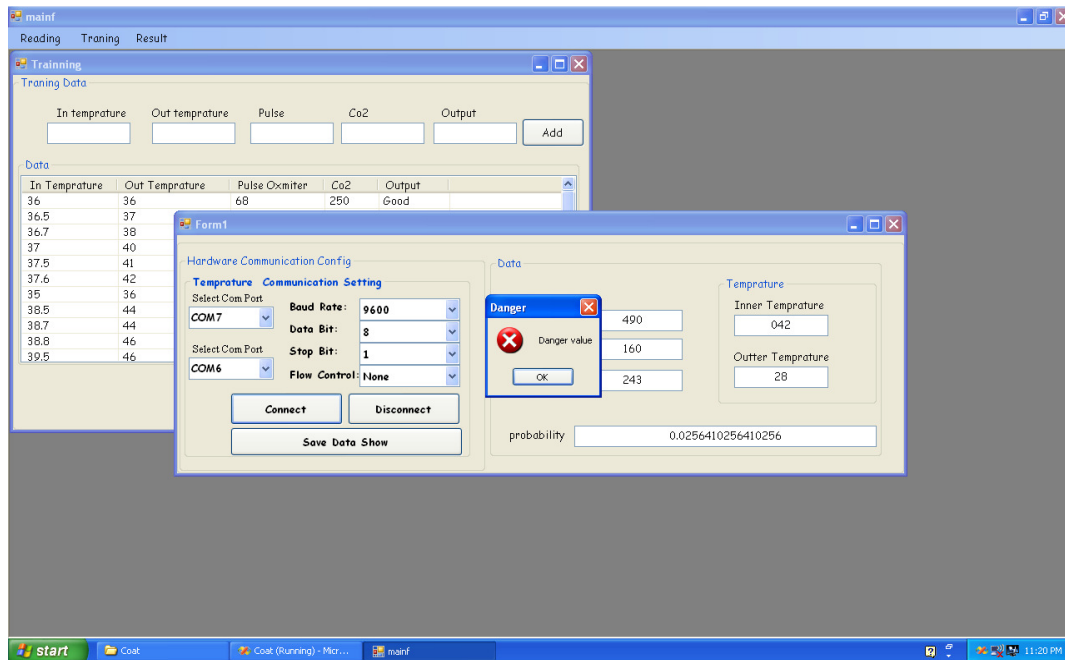


Fig. 6. With abnormality of the parameters the risk has been indicated.

In fig. 6 it can be observed that though the outer temperature is normal, there is increase in the accelerations leading to increase inner temperature indicated by the increase in pulse rate. Here it also be seen that with the increase in accelerations the increase in the CO₂ level is also noted.

This is how the parameters are co-related and play significance role in the elevation of the body temperature. Fig. 7 shows actuation of the fan mounted on the helmet as a preventive mechanism of the system. To be mounted on the suit it needs a specific designed suit with the circulatory mechanism for the air inside the suit.



Fig. 7. As a preventive mechanism the fan has been switched on.

6. CONCLUSION

An autonomous system for UHS prediction has been established with low cost by using the Bayesian Net Model as a predictor [1]. The cost of the project is minimized by using relays that switches the use of analog-to-digital convertor between various sensor but yet not affect the quality of the project. The execution time is also minimized by not calculating the mean skin temperature to get the inner temperature which requires temperatures sensors to be placed at upper and lower limbs. Getting the accurate inner temperature is not the focus here instead if there is sign of increased inner temperature then hot to cope with it is the focus, hence the increased pulse rate is taken as a indicator of the increased inner temperature above compensable rate. Implementing this concept has reduced the calculations as well as use of temperature sensors leading to reduced time for execution of the predictor.

7. FUTURE SCOPE

Multiple environment suitability : Refinement of the model parameters may allow the same prediction mechanism to be employed in a variety of other applications. Such as Coal-mining, Fire-fighting operations, soldier training scenarios etc.

ACKNOWLEDGEMENT

The author would like to thank Prof. P. S. Mohod for his guidance and valuable comments.

REFERENCES

- [1] Elena Gaura, *Member, IEEE*, John Kemp, and James Brusey, *Member, IEEE*, "Leveraging Knowledge From Physiological Data: On-Body Heat Stress Risk Prediction With Sensor Networks ", *IEEE transactions on biomedical circuits and systems* 2013.
- [2] C. W. Mundt, K. N. Montgomery, U. E. Udoh, V. N. Barker, G. C. Thonier, A. M. Tellier, R. D. Ricks, and R. B. Darling, "A multiparameter wearable physiologic monitoring system for space and terrestrial applications," *IEEE Trans. Inf. Technol. Biomed.*, vol. 9, no. 3, pp. 382–391, Sep. 2005.

- [3] R. Dilmaghani, H. Bobarshad, M. Ghavami, S. Choobkar, and C. Wolfe, "Wireless sensor networks for monitoring physiological signals of multiple patients," *IEEE Trans. Biomed. Circuits Syst.*, vol. 5, no. 4, pp. 347–346, Aug. 2011.
- [4] K. Li, S. Warren, and B. Natarajan, "Onboard tagging for real-time quality assessment of photoplethysmograms acquired by a wireless reflectance pulse oximeter," *IEEE Trans. Biomed. Circuits Syst.*, vol. 6, no. 1, pp. 54–63, Feb. 2012.
- [5] Y. Chuo, M. Marzencki, B. Hung, C. Jaggernaut, K. Tavakolian, P. Lin, and B. Kaminska, "Mechanically flexible wireless multisensor platform for human physical activity and vitals monitoring," *IEEE Trans. Biomed. Circuits Syst.*, vol. 4, no. 5, pp. 281–294, Oct. 2010.
- [6] E. Gaura, J. Brusey, J. Kemp, and C. D. Thake, "Increasing safety of bomb disposal missions: A body sensor network approach," *IEEE Trans. Syst., Man, Cybern. C, Applicat. Rev.*, vol. 39, no. 6, pp. 621–636, Nov. 2009.

AUTHORS

Kanchan M. Taiwade

M. Tech student of Computer Science and Engineering. GHRIETW.
Nagpur University 2012-2014.

Prof. Prakash S. Mohod

Head of the Dept of Computer Science and Technology G.H. Rasoni College of Engg. and Technology for Women. Nagpur.

INTENTIONAL BLANK

ORTHOGONAL DISCRETE FREQUENCY CODING SPACE TIME WAVEFORM FOR MIMO RADAR DETECTION IN COMPOUND GAUSSIAN CLUTTER

B. Roja Reddy¹ and M. Uttarakumari²

¹Department of Telecommunication Engineering,
R.V. College of Engineering, Bangalore, India.
rojareddyb@rvce.edu.in

²Department of Electronics and Communication Engineering,
R.V. College of Engineering, Bangalore, India
uttarakumari@rvce.edu.in

ABSTRACT

This paper proposes orthogonal Discrete Frequency Coding Space Time Waveforms (DFCSTW) for Multiple Input and Multiple Output (MIMO) radar detection in compound Gaussian clutter. The proposed orthogonal waveforms are designed considering the position and angle of the transmitting antenna when viewed from origin. These orthogonally optimized show good resolution in spikier clutter with Generalized Likelihood Ratio Test (GLRT) detector. The simulation results show that this waveform provides better detection performance in spikier Clutter.

KEYWORDS

Multiple Input and Multiple Output (MIMO), orthogonal Discrete Frequency Coded Space Time Waveforms (DFCSTW), Generalized Likelihood Ratio Test (GLRT), Compound Gaussian Clutter

1. INTRODUCTION

Multiple Input and Multiple Output system (MIMO) transmits multiple linearly independent probing signals via its transmit antennas and receives multiple coded waveforms from multiple locations. MIMO radar systems have many advantages including increased angle resolution [1-6], increased Doppler resolution [1, 7], reduced ground-based radar clutter levels [1], sharper airborne radar clutter notches [2,7], Lower Probability of Intercept (LPI) [1,8], and relaxed hardware requirements [1].

The performance of the transmitted waveforms is judged by their correlation properties [9-12]. The waveforms should have good autocorrelation properties for high range resolution and good cross-correlation for multiple target return separability. So, there is a need to design MIMO radar waveforms as orthogonal pulses with low correlation properties. In literature various algorithms have been proposed to design the orthogonal sequences with low autocorrelation and cross-correlation peak sidelobe levels. In [9], the focus is to design orthogonal Discrete Frequency Coding Waveforms_Frequency Hopping (DFCW_FF) for netted radar systems using simulated

annealing (SA) algorithm to optimize the frequency sequences, [10] and [11] focus on orthogonal DFCW_FF and orthogonal Discrete Frequency Coding Waveforms_Linear Frequency Modulation (DFCW_LFM) to design multiple orthogonal sequences with good correlation using a Modified Genetic Algorithm (MGA) technique. In [12] various Cyclic Algorithms (CA) for unimodular MIMO radar waveforms are designed for good correlation properties. Target Radar Cross Section (RCS) fades the received signal from the target. One way to maximize the system's processing gain is by antenna spacing [13-15]. In [16], adaptive space time waveform was proposed to improve detection performance. To improve the system performance, antenna spacing is one of the important factors.

Statistical characterization of the clutter is necessary for designing the detector. The clutter echoes result from very large number of elementary scatterers due to which the radar system has relatively low resolution capability. As the radar resolution increases, the statistics of the clutter has no longer been observed to be Gaussian. There is experimental evidence that high resolution radar systems are now plagued by target-like "spikes" that gives rise to non-Gaussian heavy tailed observations [17-18]. The high resolution sea clutter is modelled by compound Gaussian clutter which is a sample of compound K -distribution clutter. In [19-23] the focus is on the Generalized Likelihood Ratio Test (GLRT) detector to yield excellent performance and it is very much attractive for radar detection in the presence of correlated non-Gaussian clutter model.

In this paper, the proposed Orthogonal Discrete Frequency Coded Space Time Waveforms (DFCSTW) can improve the target detection in compound Gaussian (spikier) clutter. Section II illustrates the signal model and GLRT detector, section III shows the numerical simulations for the model developed in section II. Section IV concludes the paper.

2. WAVEFORM MODEL

Consider a MIMO radar system with a T_x transmitting antennas. Let x_i where $i=\{1,2,\dots,T_x\}$ denote the position of T_x transmitting antennas located at an angle θ_i when viewed from an origin. Each element may transmit N coding frequencies on each subpulse of a waveform. Each R_x receives and processes the signal from all the T_x transmitters. The received signals are the reflected signals from a target and clutter. Each element transmits N pulses with a Pulse Repetition frequency (PRF) of f_p .

2.1. Discrete Frequency Coding Space Time Waveform (DFCSTW)

Linear Frequency Modulation (LFM) is the first and probably the most popular pulse compression method. Discrete Frequency-coding Waveform (DFCW) has the large compression ratio. DFCW can lower correlation properties if sequences are coded properly. The basic idea is to sweep the frequency band (B) linearly during the pulse duration (T) and the time bandwidth product of the signal is BT . The spectral efficiency of the DFCW improves as the time-bandwidth product increases, because the spectral density approaches a rectangular shape. Here we consider the sequence length of each waveform (N) and Number of antennas (T_x).

The Discrete Frequency Coding Space Time (DFCSTW) Waveform is defined as

$$S_p(t, \Phi) = \begin{cases} \sum_{n=0}^{N-1} e^{j2\pi f_n^p(t-nT_p)} e^{j\pi k t^3} e^{j2\pi x_p \sin\Phi_p f_k / c_f}, & 0 \leq t \leq T_p \\ 0, & \text{elsewhere} \end{cases}, p = 1, 2 \dots T_x \quad (1)$$

where s is the frequency slope, $s=B/T$ and $p=1, 2, \dots, T_x$. T is the subpulse time duration. N is the number of subpulse that is continuous with the coefficient sequence $\{n_1, n_2, \dots, n_{T_x}\}$ with unique permutation of sequence $\{0, 1, 2, \dots, N\}$. $f_n^p = n \Delta f$ is the coding frequency of subpulse n of waveform p in the waveform. Δf is the frequency step. Where $x_p, i=\{1, 2, \dots, T_x\}$ denote the position of T_x transmitting antennas located at an angle θ_i when viewed from an origin. The choice of BT , $T\Delta f$ and $B/\Delta f$ values are crucial for the waveform design. Different lengths of firing sequence (N) have different values for each of the above mentioned parameters [11].

3. SIGNAL MODEL

The received signals for MIMO radar can be formulated as

$$r_i = S_i * T_i + S_i * Cl_i + V_i, \quad i = 1, 2, \dots, R_x \quad (2)$$

Where S is the transmitted code matrix. $T_i = [T_{i1}, \dots, T_{iT_x}]^T$, $i=1, 2, \dots, R_x$ are the complex values accounting for both the target backscattering. $V = [V_{i1}, \dots, V_{iT_x}]^T$, $i=1, 2, \dots, R_x$ are noise component. $r_i = [r_{i1}, \dots, r_{iN}]^T$, $i=1, 2, \dots, R_x$ are the echo signals of the i^{th} receiver antenna contaminated by the clutter. The clutter vectors n_i are assumed as compound Gaussian random vector i.e., [22]

$$Cl_i = \sqrt{\alpha_i \beta_i}, \quad i = 1, \dots, R_x \quad (3)$$

The texture α_i is non-negative random variable which models the variation in power that arises from the spatial variation in the backscattering of the clutter and the speckle components β_i are correlated complex circular Gaussian vectors and independent to one other. This α_i is independent Zero-mean complex circular Gaussian vector with covariance matrix.

$$R_i = E[n_i n_i^H] = \alpha_i r_o \quad (4)$$

Whereas $r_o = [Cl_i Cl_i^H]$ is the covariance structure and H is complex conjugate. The Compound Gaussian clutter is samples from K-distribution with pdf.

$$f(z) = \frac{\sqrt{2v/\mu}}{\Gamma(v)} \left(\sqrt{\frac{2v}{\mu}} z \right)^v K_{v-1} \left(\sqrt{\frac{2v}{\mu}} z \right) \quad (5)$$

The texture component $\sqrt{\tau_i}$ is gamma distribution with pdf

$$f(\tau_i) = \frac{1}{\sqrt{(v)}} \left(\frac{v}{\mu} \right)^v \sqrt{\tau_i}^{v-1} e^{-v/\mu \sqrt{\tau_i}} u(\sigma_i) \quad (6)$$

where $\Gamma(\cdot)$ is the Eulerian Gamma function, $v > 0$ is the parameter ruling the shape of the distribution, $u(\cdot)$ denotes the unit-step function, and $K_v(\cdot)$ is the modified second kind Bessel function with order v , which rules the clutter spikiness. The smaller the value of v , the higher is the tails of the distribution. The distribution will become Gaussian for $v \rightarrow \infty$.

The clutter has exponential correction structure of covariance matrix r_o , the (i, j) element of which is $\rho^{|i-j|}$ where ρ is one-lag correlation coefficient. The Power Spectral Density of clutter is

generally located in low frequency region & Clutter spread is controlled by v . The small the values of v the spikier is the clutter.

3.1. GLRT Detector

Suppose k ($k \geq N$) secondary data vector, sharing the same covariance structure of the primary data is available, r_i and r_{ik} , $i=1, 2, \dots, R_x$, $k=1, 2, \dots, k$ are the received signal from the primary and secondary data. Then, the detecting of a target with MIMO radar can be formulated in terms of the following binary hypotheses test.

$$\begin{aligned} H_0 : & \begin{cases} r_i = S_i * C_l + V_i, i = 1, 2, \dots, R_x \\ r_{ik} = S_{ik} * C_{lk} + V_{ik}, i = 1, 2, \dots, R_x, k = 1, 2, \dots, k \end{cases} \\ H_1 : & \begin{cases} r_i = S_i * T_i + S_i * C_l + V_i, i = 1, 2, \dots, R_x \\ r_{ik} = S_{ik} * T_{ik} + S_{ik} * C_{lk} + V_{ik}, i = 1, 2, \dots, R_x, k = 1, 2, \dots, k \end{cases} \end{aligned} \quad (7)$$

The GLRT detector [22] based on the primary data can be obtained by replacing the unknown parameters with their maximum likelihood estimates in the likelihood ratio. The GLRT detector [22] of the complex amplitude TH

$$\prod_{i=1}^r \frac{r_i^H r_o^{-1} r_i}{r_i^H (r_o^{-1} - r_o^{-1} S (S^H r_o^{-1} S)^{-1} S^H r_o^{-1}) r_o} \underset{H_0}{\overset{H_1}{>}} TH \quad (8)$$

where the TH is variable detection threshold. In a practical adaptive radar system, the covariance matrix of the clutter is estimated from a set of secondary data, which must be representative of the samples in the Cell Under Test (CUT). To make the detectors ensure the CFAR property w.r.t texture statistics, a normalized sample covariance matrix is adopted, based on the secondary data collected by the receiver antennas.

$$\hat{R}_{oi} = \frac{N}{K} \sum_{k=1}^k \frac{n_{i,k} n_{i,k}^H}{n_{i,k}^H n_{i,k}} \quad (9)$$

Substituting eq. (9) in eq. (8), we get the adaptive detector.

$$\prod_{i=1}^r \frac{r_i^H \hat{r}_{oi}^{-1} r_i}{r_i^H (\hat{r}_{oi}^{-1} - \hat{r}_{oi}^{-1} S (S^H \hat{r}_{oi}^{-1} S)^{-1} S^H \hat{r}_{oi}^{-1}) r_o} \underset{H_0}{\overset{H_1}{>}} TH \quad (10)$$

For a given value of N , as k varies the proposed adaptive detectors end up coincident with real scenario. However, for finite values of K , the performance of the estimate and eventually of the adaptive detector itself depends upon the actual values of N . Thus it is necessary to quantify the loss of the proposed decision strategy with respect to its non adaptive counterpart under situations of exact covariance matrix.

In order to compare the performance of the GLRT detector with Gaussian clutter GLRT detector (GC-GLRT),

$$\sum_{i=1}^r r_i \hat{r}_{oi}^{-1} S (S^H \hat{r}_{oi}^{-1} S)^{-1} S^H \hat{r}_{oi}^{-1} r_i > TH \quad (11)$$

In order to limit the computational burden, we assume P_{fa} as 10^{-4} and also to save the simulation time. The transmit code matrix S is the orthogonal DFCSTW and the Signal-to-Clutter Ratio (SCR) is defined as

$$SCR = \frac{\alpha^2}{NT_x} \text{tr}[S^H r_{oi}^{-1} S] \quad (12)$$

4. DESIGN RESULTS

Consider an orthogonal DFCSTW code set for 4*4 MIMO radar with code length of $N=8$. The simulation is carried out in MATLAB. The frequency code sets are optimized using ACC_PSO [24]. These sequences are considered in eq. (1) to generate the orthogonal DFCSTW set with good correlation properties. The above generated Code sequence matrix is used in the signal model mentioned in the section II to generate the DFCSTW where θ_i is the angle of the T_x transmitting antennas are generated randomly and placed linearly. Thus generated waveform is implemented in the signal model mentioned in the section III.

The P_{ds} of GLRT and of GC-GLRT are plotted versus SCR with $P_{fa}=10^{-4}$, $N=8$, $N_T=4$, $N_R=4$, $\rho=0.9$, $K=64$, $v=0.5$ in Fig.1. The performance of GLRT is better than GC-GLRT. Fig 2 shows the pds versus SCR for orthogonal DFCSTW code set and Space Time Code (STC) waveforms. It can be observed that the performance of orthogonal DFCSTW code is better than STC. The value of P_{ds} is 0.6 at -25 dB in [21] and the simulated result is 0.68 at the same dB value for DFCSTW code. This shows that the orthogonal DFCSTW code sets perform better in the spikier clutter than STC.

Fig 3 shows the pds versus SCR for orthogonal DFCSTW code set for different values of v . As the values of v decreases the clutter is spikier and the detector works better of the value of $v=0.3$ than $v=0.8$. Thus the DFCSTW code set has better detection performance in spikier environment. In Table 1, lists the values of P_{ds} versus SCR for -10 dB. It can be observed that the performance of this paper is much better than the existing statistics [20-22]. The curves show that the performance of GLRT with orthogonal DFCSTW code set is better in spikier clutter.

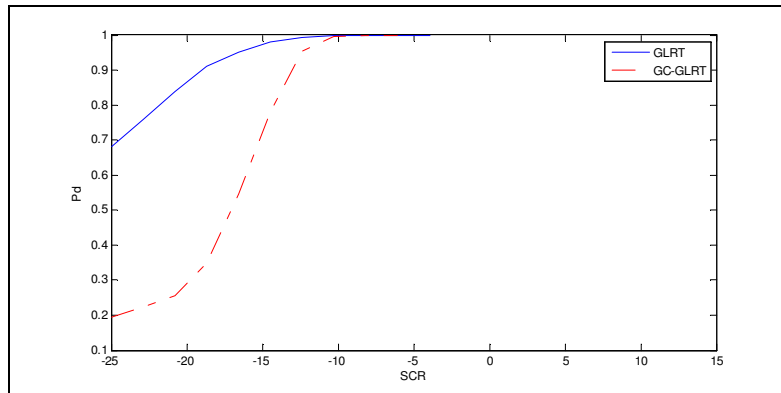


Figure 1. P_d versus SCR plots of GLRT (solid curves) and GCGLRT (dashed curves) receivers with Orthogonal DFCSTW Code set for $P_{fa}=10^{-4}$, $N=8$, $T_x=4$, $R_x=4$, $\rho=0.9$, $K=64$, $v=0.5$ parameter.

TABLE 1. Values of Pd Vs SCR

Literature	SCR (in dB)	Pd
[20]	-10	0.02
[21]	-10	0.95
[22]	-10	0.8
For the simulated result	-10	0.998

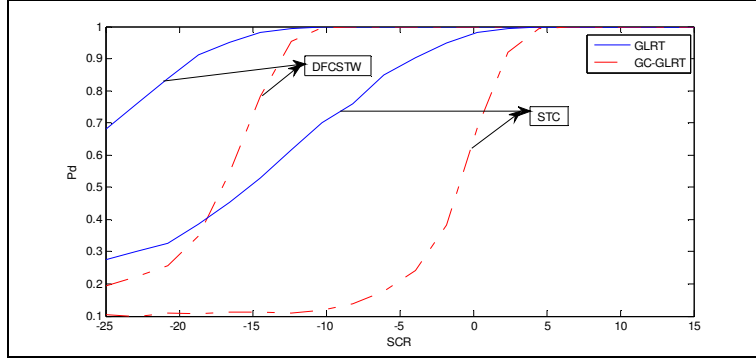


Figure 2. Pd versus SCR plots of GLRT (solid curves) and GCGLRT (dashed curves) receivers with Orthogonal DFCSTW Code set and STC for $P_{fa}=10^{-4}$, $N = 8$, $T_x=4$, $R_x=4$, $\rho = 0.9$, $K = 64$, $v=0.5$ parameter.

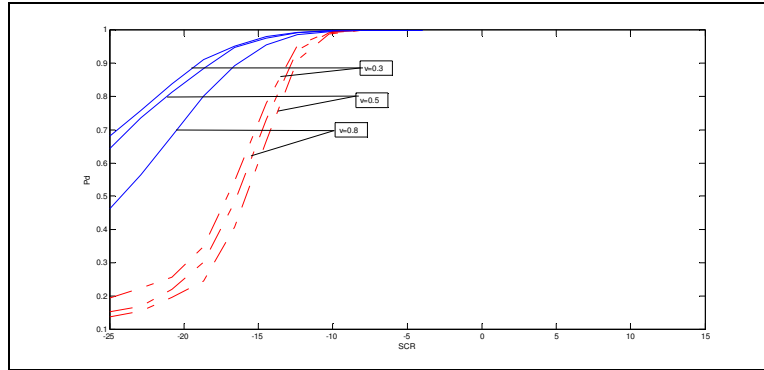


Figure 3. Pd versus SCR plots of GLRT (solid curves) and GCGLRT (dashed curves) receivers with Orthogonal DFCSTW Code set for $P_{fa}=10^{-4}$, $N = 8$, $T_x=4$, $R_x=4$, $\rho = 0.9$, $K = 64$, $v=0.3, 0.5$ and 0.8 as variable parameter.

5. CONCLUSIONS

In this paper, orthogonal DFCSTW is modelled to transmit in a spikier clutter environment. The results show that the performance of P_d is better for DFCSTW codes than STC waveform. The results also show that the DFCSTW code has the better performance in spikier clutter.

REFERENCES

- [1] D.J. Rabideau, P. Parker, (2003) "Ubiquitous MIMO multifunction digital array radar", Proc. Asilomar Conf. Signals, Syst., Comput., pp. 1057-64.
- [2] D.W. Bliss, K.W. Forsythe, (2003) "Multiple-input multiple-output radar and imaging: degrees of freedom and resolution", Proc. Asilomar Conf. Signals, Syst., Comput., vol. 1, pp. 54-9.
- [3] F.C. Robey, S. Coutts, et al, (2004) "MIMO radar theory and experimental results", Proc. Asilomar Conf. Signals, Syst., Comput., vol. 1, pp. 300-304.
- [4] Bekkerman, J. Tabrikian, (2006) "Target detection and localization using MIMO radars and sonars", IEEE Trans Sig Proc, vol. 54, pp. 3873-3883.
- [5] J. Tabrikian, (2006) "Barankin bounds for target localization by MIMO radars", IEEE Wksh. Sensor Array & Multichannel Sig. Proc., pp. 278-281.
- [6] K.W. Forsythe, D.W. Bliss, (2005) "Waveform correlation and optimization issues for MIMO radar", Proc. Asilomar Conf. Signals, Syst., Comput., pp. 1306-1310.
- [7] K.W. Forsythe, D.W. Bliss, et al, (2004) "Multiple-input multiple-output radar: performance issues", Proc. Asilomar Conf. Signals, Syst., Comput., vol. 1, pp. 310-315.
- [8] E. Fishler, A. Haimovich, R. Blum, L. Cimini, D. Chizhik, and R. Valenzuela, (2004) "MIMO radar: An idea whose time has come," in proc. of IEEE International Radar conference, Philadelphia, PA, pp No. 71-78.
- [9] Hai Deng, (2004) "Discrete frequency-coding waveform design for netted radar systems", in IEEE Signal Proces. letters, vol. 11, Issue 2, pp 179-182..
- [10] Bo Liu, Zishu He, Oian He, (2007) "Optimization of orthogonal discrete frequency-coding waveform based on modified genetic algorithm for MIMO radar", in Inter. Conf. on Commun. Circuits and Syst., Kokura, pp no. 966 – 970.
- [11] Bo Liu, Zishu He, Jun Li, (2008) "Mitigation of autocorrelation sidelobe peaks of orthogonal discrete frequency-coding waveform for MIMO radar", in proc. of IEEE Radar confer., China, Chengdu, pp 1-6..
- [12] Hao He, Petre Stoica, Jain Li, (2009) "Designing unimodular sequences sets with good correlation-Including an application to MIMO Radar", in IEEE Trans. on Signal Proces., Vol. 57, No. 11, pp 4391-4405.
- [13] Eran Fishler, Alexander Haimovich, Rick S. Blum, Leonard J. Cimini, Dmitry Chizhik, Reinaldo A. Valenzuela., (2006) "Spatial Diversity in Radars—Models and Detection Performance", IEEE Trans. on Sig Proc, Vol. 54, No. 3, pp. 823-837.
- [14] Haimovich, A.M., Blum, R.S., Cimini, L.J., (2008) "MIMO Radar with widely separated antennas", in IEEE Sig Proc Magazine, Vol 25, Issue 1, PP No. 116-129.
- [15] Jian Li, Petre Stoica., (2007) "MIMO Radar with Colocated Antennas" in IEEE Sig Proc Magazine, vol 24, Issue 5, pp No: 106-114.
- [16] Zengjiankui, Hezishu, Liubo, (2007) "Adaptive Space-time-waveform Processing for MIMO Radar", in Inter. Conf. on Commun., Circuits and Syst., Kokura, pp no. 641 – 643.
- [17] A. Farina, F. Gini, M.V. Greco, L. Verrazzani, (1997) "High resolution sea clutter data: a statistical analysis of recorded live data", IEEE Proc.-Radar, Sonar Navigation 144 (3) page: 121-130.
- [18] K.J. Sangston, K.R. Gerlach, (1994) "Coherent detection of radar targets in a non-Gaussian background", IEEE Trans. Aerospace Electron. Systems 30 (2), page: 330-340.
- [19] Rui Fa, Rodrigo C. de Lamare, (2010) "Knowledge-aided reduced-rank STAP for MIMO radar based on joint iterative constrained optimization of adaptive filters with multiple constraints", in IEEE Inter. Conf. on Acoust. Speech Sig Proc., Dallas, TX, pp. 2762 – 2765.
- [20] Antonio De Maio, Marco Lops, (2007) "Design Principles of MIMO Radar Detectors", in IEEE Trans. Aero. Elec. Syst., Vol 43, Issue 3, pp. 886-898.
- [21] Guolong Cui • Lingjiang Kong • Xiaobo Yang, (2012) "GLRT-based Detection Algorithm for Polarimetric MIMO Radar Against SIRV Clutter", in Circuits Syst Signal Process, Vol 31, pp. 1033-1048.
- [22] Guolong Cui, Lingjiang Kong, Xiaobo Yang, Jianyu Yang, (2010) "Two-step GLRT design of MIMO radar in compound-Gaussian clutter", in IEEE Radar Conf, Washington, DC, pp. 343 – 347.
- [23] A. De Maio, C. Hao, D. Orlando, (2014) "An Adaptive Detector with Range Estimation Capabilities for Partially Homogeneous Environment", in IEEE Sig. Proc. Letters, vol 21, no. 3, pp no- 325-329.

- [24] B. Roja Reddy, M. uttarakumari, (2012) "Generation of orthogonal discrete frequency coded waveform using accelerated particle swarm optimization algorithm for MIMO radar", Proceedings of the Second Inter. Conf. on Computer Science, Engg. and App. (ICCSEA 2012), New Delhi, India, Volume 1, pp 13-23.

AUTHORS

Smt. B. Roja Reddy received the B.E degree in 1998 from Gulbarga University, Karnataka and M. Tech degree in 2004 from VTU, Karnataka. Presently working at R.V. college of Engineering with an experience of 11 years in the teaching field. Her research interest lies in various areas signal Processing. Currently précising her Ph.D in Radar Signal Processing & MIMO Radar.



Dr. M Uttara Kumari received the B.E degree in 1989 from Nagarujna University, Hyderabad, Andhra Pradesh and M.E degree in 1996 from Bangalore University, Karnataka and Ph.D degree in 2007 from Andhra University. Presently working at R.V.College of Engineering with an experience of 19 years in the teaching field. Her research interest lies in various areas of radar systems, Space-time adaptive processing, speech processing and image processing.



ESTIMATION OF RECURSIVE ORDER NUMBER OF PHOTOCOPIED DOCUMENT BASED ON PROBABILITY DISTRIBUTIONS

Suman V Patgar¹, Rani K², Vasudev T²

¹P.E.T Research Foundation, P.E.S College of Engineering, Mandya, India,
571401

sumanpatgar@gmail.com

²Maharaja Research Foundation, Maharaja Institute of Technology Mysore,
Belawadi, S.R Patna, Mandya, India, 571438

ABSTRACT

Photocopy documents are very common in our normal life. People are permitted to carry and present photocopied documents to avoid damages to the original documents. But this provision is misused for temporary benefits by fabricating fake photocopied documents. Fabrication of fake photocopied document is possible only in 2nd and higher order recursive order of photocopies. Whenever a photocopied document is submitted, it may be required to check its originality. When the document is 1st order photocopy, chances of fabrication may be ignored. On the other hand when the photocopy order is 2nd or above, probability of fabrication may be suspected. Hence when a photocopied document is presented, the recursive order number of photocopy is to be estimated to ascertain the originality. This requirement demands to investigate methods to estimate order number of photocopy. In this work, a different approach based on probability distributions like normal, extreme values and exponential is proposed to estimate the recursive order number of the photocopied document under consideration. A detailed experimentation is performed on a generated data set and the method exhibits efficiency close to 84%.

KEYWORDS

fabricated photocopy documents, recursive order number, probability distributions

1. INTRODUCTION

Many authorities trust and accept the photocopied documents submitted by citizens as proof and consider the same as genuine. Few such applications like to open bank account, applying for gas connection, requesting for mobile sim card, concerned authorities insist photocopy documents like voter id, driving license, ration card, pan card and passport as proof of address, age, photo id etc to be submitted along with the application form. Certain class of people could exploit the trust of the authorities, and indulge in forging/ tampering/ fabricating photocopy document. These things would be deliberately made at the time of obtaining the photocopy of document without

damaging the original document. It is learned that in majority of the cases fabrications are made by changing/ replacing/ overwriting/ removing/ adding contents in place of authenticated content. The fabricated photocopy documents are generated to gain some short term and long term benefits unlawfully. This poses a serious threat to the system and the economics of a nation. The types of systems trusting photocopied document raise an alarm to have an expert system [1] that efficiently supports in detecting a forged photocopy document. The need of such requirement to the society has motivated us to take up research through investigating different approaches to detect fabrication in photocopy document. It is quite evident from the above discussions that the probability of fabrication is zero in the 1st photocopy obtained from the original document, where as the fabrication may be suspected in the higher order photocopy. In this direction, an attempt is made to estimate the recursive order number of the photocopy submitted. Further, based on the estimation of order number, investigations can be explored to detect the possibility of fabrication in photocopy.

Many research attempts are carried out on original documents like signature verification, detection of forged signature [2], handwriting forgery [3], printed data forgery [4], and finding authenticity of printed security documents [5]. Literature survey in this direction reveals that the above research attempts are made in the following issues: Discriminating duplicate cheques from genuine ones [5] using non-linear kernel function; Detecting counterfeit or manipulation of printed document [4] and this work is extended to classify laser and inkjet printouts; Recognition and verification of currency notes of different countries [6] using society of neural networks along with a small work addressing on fake currencies; Identification of forged handwriting [3] using wrinkles as a feature is attempted along with comparison of genuine handwriting. One of the interesting features is a measure of the variability of the handwriting on a small scale. Although one can copy the shape of another's handwriting, it is difficult to mimic the dynamic aspects, such as speed and acceleration. Because forged handwriting tends to be drawn slowly, when scanned, it might be more wiggly than the authentic handwriting. Forgery handwriting shows more wrinkliness than natural handwriting does. This wrinkliness feature can be measured using the *fractal* dimension measure.

While preparing any security document, the designers embed certain features that are considered as security features. It is generally assumed that these features are difficult to replicate or copy. Duplicity of a document is identified by checking these security features. Security features in documents like cheques, legal deeds, certificates, etc. are embedded by three attributes namely (i) color features, (ii) background artwork and logo, and (iii) paper quality. After the extracting features from a cheque document, its authentication is to be done. This is modeled as a 2-class pattern recognition problem, i.e. whether the document belongs to the genuine document class or not. Support Vector Machines(SVMs) [5] are used to verify authenticity of these cheques.

Further, in literature to the best of our knowledge no significant effort is noticed towards detecting forgery made while taking photocopy. As the domain under consideration is new for research, no standard data set is available. Hence for the purpose of experimentation sufficient numbers of sample photocopies are obtained on a RICOH AFICIO 2018D copier machine to generate different recursive order copies. The photocopies were scanned using a scanner to produce bitmap images at 300dpi. Fig 1a and 1b show the samples of 1st order and 5th order recursive photocopies of a document respectively.

With Prime Minister Manmohan Singh maintaining that the government is open to dialogue, Anna Hazare on Monday suggested that he should send his representatives for discussion on the JanLokpalBill. Kiran Bedi, one of the leading members of Team Anna, quoted Hazare as saying, "Let the government come forward to discuss the Jan Lokpal bill. Let the PM send his representatives." She also ruled out that any negotiations were on between the two groups. "(Talk of any negotiations are rumours," Bedi tweeted. The suggestion came within hours of the Prime Minister saying that the government is open to a "reasoned debate" on the Lokpal Bill and that the Parliamentary panel examining it can propose changes. "We are open to a reasoned debate on all

Fig 1a: 1st order photocopy

With Prime Minister Manmohan Singh maintaining that the government is open to dialogue, Anna Hazare on Monday suggested that he should send his representatives for discussion on the JanLokpalBill. Kiran Bedi, one of the leading members of Team Anna, quoted Hazare as saying, "Let the government come forward to discuss the Jan Lokpal bill. Let the PM send his representatives." She also ruled out that any negotiations were on between the two groups. "(Talk of any negotiations are rumours," Bedi tweeted. The suggestion came within hours of the Prime Minister saying that the government is open to a "reasoned debate" on the Lokpal Bill and that the Parliamentary panel examining it can propose changes. "We are open to a reasoned debate on all

Fig 1b: 5th order photocopy

Visual analysis performed on the recursive photocopied documents exhibits a relative degradation in the texture of the document. The degradation keeps relatively increasing on each recursive photocopy i.e., more the order of recursion, higher is the degradation in texture which is quite clear from fig 1a and 1b. This directed us to explore a texture analysis method to study the relative degradation in the recursive photocopies of documents. Earlier, two methods were proposed to find texture degradation. The first method uses Geometric Moments [7] to find texture degradation and the second method was proposed using entropy from Gray Level Co-occurrence Matrix [8]. The first approach showed an efficiency of 65%. In order to achieve higher efficiency exploration of another method was attempted on this problem to analyze texture degradation using probability distributions.

The remainder of the paper is organized as follows:. Section 2 gives introduction to Probability distributions used in the methods. Section 3 describes methodology adopted for estimation of order number of photocopy using texture feature. The experiments conducted along with analysis of results are discussed in section 4. Conclusion on the work is presented in section 5.

2. PROBABILITY DISTRIBUTIONS

In the proposed system, the recursive order of the photocopy documents is estimated based on measure of texture degradation using different distribution methodologies like Normal

distribution, Exponential distribution and Extreme value distribution. These distribution methods are applied to maximum peak values which are extracted from the distribution graphs and brief introduction to the same is given subsequently.

2.1 Normal distribution

In probability theory, the normal (or Gaussian) distribution is a very commonly occurring distribution function that tells the probability that an observation in some context will fall between any two real numbers. Normal distributions are extremely important in statistics and are often used in the natural and social sciences for real-valued random variables whose distributions are not known [9].

The Gaussian distribution is sometimes informally called the bell curve. However, many other distributions are bell-shaped (such as Cauchy's, Student's, and logistic). The terms Gaussian function and Gaussian bell curve are also ambiguous because they sometimes refer to multiples of the normal distribution that cannot be directly interpreted in terms of probabilities [10].

A normal distribution is,

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

The parameter μ in this definition is the *mean* or *expectation* of the distribution (and also its median and mode). The parameter σ is its standard deviation; its variance is therefore σ^2 . A random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate. If $\mu = 0$ and $\sigma = 1$, the distribution is called the standard normal distribution or the unit normal distribution, and a random variable with that distribution is a standard normal deviate[9].

2.2 Extreme value Distribution

The extreme value distribution has two forms. One is based on the smallest extreme and the other is based on the largest extreme. These are the minimum and maximum cases, respectively. The extreme value distribution is also referred to as the Gumbel distribution [11]. The general formula for the probability density function of the Gumbel (minimum) distribution is

$$f(x) = \frac{1}{\beta} e^{\frac{x-\mu}{\beta}} e^{-e^{\frac{x-\mu}{\beta}}} \quad (2)$$

Where μ is the location parameter and β is the scale parameter. The case where $\mu = 0$ and $\beta = 1$ is called the standard Gumbel distribution [11]. The equation for the standard Gumbel distribution (minimum) reduces to

$$f(x) = e^{-x} e^{-e^{-x}} \quad (3)$$

2.3 Exponential Distribution

The exponential distribution is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a

constant average rate. It is the continuous analogue of the geometric distribution, and it has the key property of being memory less. In addition to being used for the analysis of Poisson processes, it is found in various other contexts [12].

The probability density function (pdf) of an exponential distribution is

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x < 0 \end{cases} \quad (4)$$

Alternatively, this can be defined using the Heaviside step function [13], $H(x)$.

$$f(x; \lambda) = \lambda e^{-\lambda x} H(x) \quad (5)$$

Here $\lambda > 0$ is the parameter of the distribution, often called the rate parameter. The distribution is supported on the interval $[0, \infty]$. If a random variable X has this distribution, we write $X \sim \text{Exp}(\lambda)$. The exponential distribution exhibits infinite divisibility [14].

Mean and standard deviation are the essentials to apply probability distribution to the images. The mean is the average of the numbers and mean is used to calculate the central value of a set of numbers [15]. The mean is denoted by the symbol ' \bar{X} '. The formula for mean is,

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} \quad (6)$$

Standard deviation is the measure of how the numbers are spread out. The standard deviation is denoted by the symbol ' σ '. The formula for standard deviation is,

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad (7)$$

Probability distribution function are applied to several set of samples and get a range of values through which we can predict the recursive order of the given photocopy documents

3. METHODOLOGY

Recursive order number is the order number of photocopy which is obtained recursively. Texture degradation is one of the noticeable feature of the recursive photocopy document. Degradation of text increases as the recursive order number of the photocopy document increases. From the Fig 1b, it is evident that degradation in the recursive photocopy is maximum on right side of the document. The work focus on measuring texture degradation only on the right edge part of the recursive photocopy. In order to estimate the recursive order of the photocopy document, different distribution functions are applied to the given photocopy document and based on the values obtained by these functions, the order of the photocopy document is estimated. In the proposed work three different distribution functions Exponential, Normal and Extreme value distributions are computed. These distribution functions are applied to the training samples, range of values for every distribution function are generated and tabulated as shown in table 1.

Table 1: Range of distribution functions

Distribution Functions	Order Number					
	1 st	Overlapping between 1 st & 2 nd	2 nd	3 rd	4 th	5 th
Exponential	0.0211	0.0258	0.0258	0.0208	0.0189	0.0165
	-	-	-	-	-	-
Extreme value	0.0257	0.0308	0.0308	0.0221	0.0207	0.0188
	-	-	-	-	-	-
Normal	0.3401	0.3623	0.3995	0.2735	0.2009	0.0845
	-	0.3994	-	-	-	-
	0.3622	-	0.3400	0.3400	0.2734	0.2008
	-	-	-	-	-	-
	0.2107	0.2385	0.2567	0.1126	0.0801	0.0230
	-	0.2566	-	-	-	-
	0.2384	-	0.2886	0.2106	0.1125	0.0802
	-	-	-	-	-	-

From the table 1 it is noticeable that there are considerable overlapping values for order numbers one and two. This is because the texture degradation is quite narrow in these two recursive orders where as other order number classification has distinct range values. Hence, a decision system is required to make the classification in case of overlap to resolve the conflicts. Fig 2 shows the flow while making the decision.

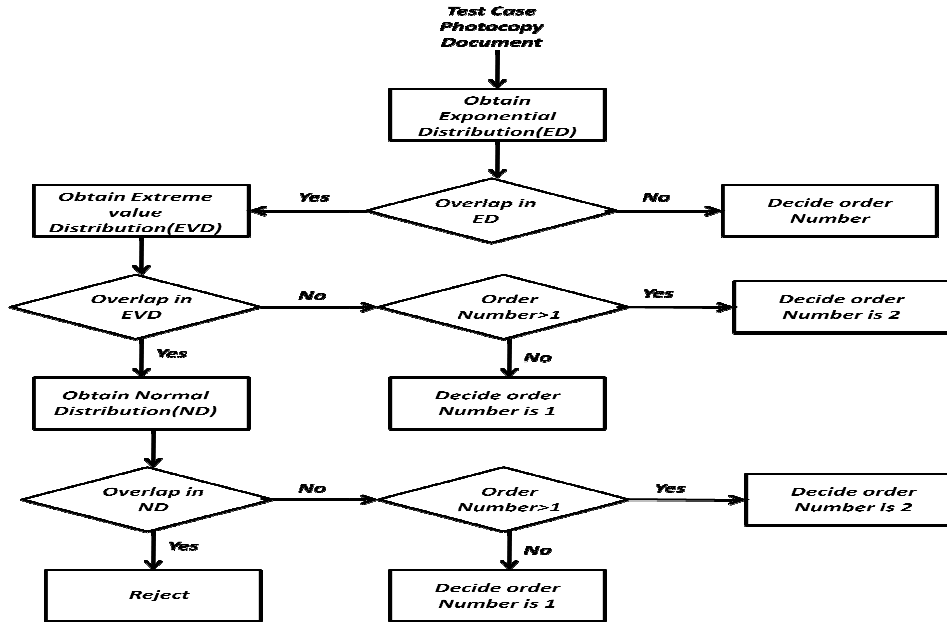


Fig2: Flow Diagram for Decision making in the proposed method

The photocopy document under consideration is subjected to obtain Exponential distribution.

The order number of the photocopy is decided if the Exponential distribution value falls within the classified range of values. In case the value is in overlap range then the Extreme value distribution is obtained. The order number is decided as 2 if the Extreme value distribution value

is in any of the range 2 or higher. The order number is decided as 1 if the value clearly falls in range of copy 1. In case of Extreme value distribution value also falls in overlap range, then Normal distribution is computed. The order number is decided as in Extreme value distribution and considered as rejection case if the value falls in the overlap range.

4. EXPERIMENTAL RESULTS

Experimentation is performed through testing the proposed method using synthetically generated samples of photocopy documents from different photocopying machines. Testing is carried out with sufficient number of test samples up to 5th order since higher order copies are not suitable for fabrication. The test samples include different sizes, different contents, figures, tables etc. The results of the testing are tabulated in table 2. The experiments conducted using test samples show an average classification efficiency of 83.5% with the misclassification of 11.5% and rejection is 5%. The misclassifications are mainly due to the presence of noise and dirt in the document, toner quality and machine's quality used in production of photocopies.

Table 2: Results of testing

Order Number	No. of samples	Classification			Efficiency
		Correct	Incorrect	Rejection	
1 st	45	40	03	02	88.8%
2 nd	54	42	09	03	77.7%
3 rd	65	50	10	05	76.9%
4 th	42	34	06	02	80.9%
5 th	55	52	02	01	94.5%
Total	261	217	30(11.5%)	13(5%)	83.5%

5. CONCLUSION

The implemented method provides a supervised system for estimating the recursive order of photocopy submitted. The method is essentially based on the probability distribution methods. The method shows average classification efficiency close to 84%. The misclassification is due to photocopies obtained from different machines and their quality. The proposed work is design of an efficient method to estimate the recursive order of photocopy and is a base work to continue research for better efficiency through investigating methods to find rate of degradation using variations in character thickness and line orientations. The rejection case indicates the proposed model cannot resolve the conflict and suggested for physical verification. The proposed work is prerequisite to continue the research to detect the photocopy under consideration is a fabricated or not.

REFERENCES

- [1] Rich Kevin Knight, Artificial Intelligence, 2nd Edition, McGraw-Hill Higher Education.
- [2] Madasu Hanmandlu, Mohd. Hafizuddin Mohd. Yusof, Vamsi Krishna Madasu off-line signature verification and forgery detection using fuzzy modeling Pattern Recognition Vol. 38, pp 341-356, 2005
- [3] Cha, S.-H., & Tapert, C. C., Automatic Detection of Handwriting forgery, Proc. 8thInt.Workshop Frontiers Handwriting Recognition(IWFHR-8), Niagara, Canada, pp 264-267, 2002
- [4] Christoph H Lampert, Lin Mei, Thomas M Breuel Printing Technique Classification for Document Counterfeit Detection Computational Intelligence and Security, International Conference, Vol. 1, pp 639-644, 2006
- [5] Utpal Garian, Biswajith Halder, On Automatic Authenticity Verification of Printed Security Documents, IEEE Computer Society Sixth Indian Conference on Computer vision, Graphics & Image Processing, pp 706-713, 2008
- [6] Angelo Frosini, Marco Gori, Paolo Priami, A Neural Network-Based Model For paper Currency Recognition and Verification IEEE Transactions on Neural Networks, Vol. 7, No. 6, Nov 1996
- [7] Suman Patgar, Vasudev T, 2012, Estimation of order number from successively photocopied document using Geometric moments, SACAIM 2012.
- [8] Suman. V. Patgar, Vasudev. T, An unsupervised intelligent system to detect fabrication in photocopy document using Geometric Moments and Gray Level Co-Occurrence Matrix, 2013.
- [9] Weisstein, Eric W. "Normal Distribution." From MathWorld--A Wolfram Web Resource.
- [10] Dr. B. S. Grewall, "Higher Engineering Mathematics" 39th edition.
- [11] www.mathworks.in/help/stats/extreme-value-distribution.html.
- [12] <http://mathworld.wolfram.com/ExponentialDistribution.html>.
- [13] Weisstein, Eric W., "Heaviside Step Function", MathWorld.
- [14] B. V. Ramana, "Higher Engineering Mathematics" McGraw-Hill.
- [15] Erwin Kreyszing, "Advanced Engineering Mathematics" 8th edition.

AUTHORS

Vasudev T is Professor, in the Department of Computer Applications, Maharaja Institute of Technology, Mysore. He obtained his Bachelor of Science and post graduate diploma in computer programming with two Masters Degrees one in Computer Applications and other one is Computer science and Technology. He was awarded Ph.D. in Computer Science from University of Mysore. He is having 30 years of experience in academics and his area of research is Digital Image Processing specifically document image processing.



Suman V Patgar, is Research Scholar, P.E.T Research Center Mandya. She obtained her Bachelor of Engineering from Kuvempu University in 1998. She received her Masters degree in Computer Science and Engineering from VTU Belgaum in 2004. She is pursuing doctoral degree with the supervision of Vasudev T under University of Mysore.



Rani K, obtained her Bachelor of Engineering in Computer Science from VTU Belgaum in 2012. She is pursuing Master degree in Computer Science and Engineering under VTU Belgaum.



COST-EFFECTIVE STEREO VISION SYSTEM FOR MOBILE ROBOT NAVIGATION AND 3D MAP RECONSTRUCTION

Arjun B Krishnan and Jayaram Kollipara

Electronics and Communication Dept.,
Amrita Vishwa Vidyapeetham, Kerala, India

abkrishna39@gmail.com
kollipara.jayaram@gmail.com

ABSTRACT

The key component of a mobile robot system is the ability to localize itself accurately in an unknown environment and simultaneously build the map of the environment. Majority of the existing navigation systems are based on laser range finders, sonar sensors or artificial landmarks. Navigation systems using stereo vision are rapidly developing technique in the field of autonomous mobile robots. But they are less advisable in replacing the conventional approaches to build small scale autonomous robot because of their high implementation cost. This paper describes an experimental approach to build a cost- effective stereo vision system for autonomous mobile robots that avoid obstacles and navigate through indoor environments. The mechanical as well as the programming aspects of stereo vision system are documented in this paper. Stereo vision system adjunctively with ultrasound sensors was implemented on the mobile robot, which successfully navigated through different types of cluttered environments with static and dynamic obstacles. The robot was able to create two dimensional topological maps of unknown environments using the sensor data and three dimensional model of the same using stereo vision system.

KEYWORDS

Arduino, Disparity maps, Point clouds, Stereo vision, Triangulation

1. INTRODUCTION

The future will see the deployment of robots in the areas of indoor automation, transportation and unknown environment exploration. Implementation of Robotic systems in such tasks is widely appreciated technique as they handle these tasks more efficiently and reliably. Currently, a growing community of researchers are focusing on the scientific and engineering challenges of these kinds of robotic systems.

This project tries to address the main challenges in the field of autonomous robots – Autonomous Navigation. There are several techniques for effective autonomous navigation, among which Vision based navigation is the most significant and popular technique which experiences rapid

developments. Other techniques include navigation using ultrasound sensors, LIDAR (Light Detection and Ranging) systems, preloaded maps, landmarks etc. Navigation which uses ultrasound sensors will not detect narrow obstacles such as legs of tables and chairs properly, and hence leads to collision. LIDAR systems are perfect tools for Indoor navigation because of their accuracy and speed but they are less impressive for large scale implementation due to their high cost [1]. Navigation based on landmarks and preloaded maps become valid options only when there is prior information about the environment and thus, it does not give a generic solution to the problem of autonomous navigation. Vision can detect objects just as in the case of human vision and it gives the sense of intelligence to the robots. Out of all vision based techniques, stereo vision is the most adoptable technique because of its ability to give the three dimensional information about how the environment looks like and decide how obstacles can be avoided to safely navigate through that environment. Commercially available stereo cameras are expensive and require special drivers and software to interface with processing platforms which again adds up the cost of implementation. In this scenario, building a cost-effective stereo vision system using regular webcams which is able to meet the performance of commercially available alternatives is highly appreciable and this fact makes the theme of this project.

2. RELATED WORKS

Several autonomous mobile robots equipped with stereo vision, were realized in the past few years and deployed both industrially and domestically. They serve humans in various tasks such as tour guidance, food serving, transportation of materials during manufacturing processes, hospital automation and military surveillance. The robots Rhino [2] and Minerva [3], developed by the researchers from Carnegie Mellon University (USA) and University of Bonn (Germany) are famous examples of fully operational tour guide robots used in museums. These robots use stereo vision along with sonar sensors for navigate and building the map. The robot Jose [4], developed in University of British Columbia (Canada), uses Trinocular stereo vision - which is a combination of vertical and horizontal binocular stereo vision - to accurately map the environment in all three dimensions. PR2 robot designed by Willow Garage research laboratory is one of the most developed home automation robot [5]. This uses a combination of stereo vision and laser range finders for navigation and grasping of objects.

According to [6] there are two essential algorithms for every stereo vision systems: Stereo Calibration algorithm and Stereo Correspondence algorithm. Calibration algorithm is used to extract the parameters of the image sensors and stereo rig, hence has to be executed at least once before using the system for depth calculation. Stereo correspondence algorithm gives the range information by using method of triangulation on matched features. A stereo correspondence algorithm based on global matching is described in [7] and [8], [9], [10] are using correspondence search based on block matching. Considering these techniques as a background, an algorithm is designed for this project, which uses horizontal stereo vision system by block matching for obtaining stereo correspondence. Low cost ultrasound sensors and infrared sensors are chosen for overlapping with visual information.

3. ROBOT PLATFORM

Our experimental platform is a six wheeled differential drive rover that can carry a portable personal computer. Two wheels are free rotating wheels with optical encoders attached for keeping track of the distance travelled. Other four wheels are powered by high torque geared motors of 45 RPM each, which gives the robot a velocity of 20cm/sec. Three HC-SR04 ultrasound sensors are attached in the front for searching obstacles in 4m range. Two Infrared range finders are employed to monitor the vertical depth information of the surface on which

robot operates and hence avoid falling from elevated surfaces. A digital compass – HMC5883L – is used to find the direction of robot's movement. The core elements of the embedded system of this robot are two 8 bit ATmega328 Microcontroller based Arduino boards. One Arduino collects information from optical wheel encoders based on interrupt based counting technique, whereas the second Arduino collects data from all other sensors used in the platform and also controls the motion of motors through a motor driver. Heading from compass and distance data from wheel encoders gives reliable odometric feedback to the control system. A PID algorithm has been developed and implemented in order to keep the robot in straight line motion in an obstacle free region. Both the Arduino boards continuously transfer the collected data from the sensors to the on-board PC for storage and receive decisions from vision system implemented in on-board PC. Figure 1 shows the overall architecture of the mobile robot used in this project.

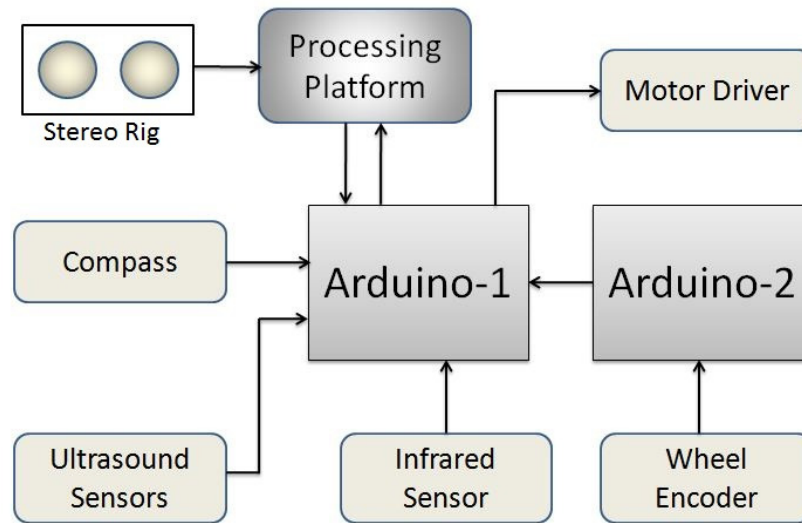


Figure 1. The architecture of the mobile robot

4. STEREO VISION SYSTEM

Stereo vision is a technique for extracting the 3D position of objects from two or more simultaneous views of a scene. Stereo vision systems are extensively used in object classification and object grasping applications because of its ability to understand the three dimensional structure of objects. Mobile robots can use a stereo vision system as a reliable and effective primary sensor to extract range information from the environment.

In ideal case, the two image sensors used in a stereo vision system has to be perfectly aligned along a horizontal or vertical straight line passes through the principle points of both images. Cameras are prone to lens distortions, which are responsible for introducing convexity or concavity to the image projections. The process called Stereo-pair rectification is adopted to remap distorted projection to undistorted plane. The obtained rectified images from both the sensors are passed to an algorithm which searches for the matches in the images along each pixel line. The difference in relative positions of an identified feature is called as the disparity associated with that feature. Disparity map of a scene is used to understand the depth of objects in the scene with respect to the position of the image sensors through the process called Triangulation. Figure 2.a shows the arrangement of image planes and Figure 2.b describes the Pinhole model [11] of two cameras to illustrate the projection of a real world object is formed in left and right image planes. The formation of disparity is shown in Figure 2.c

Accuracy of depth perception from a robust stereo vision system is sufficient for segmenting out objects based on their depth, in order to avoid collisions during navigation in real time. The following sections describe the details of hardware and software implementations of stereo vision system in this project.

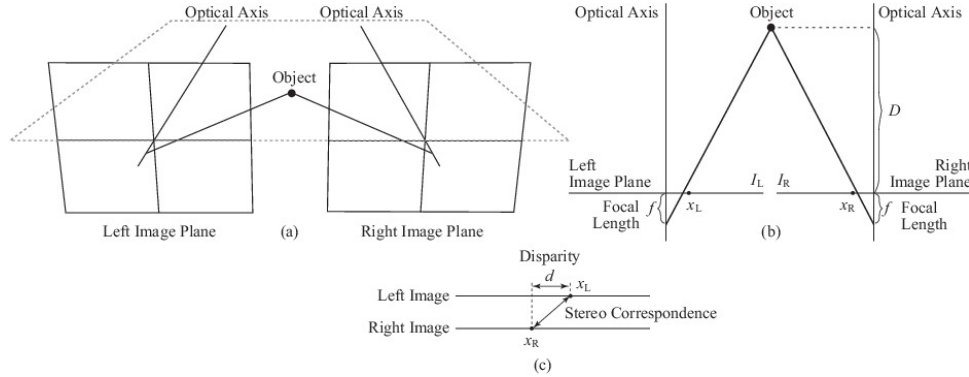


Figure 2. Modelling of stereo rig and disparity formation using pinhole model of cameras

4.1. Building the Stereo Rig

A stereo camera is a type of camera with two or more lenses with separate image sensors for each lens. This allows the camera to simulate human binocular vision, therefore giving it the ability to capture three-dimensional images, a process known as stereo imaging. In this project two CMOS VGA web Cameras (6 USD per camera) of resolution 640×480 with USB2.0 high speed (UVC) interface are used to build the Stereo Rig.

The distance of separation between two cameras which is also known as baseline length, is a crucial parameter of a stereo vision system, which decides the range of reliable depth perception. A longer baseline length increases both the minimum and maximum bounds of this range whereas a shorter baseline length decreases the bounds [12]. Hence the choice of baseline length of a stereo rig is mostly application dependent and limits the usable information available from the rig. Since a mobile robot in an indoor environment is similar to a human navigating in indoor, the most adoptable option for baseline length is the distance of separation between the eyes. A detailed study on human binocular vision system was conducted and the results were recorded. The typical interpupillary distance of humans varies between 50-75mm. The mean interpupillary distance for a human is found out to be 63.2mm [13] and hence the distance of 63mm is selected for the stereo rig used in this project. The mechanical setup was designed using CAD tool and the design is manufactured on acrylic sheet using CNC machine. The cameras were fixed on the rig precisely by monitoring the collinearity of the left and right images obtained. The manufactured stereo rig is shown in Figure 3.



Figure 3. Stereo camera rig made from two webcams.

4.2. The software for stereo vision system

Software required for this project has been developed in C++ language using Microsoft Visual C++ IDE. OpenCV, which is a popular open source computer vision library, is used to implement image processing algorithms. The stereo vision system modelled with Pinhole model is described with the help of two entities, Essential matrix E and Fundamental matrix F . The matrix E includes information about relative translation and rotation between two cameras in physical space whereas matrix F contains additional information related to the intrinsic parameters of a both cameras. Hence Essential matrix relates two cameras with their orientation and Fundamental matrix relates them in pixel coordinates.

The stereo camera will provide simultaneously taken left and right image pairs as an input to the processing unit. The initial task for a stereo vision system implementation is to find above mentioned fundamental and essential matrices. OpenCV provides predefined functions to find these matrices using RANSAC algorithm [14] and hence calibrate cameras and the rig. Calibration requires a calibration object which is regular in shape and with easily detectable features. In this project, the stereo camera calibration is performed using a regular chessboard as it gives high contrast images which contain easily detectable sharp corners which are separated with equal distances. Several left and right image pairs were taken at different orientations of the chessboard and the corners were detected as shown in Figure 4.



Figure 4. Stereo camera and rig calibration using chessboard as a calibrating object. Detected chessboard corners are marked in simultaneously taken left and right images.

The calibration algorithm computes intrinsic parameters of both the cameras and extrinsic parameters of the stereo rig and stores the fundamental and essential matrixes in a file. This information is used to align image pairs perfectly along the same plane by a process called Stereo Rectification. Rectification enhances both reliability and computational efficiency in depth perception. This is a prime step in the routine if the cameras are misaligned or with an infirm mechanical setup. The custom made stereo setup used in this project showed a negligible misalignment which suggested no requirement of rectification of image pairs for reliable results needed for safe indoor navigation. An example of rectified image pair obtained from the algorithm is shown in Figure 5.

The image pair is passed through a block-matching stereo algorithm which works by using small Sum of Absolute Difference (SAD) windows to find matching blocks between the left and right images. This algorithm detects only strongly matching features between two images. Hence the algorithm produces better results for scenes with high texture content and often fails to find correspondence in low textured scenes such as an image of a plane wall. The stereo correspondence algorithm contains three main steps: Pre-filtering of images to normalize their brightness levels and to enhance the texture content, Correspondence search using sliding SAD window of user defined size along horizontal epipolar lines, and post-filtering of detected matches to eliminate bad correspondences.

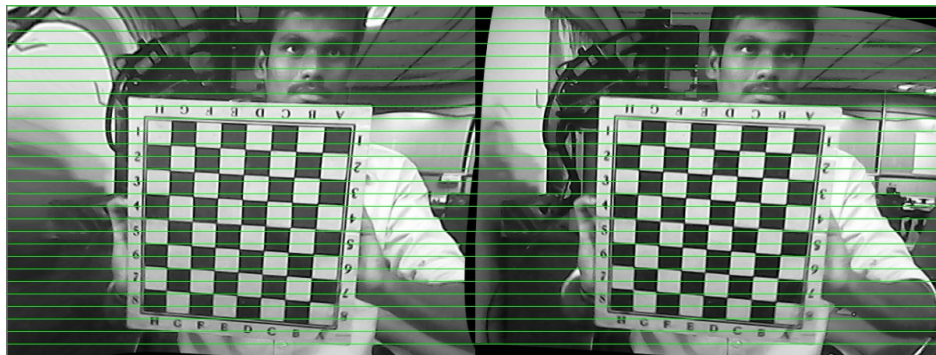


Figure 5. The rectified image pairs

The speed of the algorithm depends on the size of SAD window and the post-filtering threshold used in the algorithm. Larger SAD windows produce poorer results but elapses less time and vice versa. The choice of window size exhibits a trade-off between quality of the results and algorithm execution time, which leads to the conclusion that this parameter is completely application specific. The window size of 9x9 was selected empirically for the algorithms used in this project. Other parameters associated with the correspondence search algorithm are minimum and maximum disparities of searching. These two values establish the Horopter, the 3D volume that is covered by the search of the stereo algorithm. If these values are fixed, the algorithm limits the search for a match in the range between these two values which indirectly confines the real world depth perception between two well defined distances. The formation of horopter is shown in Figure 6 [15]. Each horizontal line in the Figure 6 represents a plane of constant disparity in integer pixels 20 to 12. A disparity search range of five pixels will cover different horopter ranges, as shown by vertical arrows. Considering the velocity of the robot the disparity limits are chosen such that a horopter is formed from 40cm to 120 cm from the frontal plane of the camera.

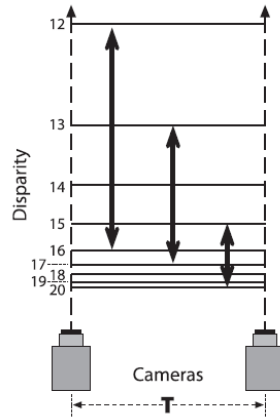


Figure 6. The formation of horopter for different disparity limits

The stereo correspondence algorithm generates a greyscale image in which intensity of a pixel is proportional to disparity associated with corresponding pixel location. The obtained disparity values in the image are mapped to real world distances according to the triangulation equation 1.

$$Z = \frac{f \times T}{d} \quad (1)$$

Where f is the known focal length, T is the distance of separation between cameras, d is the disparity obtained.

Figure 7 shows the disparity map of a testing image taken during the camera calibration process using chessboard. The low intensity (dark) portions are distant objects whereas high intensity (light) portions are objects which are closer to the camera.

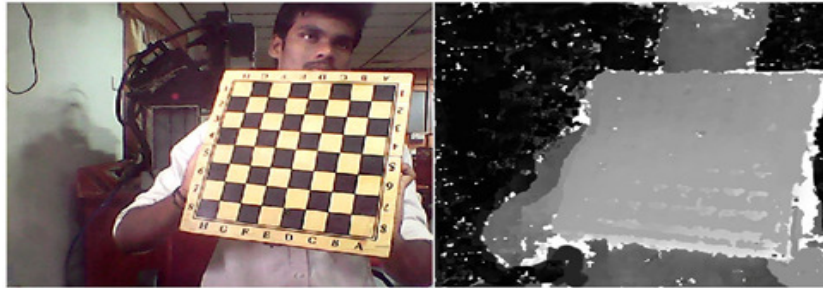


Figure 7: Image from the left camera (left), Computed disparity map (right)

The disparity results are obtained as expected only in particular range of distances from the Stereo Rig because of the nonlinear relationship between disparity and distance [15].

4.3. Depth based Image segmentation and obstacle avoidance

The disparity maps generated by above mentioned algorithm plays a vital role in obstacle avoidance during navigation. The segmentation based on the intensity levels is same as segmentation based on depth. A segmentation algorithm is used to detect near objects which isolates regions which are having high intensity range and searches for connected areas that can form blobs within the segmented regions. The contours of these blobs are detected and bounding

box coordinates for each blob are calculated. The centres of the bounding boxes as well as the bounding boxes are marked on the image. The input image from left camera is divided into two halves to classify the position of the detected object to left or right. The centre of the contour is tracked and if it is found out to be in the left half of the image, algorithm takes a decision to turn the robot to the right side and vice versa. If no obstacles are found in the search region robot will continue in its motion along the forward path. In case of multiple object occurrences in both halves, robot is instructed to take a 90 degree turn and continue the operation. Figure 8 shows the disparity map of several obstacle conditions and the corresponding decisions taken by the processing unit in each case.

Instruction from processing unit is communicated with robot's embedded system through UART communication. Instruction to move forward will evoke the PID algorithm implemented and robot follows exact straight line path unless the presence of an obstacle is detected by the vision system. Our algorithm elapses 200 ms for a single decision making. Dynamic obstacles such as moving humans may not be properly detected by the stereo vision. But this issue is handled by giving high priority for ultrasound sensors and the robot is able to stop instantly.

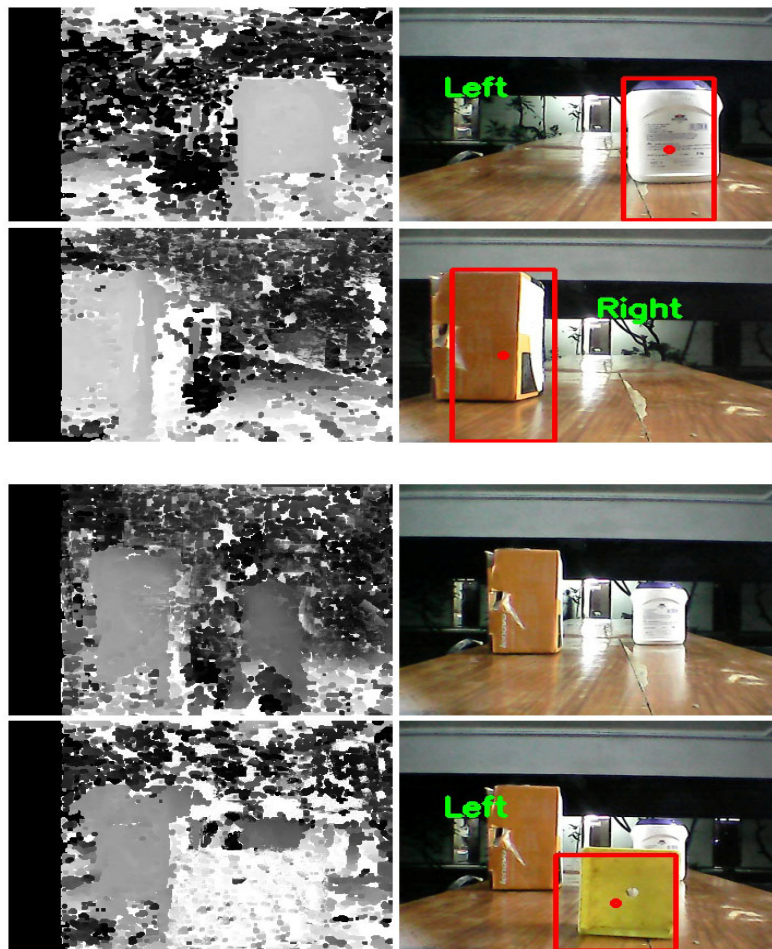


Figure 8. Disparity map of several obstacle conditions in an indoor environment (left). Detected obstacles in the specified distance range and corresponding decisions taken by the processing unit are shown (right)

4.4 3D Reconstruction

Three Dimensional reconstruction is the process of generating the real world model of the scene observed by multiple views. Generated disparity maps from each scene can be converted into corresponding point clouds with real world X, Y and Z coordinates. The process of reconstruction of 3D points requires certain parameters obtained from the calibration of Stereo rig. An entity called Re-projection matrix is formed from the intrinsic and extrinsic parameters and it denotes the relation between real-world coordinates and pixel coordinates. The entries of re-projection matrix are shown in Figure 9.

$$Q = \begin{bmatrix} 1 & 0 & 0 & -c_x \\ 0 & 1 & 0 & -c_y \\ 0 & 0 & 0 & f \\ 0 & 0 & -1/T_x & (c_x - c'_x)/T_x \end{bmatrix}$$

Figure 9. Re-projection matrix of a Stereo Rig

(c_x, c_y) – is the principal point of the camera. The point at which, the image plane coincides with the middle point of the lens.

f – Focal length of the camera, as the cameras in the stereo rig are set to same focal length thus the Re-projection matrix has a single focal length parameter.

T_x – Translation coefficient in x –direction.

The Re-projection matrix thus generated converts a disparity map into a 3D point cloud by using the matrix computation shown in equation 2.

$$Q \begin{bmatrix} x \\ y \\ d \\ 1 \end{bmatrix} = \begin{bmatrix} X \\ Y \\ Z \\ W \end{bmatrix} \quad (2)$$

Where x and y are the coordinates of a pixel in the left image, d is the corresponding disparity associated with that pixel and Q is the re-projection matrix. The real world coordinates can be computed by dividing X, Y and Z by W present in the output matrix.

The calculated 3D point clouds along with corresponding RGB pixel values are stored in the memory in a text file along with the odometric references at each instance of point cloud generation. After the successful completion of an exploration run in an unknown environment, the stored point cloud is retrieved and filtered using Point Cloud Library integrated with C++. Point clouds groups which are having a cluster size above a particular threshold level only is used for 3D reconstruction and thus inherently removes the noisy point clusters. Since error of projection increases with increasing real world distance, point clouds which lie beyond a threshold distance is also removed. 3D reconstructions of each scene are generated and stored according to the alignment of robot at that corresponding time. The visualised 3D reconstruction examples are shown in Figure 10. The overlapped re-projection of continuous scenes can be done to obtain the

complete 3D mapping of the environment. This 3D map can be used as a powerful tool in further navigations in the same environment. It can also be used to plan the path if a destination point in the environment is given to the robot.

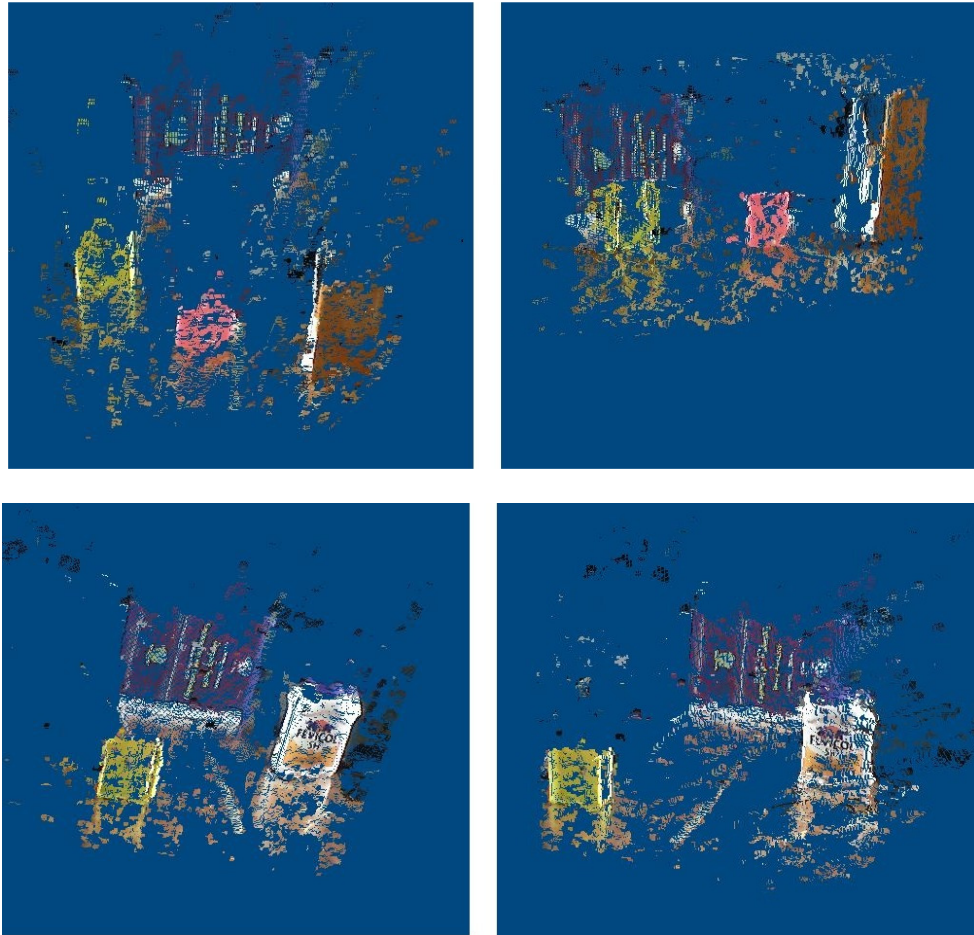


Figure 10. 3D Reconstructions of filtered Point clouds

5. RESULTS

Stereo vision based SLAM architecture is one of the least pondered but rapidly developing research area which has been dealt in this project and we have successfully implemented a cost effective prototype of the stereo camera and robotic platform. The Stereo Vision System produces comparable outputs with that of commercially available alternatives. The total stereo matching program is able to process five frames per second in a 1.6Ghz Intel Atom processor board equipped with 2GB RAM. This performance is adequate for safe indoor navigation for slowly moving robots. The overlapping of vision perception with other information from sensors ensures a nearly error-proof navigation for robot in indoor environments. Accurate 2D mapping of the environment based on the ultrasound data is implemented along with the 3D mapping using the stereo vision. 3D reconstruction elapses 25 to 80ms per frame whereas 2D mapping requires less than 50ms for a sample data collected from a test run timed four minutes. Vision can detect objects just as in the case of human vision and gives the sense of intelligence to the robot. The choice of mechanical parameters of stereo rig, range of the horopter, stereo correspondence

algorithm parameters and filter parameters used for reconstruction were proved to be sufficient for the successful accomplishment of tasks identified during project proposal. The images of robot navigating in the indoor environment are shown in Figure 11.



Figure 11. Robot operates in cluttered indoor environment

6. CONCLUSION AND FUTURE WORK

This paper outlines the implementation of a cost-effective stereo vision system for a slowly moving robot in an indoor environment. The detailed descriptions of algorithms used for stereo vision, obstacle avoidance, navigation and three dimensional map reconstruction are included in this paper. The robot described in this paper is able to navigate through a completely unknown environment without any manual control. The robot can be deployed to explore an unknown environment such as collapsed buildings and inaccessible environments for soldiers during war. Vision based navigation allows robot to actively interact with the environment. Even though vision based navigation systems are having certain drawbacks when compared with other techniques. Stereo vision fails when it is being subjected to surfaces with less textures and features, such as single colour walls and glass surfaces. The illumination level of environment is another factor which considerably affects the performance of stereo vision. The choice of processing platform is crucial in the case of processor intense algorithms used in disparity map generation. Point clouds generated are huge amount of data which has to be properly handled and saved for better performances.

The future works related to this project are developing of a stereo camera which has reliable disparity range over longer distance, implementing the stereo vision algorithm in a dedicated processor board and further development of the robot for outdoor navigation with the aid of Global Positioning System.

REFERENCES

- [1] Borenstein, J., Everett, B., and Feng, (1996) Navigating Mobile Robots: Systems and Techniques, A.K. Peters, Ltd.: Wellesley, MA.
- [2] J. Buhmann, W. Burgard, A.B. Cremers, D. Fox, T. Hofmann, F. Schneider, J. Strikos, and S. Thrun, (1995) "The mobile robot Rhino," AI Magazine, Vol. 16, No. 1.
- [3] S. Thrun, M. Bennewitz, W. Burgard, A.B. Cremers, F. Dellaert, D. Fox, D. Hähnel, C. Rosenberg, N. Roy, J. Schulte and D. Schulz, (1999) "MINERVA: A second generation mobile tour-guide robot," in Proc. IEEE International Conference on Robotics and Automation (ICRA), vol.3, No., pp.1999.
- [4] Don Murray, and Jim Little, (2000) "Using real-time stereo vision for mobile robot navigation," Autonomous Robots, Vol. 8, No. 2, pp.161-171.
- [5] Pitzer, B., Osentoski, S., Jay, G., Crick, C., and Jenkins, O.C., (2012) "PR2 Remote Lab: An environment for remote development and experimentation," Robotics and Automation (ICRA), vol., no., pp.3200 – 3205.

- [6] Kumar S., (2009) "Binocular Stereo Vision Based Obstacle Avoidance Algorithm for Autonomous Mobile Robots," Advance Computing Conference, IACC 2009. IEEE International, vol., no., pp.254-259.
- [7] H. Tao, H. Sawhney, and R. Kumar. (2001) "A global matching framework for stereo computation," In Proc. International Conference on Computer Vision, Vol. 1.
- [8] Iocchi, Luca, and Kurt Konolige. (1998) "A multiresolution stereo vision system for mobile robots," AIIA (Italian AI Association) Workshop, Padova, Italy.
- [9] Schreer, O., (1998) "Stereo vision-based navigation in unknown indoor environment," In Proc. 5th European Conference on Computer Vision, Vol. 1, pp. 203-217.
- [10] Yoon, K.J., and Kweon, I.S, (2006) "Adaptive support-weight approach for correspondence search," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, No. 4, pp.650-656.
- [11] Z. Zhang, G. Medioni and S.B. Kang, (2004) "Camera Calibration", Emerging Topics in Computer Vision, Prentice Hall Professional Technical Reference, Ch. 2, pp.443.
- [12] M. O kutomi and T . K anade, (1993) "A multiple-baseline stereo," IEEE Transactions on Pattern Analysys and Machine Intelligence, Vol. 15, No. 4, pp.353-363.
- [13] Dodgson, N. A, (2004) "Variation and extrema of human interpupillary distance," In A. J. Woods, J. O. Merritt, S. A. Benton and M. T. Bolas (eds.), Proceedings of SPIE: Stereoscopic Displays and Virtual Reality Systems XI, Vol. 5291, pp.36-46.
- [14] M.A. Fischler and R.C. Bolles, (1981) "Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography". Communication of ACM, Vol. 24, No. 6, pp.381-95.
- [15] G. Bradski and A. Kaehler, (2008) "Learning OpenCV: Computer Vision with the OpenCV Library," O'Reilly Media, Inc.
- [16] Martin Humenberger, Christian Zinner, Michael Weber, Wilfried Kubinger, and Markus Vincze, (2010) "A fast stereo matching algorithm suitable for embedded real-time systems", Computer Vision and Image Understanding, Vol. 114, No. 11, pp.1180-1202.
- [17] Murray, D. and Jennings, C., "Stereo vision based mapping and navigation for mobile robots," in Proc. 1997 IEEE International Conference on Robotics and Automation, Vol. 2, pp.1694-1699.
- [18] Z. Zhang, and G. Xu, (1996) "Epipolar Geometry in Stereo, Motion and Object Recognition," Kluwer Academic Publisher, Netherlands.

AUTHORS

Arjun B Krishnan received Bachelor of Technology degree in Electronics and Communication Engineering from Amrita Vishwa Vidyapeetham, Kollam, India in 2014. Currently, he is working as a researcher in Mechatronics and Intelligent Systems Research Laboratory under Mechanical Dept. of Amrita Vishwa Vidyapeetham. His research interests include Autonomous mobile robotics, Computer vision and Machine learning.



Jayaram Kollipara received Bachelor of Technology degree in Electronics and Communication Engineering from Amrita Vishwa Vidyapeetham, Kollam, India in 2014. He joined as a Program Analyst in Cognizant Technology Solutions, India. His research interests are Image and Signal processing, Pattern recognition and Artificial intelligence.



PERFORMANCE COMPARISON OF ONLINE HANDWRITING RECOGNITION SYSTEM FOR ASSAMESE LANGUAGE BASED ON HMM AND SVM MODELLING

Deepjoy Das, Rituparna Devi, SRM Prasanna, Subhankar Ghosh,
Krishna Naik

Department of EEE, IIT Guwahati, Assam
{deepjoy2002@gmail.com, rituparna.sarma5@gmail.com,
prasanna@iitg.ernet.in, ghoshsubhankar@gmail.com,
krishnanaik.35@gmail.com}

ABSTRACT

This work emphasises on the development of Assamese online character recognition system using HMM and SVM and performs a recognition performance analysis for both models. Recognition models using HTK (HMM Toolkit) and LIBSVM (SVM Toolkit) are generated by training 181 different Assamese Stokes. Stroke and Akshara level testing are performed separately. In stroke level testing, the confusion patterns of the test strokes from HMM and SVM classifiers are compared. In Akshara level testing, a GUI (provided by CDAC-Pune) which is integrated with the binaries of HTK/LIBSVM and language rules (stores the set of valid strokes which makes a character) are used, manual testing is done with native writers to test the Akshara level performance for both models. Experimental results show that the SVM classifier outperforms the HMM classifier.

KEYWORDS

Support Vector Machines, Hidden Markov Models, Handwriting Recognition, Assamese, LIBSVM, HTK

1. INTRODUCTION

Of the various handwriting recognition systems available, there exists two basic handwriting recognition domains distinguished primarily by nature of the input signal-online and offline. In offline system the digitised information is in the static form whereas in the online system, information is acquired during production of the handwriting using equipments such as Tablet PC which captures the trajectory of the writing tool. The information captured undergoes some filtration, pre-processing and normalisation process after which the handwriting is segmented into basic units which are usually a character or part of a character. Finally each segment is classified and labelled. In our system, we examine the effectiveness of using Hidden Markov Models (HMM) and Support Vector Machines (SVM) for modelling the classifier. HMM has been used for Bangla [2], Telugu [3], Tamil [4], Malayalam [5] and in previous works for Assamese [6] [7]. Support vector machines (SVMs) have also been used in [8] for Telegu and Devnagiri scripts while [9] compares the performance between systems developed using HMM and SVM for Telegu script. The classifiers are built individually using HMM and SVM and the recognition accuracies of both the systems are analysed for comparison. In our work, the two coordinate trace

namely strokes, substrokes and suprastrokes. The strokes consist of a typical basic component of aksharas agreed naturally by majority of non-cursive writers. The substrokes contained components are formed by merging several components or strokes. Again if the infrequent writers break the components of the stroke into more than one component then splitted components form the suprastrokes. A list is prepared combining all the above strokes of different data groups and we have a final list of 203 distinct strokes in Assamese handwriting. The final strokes list is depicted below in Figure 2.

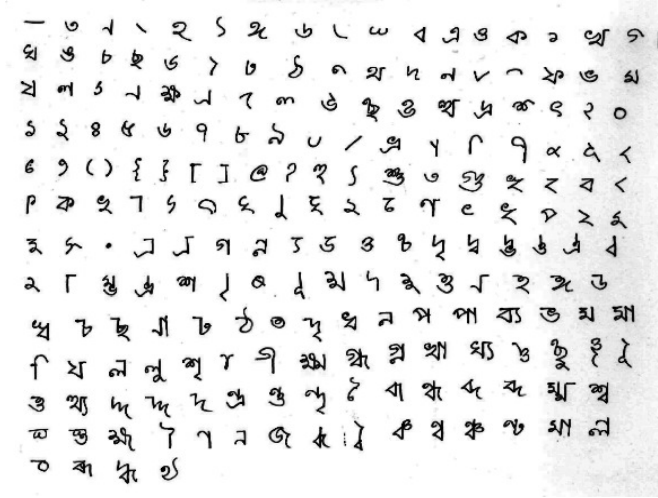


Figure 2 Isolated Assamese strokes

2. CHARACTER RECOGNITION SYSTEM

The schematic block diagram of Assamese Online Stroke Recognition system using HMM & SVM Modelling is shown below in Figure 4.

2.1 Database

The Assamese data set consists of a set of 203 isolated strokes or basic components as shown in Figure 2. A set of 147 Assamese Aksharas have been finalised to acquire the stroke data required to design a robust stroke recogniser. The Akshara examples have been collected from 100 users in two sessions in an HP Tablet PC using an open source tool developed by HP with a sampling rate of 120 Hz. From each of the Akshara sample, the basic component or stroke is extracted from one pen down to pen up which results in about 1000 examples for each stroke of which 50 % are used for training and 50 % for testing. During data collection no limitation or restriction is enforced on the style of writing and hence, we have a large variation among the samples of a given stroke.

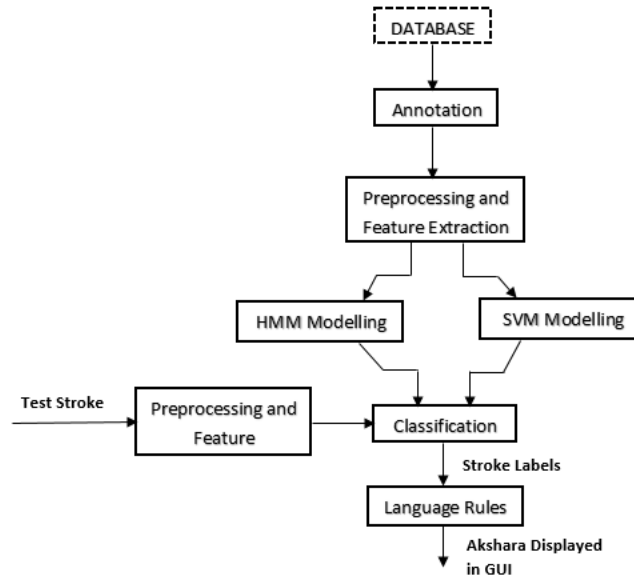


Figure 3: Assamese Online Stroke Recognition System Using HMM & SVM Modelling

2.2 Annotation

Annotation refers to the labelling of collected database in order to arrange them into groups of analogous patterns which in the milieu of our work are the strokes, substrokes and suprastrokes of Assamese language. The classifiers are then trained with these patterns which are then used for recognition purpose. The desired outcome after annotation is a database completely labelled at the sentence, word, character and stroke labels. The annotated data is analysed to finalise the set of strokes, substrokes and suprastrokes and finally only those patterns or strokes are retained which are used by more than 5 % users.

2.3 Pre-processing and Feature Extraction

The pre-processing stage consists of size normalization, smoothing, interpolation of missing points, removal of duplicate points and resampling of the captured coordinates [10].

2.3.1 Size Normalisation

The size of each individual data sample is normalised by scaling the pattern both horizontally and vertically [11]

2.3.2 Smoothing

Smoothing excludes the noise captured during the data collection process and performed using a moving average filter of size three. Each pattern is smoothed both in horizontal and vertical directions discretely [12]

2.3.3 Removal of duplicate points

Duplicate points do not contain any information and only cause data redundancy and hence these points are removed before feature extraction [12]

2.3.4 Resampling

Resampling eliminates the disparities in the data due to the writing speed of the writers. It is performed by linear interpolation of missing points which results in a sequence of equidistant points [12]

2.4 Feature Extraction

The pre-processed horizontal & vertical coordinates and their first and second derivatives can be used as features for the modelling of the stroke classifier. The first derivative gives the change and the second derivative gives the change of change in horizontal and vertical coordinates. The first derivative is calculated to observe the change in the trajectory at current point. The second derivative is calculated in order to examine the change of change in trajectory at current point. A window size of two is considered in both the cases. The method of extracting feature vectors is identical for both the classification models used in our work.

2.5 Classification Models

The efficiency of HMM and SVM are studied for developing the stroke classifiers.

2.5.1 HMM Modelling and testing

HMM models a doubly stochastic process, one observable and the other hidden [14]. In our work, the sequence of feature vectors from the online handwriting is the visible stochastic process and the underlying hand movement is the later. In the present work, for modelling each stroke, one left to right, continuous density HMM is developed. The left to right structure is used supposing distinctive directions of handwriting movements [15]. After collected database is annotated at stroke level, six dimensional features are extracted from the pre-processed coordinates. 203 strokes are finalised during script analysis of the Assamese language and hence 203 HMM models are built for each of the 203 strokes. All the test examples corresponding to each stroke class are tested against all the stroke models and if misclassification arises due to resemblance in pattern shape between two strokes those stroke classes are merged. Hence the final stroke classifier is developed with 181 stroke classes. The HMM models are trained using 7 states and 20 mixtures and HMM Toolkit (HTK) is used for training and testing.

A. HMM Training

The feature vectors described in the previous section are used for training the HMM which comprises of a set of states and the alterations linked with it and are trained using Baum-Welch re-estimation or expectation maximization (EM) approach [14]. In this procedure an initial model is taken and improved model parameters are re-estimated using the given set of feature vectors. The most recent model is the initial model for the next iteration and again re-estimation is done using the same set of feature vectors. This procedure is recurring until model parameters become static and the model of the last reiteration is stored as the model for the given class [12]. The process is repeated for all stroke classes.

B. HMM Testing

During testing the class information for the examples that are unknown to the trained model are found out. The likelihood probability of the given test example against each of the trained HMM models are determined and the model with the highest likelihood is theorised as the class. The process is repeated for all the testing examples and the class information is noted [12], [15].

2.5.2 SVM Modelling

Similar to HMMs, if given a set of training samples, SVMs will attempt to build a model. Each training data instance is marked as belonging to one of two categories. The SVM will attempt to separate the data instances into those two categories with a $p-1$ dimensional hyperplane, where p is the size of each data instance. This model can then be used on a new data instance to predict which category it would fall onto. The maximum margin hyperplane can be represented as [17]

$$y(x) = b + \sum \alpha_i y_i K(x(i), x)$$

Vector x is a test case and y_i is the class value of the training example $x(i)$. In the equation, the parameters of the hyper plane are b and α_i . b is a real constant, and α_i are non-negative real constants. The function $K(x(i), i)$ is a kernel function and SVMs are powerful in the sense that one can substitute different kernel functions. The four basic kernel functions are [18]

1. Linear: $K(x_i, x_j) = x_i^T x_j$
2. Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
3. Radial(RBF): $K(x_i, x_j) = e^{(-\gamma \|x_i - x_j\|^2)}, \gamma > 0$
4. Sigmoid: $K(x_i, x_j) = (\gamma x_i^T x_j + r)$

The classifier can be constructed as follows: [19]

$$\begin{aligned} w^T \phi(x(i)) + b &\geq 1, \text{ if } y_i = 1 \\ w^T \phi(x(i)) + b &\leq -1, \text{ if } y_i = -1 \\ y_i [w^T \phi(x(i)) + b] &\geq 1, i = 1, 2, \dots, N \end{aligned}$$

Where $\phi(-)$ is a nonlinear function that maps the given inputs into some higher dimensional space. In case we cannot find the separating hyperplane in this space, we introduce additional variables: ξ_i , Where $i = 1, \dots, N$. After this, we will attempt to solve this minimization problem:

$$\min_{w, b, \xi_i} J(w, \xi_i) = \frac{1}{2} w^T w + c \sum_{i=1}^N \xi_i \quad \text{s.t.} \quad y_i [w^T \phi(x(i)) + b] \geq 1 - \xi_i \quad \xi_i \geq 0, i = 1, 2, \dots, N$$

The solution to the above model will be the optimal separating hyperplane.

A. SVM Training and Testing

We use the LIBSVM defaults “radial basis” kernel for mapping a given set of input vectors into a higher dimensional space. The pre-processing step involves extraction of four dimensional feature vectors namely the horizontal & vertical coordinates and their first derivatives. The second derivatives are not used as they have been found to give reduced recognition accuracy. As SVM works with fixed sized vectors, we choose 60 equidistance handwritten points, which span the whole handwritten curvature (choose more points in high curvatures).

3. GRAPHICAL USER INTERFACE (GUI)

The GUI of the testing tool has been developed by Centre for Development of Advanced Computing Graphics and Intelligence based Script Technology Group (CDAC), Pune, India and is provided with API Calls. A Particular API call was used to get certain service out of the GUI. We have integrated the GUI with our HMM and SVM training models separately one at a time, along with valid set of language rules (stored in text file) using a DLL.

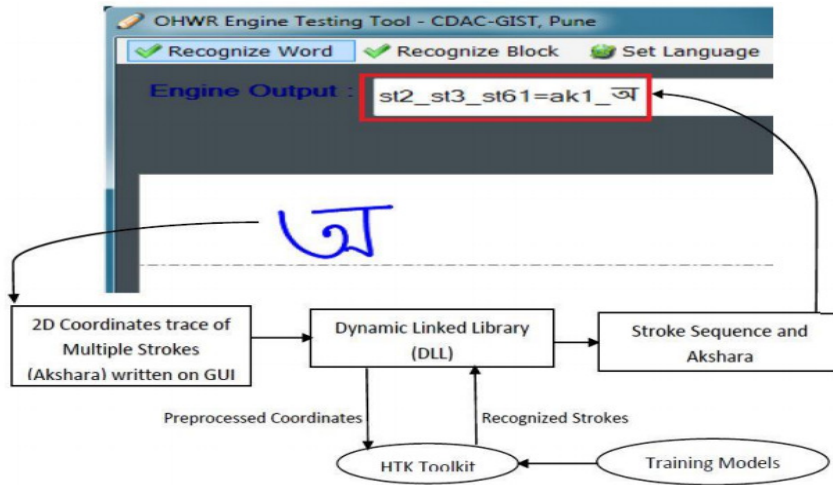


Figure 4: Block Diagram of Akshara Recognition with GUI and DLL. Akshara 1 recognized with stroke 2, 3 and 61.

When a stroke is written on the GUI, the parameters of API provide the basic data like 2-dimensional coordinate traces of the stroke. The dynamic linked library when provided with raw handwritten trace, it initiates pre-processing for refinement of the 2-dimensional trace and then performs classification tasks with the integrated HMM testing model and outputs the set of labels of recognized stroke. The set of strokes are then checked with language rules to verify whether there exist a valid Akshara for the respective strokes. If yes, then output the valid Akshara in GUI text box.

4. COMPARISON RESULTS BETWEEN HMM AND SVM CHARACTER RECOGNITION SYSTEM

During testing, log likelihood values are obtained from HMM classifier while SVM classifier gives probability estimates as output. The output from both the stroke recognizers are then compared using two approaches. In the first approach, the confusion matrix is obtained from both the HMM classifier and SVM classifier and the confusion patterns are analysed. The confusion matrix for the first 10 classes out of the 181 classes obtained using HMM is shown in Figure 6 and the confusion matrix for SVM classifier is shown in Figure 7. In the second approach, users are allowed to write the stroke patterns in the testing tool obtained by integrating the GUI provided by CDAC, Pune once with the HMM classifier and once with the SVM classifier and accuracy of both the classifiers are studied manually. The developed stroke classifier gives average recognition accuracy of about 94 % in case of HMM and 96 % in case of SVM. The akshara level average performance is 84.67 % in case of HMM and 86.23 % in case of SVM

stroke no.	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10
	—	୨	୩	୪	୫	୬	୭	୮	୯	୧୦
st1	69.68	0	0	8.30	0	1.21	0	0	0.75	0.15
st2	0	92.58	0	0	0	0	0	0	0	0
st3	0	0	94.67	0	0	0	0	0	0	0
st4	0.15	0	0	94.93	0	0.31	0	0	0.77	0
st5	0	0	0	0	85.71	0	0	0	0	0
st6	0.24	0	0	0	0	86.03	0	0	0.12	0
st7	0	0	0	0	0	0	95.06	0	0	0
st8	0	0	0	0	0	0	0	90.03	0	0
st9	0	0	0	0	0	0	0	0	98.31	0
st10	0	0	0	0	0	0	0	0	0	96.98

Figure 5: confusion percentage matrix for the first 10 strokes using HMM classifier

stroke no.	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10
	—	୨	୩	୪	୫	୬	୭	୮	୯	୧୦
st1	69.71	0	0	4.50	0	0	0	0	0.75	0
st2	0	92.58	0	0	0	0	0	0	0	0
st3	0	0	94.68	0	0	0	0	0	0	0
st4	0.09	0	0	95.01	0	0.31	0	0	0.77	0
st5	0	0	0	0	85.71	0	0	0	0	0
st6	0.11	0	0	0	0	86.58	0	0	0.12	0
st7	0	0	0	0	0	0	95.08	0	0	0
st8	0	0	0	0	0	0	0	90.03	0	0
st9	0	0	0	0	0	0	0	0	98.31	0
st10	0	0	0	0	0	0	0	0	0	97.02

Figure 6: confusion percentage matrix for the first 10 strokes using SVM classifier

5. CONCLUSION

We have observed that the feature vector namely the coordinate trace, first derivate and second derivative works perfectly with HMM based system, however the recognition accuracy reduces significantly when the second derivate is used as a feature in SVM based system. Therefore in SVM based system only the coordinate trace and first derivate is used as a feature. In HMM based system the recognition accuracy reduces if second derivate is not used in the feature set. The SVM based system outperforms HMM based system by 2% in stroke accuracy and 1.56% in akshara case. The performance is almost similar. The performance might improve if we consider a larger set of database than currently used in SVM case.

REFERENCES

- [1] N. Arica and F. T. Yarman-Vural, "An Overview of Character Recognition Focused on off-line Handwriting," IEEE Trans. Systems, Man, Cybernetics Part C: Applications and Reviews, vol. 31, no. 2, pp. 216-233, May 2001
- [2] S. K. Parui, K. Guin, U. Bhattacharya, and B. B. Chaudhuri, "Online Bangla Handwritten Character Recognition using HMM," in Proc. 19th Int. Conf. on Pattern Recognition (ICIP), pp. 1-4, Tampa FL, 2008
- [3] V. J. Babu, L. Prasanth, R. R. Prasanth, R. R. Sharma, G. V. P. Rao and A. Bharath, "HMM-based online handwriting recognition system for telugu symbols," in Proc. 9th Int. Conf. on Document Analysis and Recognition (ICDAR), Curitiba, Brazil, 2007, pp. 63-67
- [4] K. Shashikiran, K. S. Prasad, R. Kunwar, A. G. Ramakrishnan, "Comparision of HMM and SDTW for Tamil handwritten character recognition." in Proc. Int. Conf. on Signal Processing and Communications, pp. 1-4, IISc Bangalore, India, 2010
- [5] A. Arora and A. M. Namboodiri, "A Hybrid Model for Recognition of Online Handwriting in Indian Scripts," in Proc. of Int. Conf. on Frontiers in handwriting Recognition, pp. 433-438, Kolkata, 2010

- [6] G. S. Reddy, P. Sharma, S. R. M. Prasanna, C. Mahanta and L. N. Sharma, "Combined Online and Offline Assamese Handwritten Numeral Recognizer", in Proc. of 18th National Conference on Communications, pp. 1-5, 2011
- [7] G. S. Reddy, B. Sarma, R. K. Naik, S. R. M. Prasanna and C. Mahanta, "Assamese Online Handwritten Digit Recognition System using Hidden Markov Models", accepted at the Workshop on Document Analysis & Recognition, 2012
- [9] A. Jayaraman, C. Chandra Sekhar and V. S. Chakravarthy, "Modular Approach to Recognition of Strokes in Telugu Script", in Proc. Of 9th International Conference on Document Analysis and Recognition, pp. 501-505, 2007
- [10] V. J. Babu, L. Prasanth, R. R. Prasanth, R. R. Sharma, G. V. P. Rao and A. Bharath, "HMM-based online handwriting recognition system for telugu symbols," in Proc. Of 9th Int. Conf. on Document Analysis and Recognition Brazil, 2007, pp. 63-67.
- [11] X. Li and D. Y. Yeung, "Online handwritten alphanumeric character recognition using dominant points in strokes", Pattern Recognition, vol. 30, no. 1, pp. 31-44, 1997.
- [12] S.R.M Prasanna, Rituparna Devi, Deepjoy Das, Subhankar Ghosh, Krishna Naik, "Online Stroke and Akshara Recognition GUI in Assamese Language Using Hidden Markov Model" International Journal of Scientific and Research Publication, ISSN 2250-3153
- [13] <http://htk.eng.cam.ac.uk/>
- [14] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," Proc. Of IEEE, vol. 79, no. 2, pp. 257-286, 1989.
- [15] G. Siva Reddy, Bandita Sharma, R. Krishna Naik, S.R.M Prasanna, Chitralekha Mahanta, "Assamese Online Handwritten digit recognition system using Hidden Markov Models" in Proc Of the workshop on Document Analysis and Recognition, pp. 108-113, DAR'12.
- [16] Nello Cristianini and John Shawe-Taylor, —An Introduction To Support Vector Machines And Other Kernel-Based Learning Methods, Cambridge University Press, 2000.
- [17] K. Kim, Financial Time Series Forecasting Using Support Vector Machines, Elsevier, March 2003
- [18] C. Hsu, C. Chang, C. Lin, A Practical Guide to Support Vector Classification, National Taiwan University, 2003
- [19] Z. Hua, Y. Wang, X. Xu, B. Zhang, L. Liang, Predicting Corporate Financial Distress Based on Integration of Support Vector Machine and Logistic Regression, Expert Systems with Applications, 2007

INTENTIONAL BLANK

GRAPHITE: A GRAPH SEARCH FRAMEWORK

Sushanta Pradhan

Talentica Software Pvt Ltd., Pune, Maharashtra
sushanta.pradhan@talentica.com

ABSTRACT

With the recent data deluge, search applications are confronted with the complexity of data they handle in terms of volume, velocity and variety. Traditional frameworks such as Lucene [1], index text for efficient searching but do not consider relationships/semantics of data. Any structural change in data demands re-modelling and re-indexing. This paper presents an indexing model that addresses the challenge. Data is modelled as graph in accordance to object-oriented principles such that the system learns the possible queries that can be executed on the indexed data. The model is generic and flexible enough to adapt to the structural changes in data without the need of additional re-modelling & re-indexing. The result is a framework that enables applications to search domain objects, their relationships and related objects using simple APIs without the structural knowledge of underlying data.

KEYWORDS

Graph Search, Semantic Web, Object Oriented Programming, Data Modelling, Algorithms

1. INTRODUCTION

Search is a two-step process:

1. Formulation of a query
2. Execution of the query to extract results.

How does a traditional search application tackle query formulation? Traditional search application left query formulation to the user; it either does not offer or offers little help in this area. Whatever small help it offers, in terms of ‘auto text completion’ and ‘related searches’, is based on search patterns of other users. Such help does not work very well as it does not consider semantics of data or user’s personality and preferences. In a typical search application users either need to have knowledge of the kind of queries supported or will need to manually filter through the results to extract relevant data. Having the search application prompt the user with relevant queries that can be answered not only lowers the burden on the user but also helps the search application to provide relevant results.

How does a traditional search application tackle extracting results? Traditional search application deploys text based information retrieval techniques to retrieve results and hence only explicit results are returned. For example, first few result pages for the query – ‘mobile phone’ on popular web search engines, such as Google, Bing, and Yahoo etc. do not have any mention of the inventor of mobile phone – “Martin Cooper”. That is to say, search engines just looked at the pages, which contained the text ‘mobile phone’ but did not consider the relation between ‘mobile

phone' and 'Martin Cooper' while processing the query – 'mobile phone'. Having search engines consider these relations will enable users to discover more information and gain more knowledge – the most valuable asset in knowledge economy.

With the current data explosion on the Internet, users expect search applications to not only help with query formulation but also provide them relevant information/knowledge that they did not explicitly ask for. Social Graph [2], Knowledge Graph [3], Interest Graph [4], Linked Data [5] are concepts that are currently echoed in the world of Internet. All these concepts focus on the relationships that exist between data and not just the data itself. As applications running on Internet begin to embrace these concepts, it becomes imperative for search applications to also do the same. Graphite is a framework that enables search applications to cater to this demand in today's evolving Internet ecosystem.

The rest of the paper is structured as follows:

Section 2 describes the data model.

Section 3 describes the core components of the system.

Section 4 includes a reference Java implementation.

Section 5 describes the APIs exposed by graphite.

Section 6 briefly describes a graphite backed search application.

Conclusions and related work & future directions are discussed in Section 7 and Section 8 respectively.

2. DATA MODEL

In order to store and make queries based on relationships, we have chosen a 'graph' data structure to model the data. Object Oriented Principles (OOP) are widely used for developing software applications as it closely models the real world and has an inherent graph structure, which we choose to harness. The following entities are modeled in graphite:

1. Domain Class – defined as Class Node
2. Domain Property – defined as Property Node
3. Domain Object/Instance – defined as Object Node

2.1. Class Node (Domain Class)

In an object-oriented environment, application is modeled as classes with certain properties and behavior. An object/instance of such a class, with which the user interacts, is defined as Class Node. For e.g. an online booking system will have objects of type Ticket, Place, Person etc. A blogging application will have objects of type Blog, Author, Topic etc. Graphite looks into the structure of a Class Node to formulate queries that a user might want to ask about objects of this class. Each class node has a unique name and is represented as a node in the underlying store, which has an outward 'name' link pointing to a string. For e.g. class – 'Person' (see figure 1) is represented as shown below:

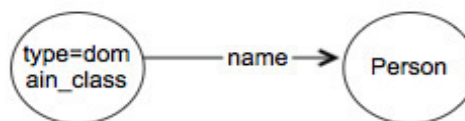


Figure 1: Class Node Structure

2.2. Property Node (Domain Property)

Each property of a class node is defined as Property Node. It has two properties:

1. Range – The type of the property is defined as range.
2. Domain – The class to which the property belongs is defined as domain.

For e.g. Person class (see figure 2) will have properties such as email, location etc.

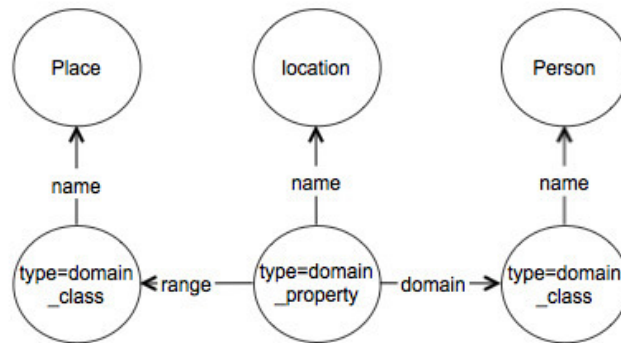


Figure 2: Property Node Structure

2.3. Object Node (Domain Object)

An instance of a class node is defined as Object Node. Each object is represented by a node and has outward property links whose name is same as that of the property name. As shown in figure 3, a domain object named 'ram' has a property named 'location' whose value is 'pune'.

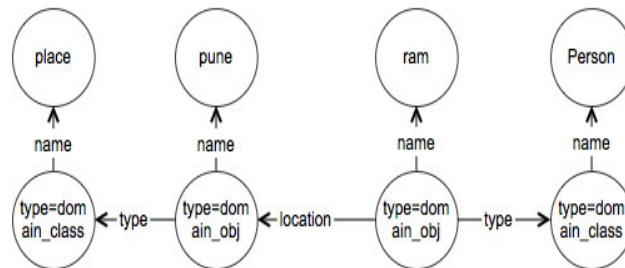


Figure 3: Object Node Structure

3. COMPONENTS

Graphite consists of the following components:

1. Indexer – Stores/Indexes raw data as per the Data Model.
2. Query Advisor – Formulates queries that could be answered by underlying data.

3. Object Searcher – Executes queries formulated by query advisor on indexed data.
4. API – Interface for applications to interact with above three components.

3.1. Indexer

Indexer transforms raw data (classes & instances) to nodes as defined by the data model. Indexing is a two-step process of creating nodes and bonds between them. In the first stage all defined domain classes and their respective properties are added to the index, i.e. Class Nodes and Property Nodes are created. Classes and properties are indexed not just by their name (as is done by a text based indexer) but also by their relationships in the form of ‘domain’, and ‘range’ links (as described in Section 2.2). In the second stage instances of an indexed class are added to the index in conformance with the class structure, i.e. Object Nodes are created. Nodes are indexed not just by their names but also by their relationships in the form of ‘type’ and property links (as described in Section 2.3).

3.2. Query Advisor

This is the gateway to access queries that can be run on the index created by the Indexer. It takes a literal as input, looks into the index and returns possible complete or incomplete [incomplete query isn’t syntactically complete to be executed on the indexed data whereas ‘Complete Query’ can be executed on the indexed data] query objects. Resultant query objects have two types of query embedded in them:

1. Human Readable Query- An English sentence that can be understood by a non-technical end user.
2. Machine Readable Query- A SQL like query that is understood and can be executed by the underlying store, is empty for incomplete queries.

Query Advising is a multi-step process to create possible queries that can be executed on the indexed data. First stage is the seeding stage wherein a literal is fed to the Query Advisor. Here a broad list of relevant complete/incomplete queries related to the inputted literal is returned. In the subsequent stages feedbacks, in the form of expand & contract commands (details discussed in section 4.1.3), from the user is used to complete the selected query.

3.3. Object Searcher

Object Searcher takes a query object as input and returns a list of object nodes that satisfies the given query. It also takes an object node as input and returns its connections, along with the connection strength, for a given level and a given query.

4. IMPLEMENTATION

This section describes a reference Java implementation of graphite.

4.1. Architecture

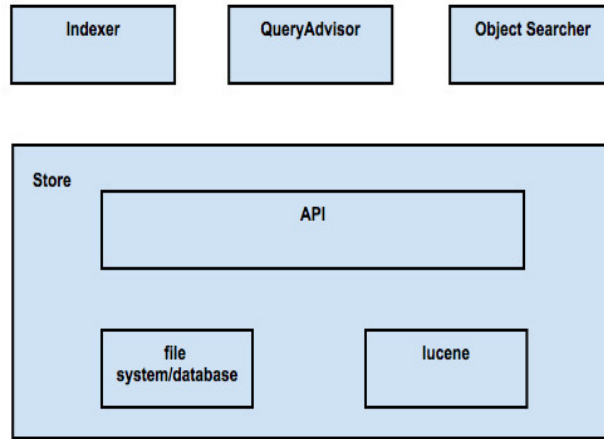


Figure 4: Architecture

4.1.1. Store

This component manages the store and retrieval of nodes. Nodes are persisted either in flat files or database. We have chosen neo4j [6] – a graph database to persist indexed data. Neo4j has support for property graphs, where each node and relationship can have arbitrary properties. To support the ability to have machines interpret data and suggest queries, we have to have a standard and hence we choose RDF [7] graph over property graph. RDF graph has the following advantages:

1. Since RDF is schema less, addition and deletion of indexed data does not demand changes in Graphite application, thus reducing development and maintenance effort for graphite applications.
2. All graphite applications can seamlessly interact and collaborate with each other to produce better search results, thus allowing search application to expand their search horizon without the need for additional integration effort.
3. Object Oriented Principles seamlessly get translated into RDF graph, hence making it easy for the application developers to develop graphite applications.

Each entity (as described in section 2) constitutes a node in the graph and nodes are logically structured into three layers similar to the structuring of carbon nodes in the mineral-‘graphite’. Nodes at Layer1 are literals and indexed using Lucene for faster retrieval.

Layer1: The top layer consists of string and number nodes – objects of primitive types.

Layer2: The middle layer consists of class and property nodes.

Layer3: The bottom most layer consists of object nodes.

Nodes of a given layer can be linked (bonded) with another node residing in the same or a different layer. For example, Class definitions can be represented as a graph by considering two classes – ‘Book’ & ‘Author’ and their respective instances ‘C++’ and ‘Stroustrup’.

```

class Book{
    Author author;
}
class Author{
    String fullname;
}

```

As seen in the Figure 5 'Book' (#22) and 'Author' (#23) are class nodes and their properties 'author' (#24) and 'fullname' (#21) respectively are property nodes in Layer2. Both the class node and the property node are linked (bonded) by the 'name' link to the string node residing in Layer1. Class node is linked by 'domain' link to their respective property nodes, for e.g. 'Book' (#22) & 'author' (#24) are linked and 'Author' (#23) & 'fullname' (#21) are linked. Property node is linked by 'range' link to their respective type (class) node, for e.g. 'author' (#24) and 'Author' (#23) are linked. Object nodes 'C++' (#31) and 'Stroustrup' (#32) reside in Layer3. These object nodes are linked to nodes, in Layer1 by name link, in Layer2 by type link, in Layer3 by property links. All links except 'name' and 'type' are created in adherence with the links of class node (instance type) and property nodes in Layer2.

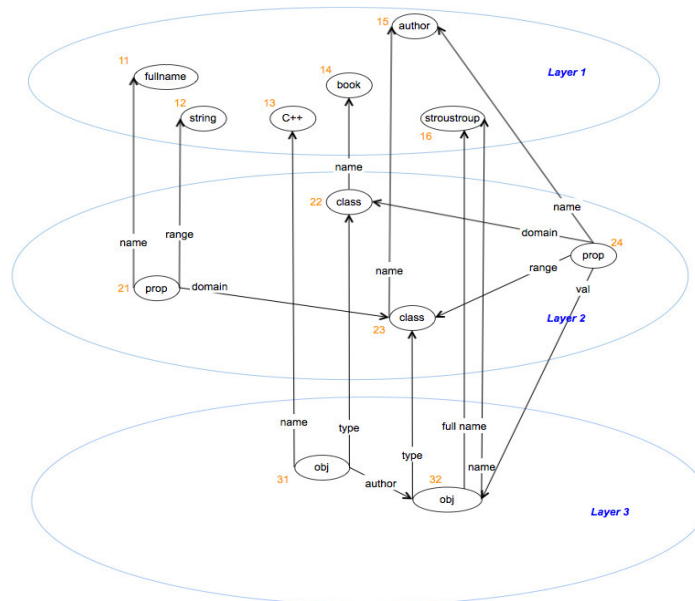


Figure 5: Data Model

4.1.2. Indexer

This is the gateway to write nodes to store; a graphite application interacts with this component to persist nodes as per the data model described in Section 2. It exposes APIs (discussed in section 5) to define domain classes, domain properties and domain objects and add them to store.

4.1.3. Query Advisor

Query Advisor manages the formulation of queries starting from a keyword and subsequent commands: expand or contract. Since the starting point is a keyword, it starts from a node of Layer1 and follows the name link to fetch the first node of query stack (described below). Node at

Layer1 is linked (bonded) to either class, property node of Layer2 or object node of Layer3, hence Query Advisor produces two types of queries based on type of node it finds.

Diverging Query: Following the name link, if the query advisor lands on Layer3, it produces a diverging query. This type of query is called diverging query because the graph structure that gets outputted after execution is of diverging nature as seen in figure 6. The output nodes for such queries are the nodes at level 1.

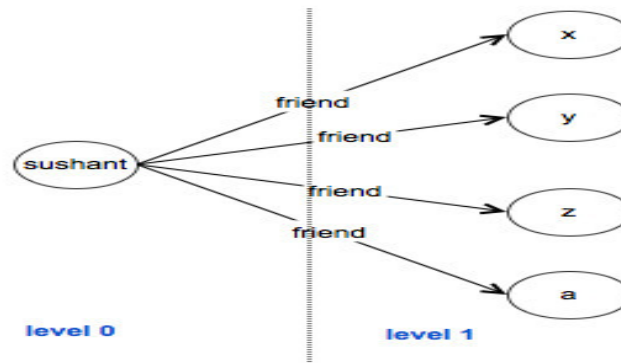


Figure 6: Diverging Query Result Structure

Converging Query: Following the name link, if the query advisor lands on Layer2, it produces a converging query. The graph structure that gets outputted is converging in nature as seen in figure 7 and hence the name. Nodes of level 0 are outputted for such queries.

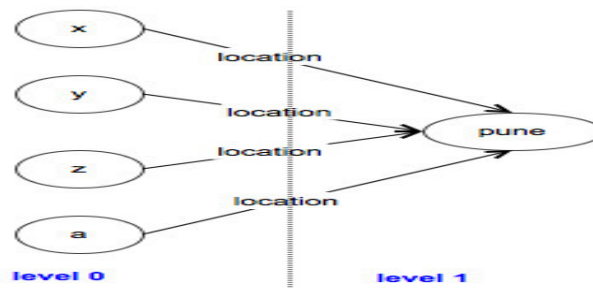


Figure 7: Converging Query Result Structure

Query Stack: It is a collection of nodes stacked one above another in accordance to the structure of class nodes residing in Layer2. It is initialized by name of either a class node or an object node, which becomes the first element of the stack. After a query chain is initialized which can be either expanded or contracted. On expansion, Query Advisor checks the type of node at top of stack and performs the following actions.

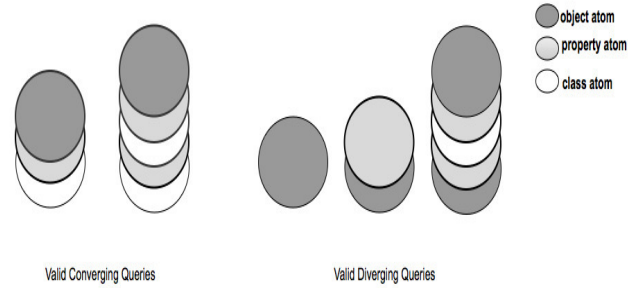


Figure 8: Query Stacks

If the top node is found to be a class node, Query Advisor follows the domain link and fetches all property nodes linked to the class node. Since there could be more than one property node, it creates as many copies of the original Query Stack as the number of property nodes found and pushes one property node per copy. Finally this list of Query Stack is returned.

If the top node is found to be an object node, Query Advisor follows the type link and fetches the class node from Layer2. It assumes this class node to be at the top of query stack and performs the expansion as explained above.

If the top node is found to be a property node, Query Advisor follows the 'val' link to find the linked object nodes at Layer3 and range link to find linked class nodes. Since there could be more than one node found, it creates as many copies of original Query Stack as the number of nodes found and pushes one node per copy. Finally the list of Query Stack is returned. Contraction of a query stack simply removes the top node.

Human Readable Query: Human Readable Query is translation of Query Stack to an English sentence by the Query Advisor that a non-technical user can understand. A property of a class can be considered as possessive adjective [8] for the noun [9] - class and hence the class, property and its value can be translated into an English language sentence as "*<class_name> whose <property_name> is <property_value>*". Using this principle a Query Stack of Diverging Query is translated into perceivable human readable query. In case of a Converging Query, the sentence is of the form "*<object_name>'s <property_name>*". A Human Readable Query can be generated for a Query Stack irrespective of its current state.

Machine Readable Query: Similar to Human Readable Query, Machine Readable Query is a translation of a Query Stack into a form that is executable by Object Searcher. As nodes in the query stack are stacked in accordance with links that exist between them, they are parsed to form cypher [10] query. For e.g. to find 'all books authored by 'Stroutstrup' the cypher query would be (see Converging Query of figure 9) - "*start book_class = node(22), author = node(32) match (book_class)-[:type]-(book)-[:author]-(author) return book;*". Similarly the author of the book named 'C++' the cypher query (see Diverging query of figure 8) would be - "*start book = node(31) match (book)-[:author]->(authors) return author;*". Unlike Human Readable Queries, Machine Readable Query is generated only when Query Stack is in complete state.

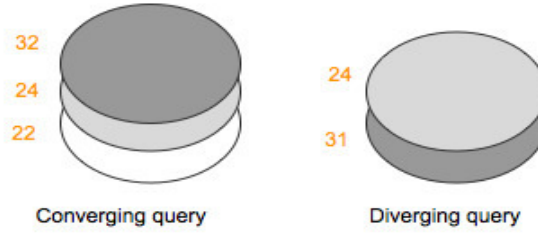


Figure 9: Query Stack Example

4.1.4. ObjectSearcher

Object searcher is primarily responsible for searching nodes of Layer3 that satisfies a given query. It does so by executing the Machine Readable Query extracted from the corresponding Query Stack. Two lists are returned when a Machine Readable Query gets executed:

1. List of ids corresponding to object nodes that satisfy the given query. For e.g. for the converging query (of Figure 9), the list: [31,] is returned.
2. List of related class node ids. For e.g. for the converging query (of Figure 9), the list: [23,] is returned.

Object Searcher also extracts connections of a given object node along with their routes and strength. Strength (S_c) of a connection gives a measure of how strongly two objects are connected to each other. It is given by the formula:

$$S_{(c,o)} = \sum_r S_{(c,o)}^r$$

$S_{(c,o)}$ – Strength of a connection (c) of object node (o).

$S_{(c,o)}^r$ – Strength of a route (r) of a connection (c) of object node (o). Route is simply the edges and nodes that are traversed to reach the connection from the object node. Strength of a route is given by the formula:

$$S^r = \frac{\sum_a R_a}{L_r^2}$$

R_a – Rank of node (a).

L_r – Path length of route (r).

5. API

This is an interface for applications to interact with Graphite i.e. to index data and retrieve it back. Annotations are used to annotate domain classes whose instances are to be indexed. Indexer and Search APIs are used to index and retrieve data respectively.

5.1 Annotations

@ClassNode: A class level annotation that is used to define a Domain Class. Its arguments define 'name' and 'rank' of the annotated java class. Default value of 'name' is the name of the class itself and default value for 'rank' is 1. The rank value is used for ranking the search results, which is described in detail in section 4.1.4.

@PropertyNode: A property level annotation used to annotate class properties. Its arguments define 'name' and 'rank' of the annotated property.

Writing a Graphite program involves the following steps:

1. Indexing
2. Searching

5.2. Indexing

At this stage, domain classes are defined along with their relations to other domain classes and added to the index. Thereafter objects/instances of domain class are added to the index.

5.2.1. Adding a domain class to the index

Domain class is defined by sub classing the predefined 'ObjectNode' class and annotating it with @ClassNode annotation (see Figure 10). Index-able properties are annotated with the @PropertyNode annotation. The Class- GraphiteAnnotationParser searches a given package and adds the domain classes along with their properties to the index (see Figure 10).

```
@ClassNode(rank=2)
public class Author extends ObjectNode{
    public Author(String name) {
        super(name);
    }
    @PropertyNode(name="author", rank = 2)
    private String fullname;
}
```

Figure 10: Class Node Definition

5.2.3. Adding a domain object to the index

ObjectNodeWriter.write(object) API is called to add a domain object to the index (see Figure 11). Only instances of a domain class can be indexed.

5.3. Searching

At this stage user makes a query to get valid queries and passes one of these queries to the search API to fetch domain objects honoring the given query. It is a three-step process:

1. Querying for queries
2. Querying for objects
3. Querying for connections

5.3.1. Querying for queries

The kind of queries that can be run on a given set of data is best known by the data itself. Graphite inspects the indexed data it handles and suggests queries to the user. QuerySuggestor.getQuery(keyword) API is used to get the list of possible queries (see Figure 11).

```
StoreResources sr = new StoreResources ("/tmp", "/tmp/db.prop");
GraphiteAnnotationParser gap = new
GraphiteAnnotationParser("com.talentica.graphite.domain", sr);
//add class nodes to index
gap.parse();
ObjectNodeWriter writer = new ObjectNodeWriter(sr);
Author author = new Author("Stroustrup");
//add object node to index
writer.write(author);
QuerySuggestor qs = new QuerySuggestor(sr);
//get list of queries for the keyword 'author'
List<Query> queries = qs.getQuery("author");
ObjectSearcher searcher = new ObjectSearcher(sr);
//search object nodes for a query
List<ObjectNode> result = searcher.search(queries.get(0));
```

Figure 11: Index and search example

5.3.2. Querying for objects

At this stage the API- ObjectSearcher.search (query) is called with a query to fetch object nodes honoring the given query (see Figure 11).

5.3.3. Querying for connections

ObjectSearcher.getConnections (level, objectId, query) API is called to retrieve the related object nodes for a given object node. The 'level' argument defines the depth to which related objects are to be searched, 'objectId' is the id of the object node whose connections will be searched for and the 'query' argument defines the criteria to be honored by the related object nodes.

6. SEARCH APPLICATION

Graphite is primarily a graph search tool; hence search becomes the primary application of Graphite. By using graphite as a backend, search can be enhanced from text-based to object-based; results that were not explicitly queried for can be outputted. For e.g. search with keyword – 'book', outputs all books in the 'primary pane', while related objects of type - 'author' & 'publisher' are displayed in 'connection pane' (see figure 12). Clicking or hovering over a connection shows the path that was traversed to reach to the connection from the object in the primary pane (arrows in blue in the below figure).

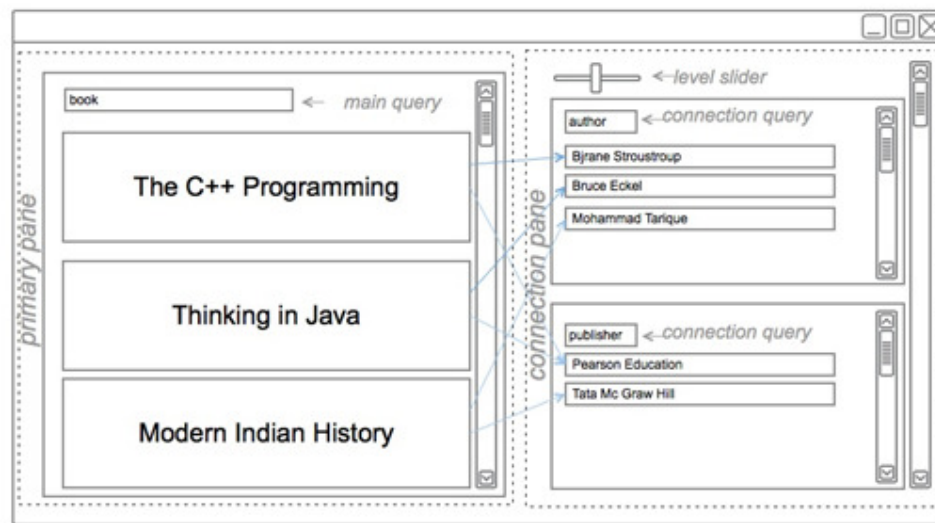


Figure 12: UI of a search application backed by graphite

By modifying a connection query, user can filter/expand his main query results (primary pane). For e.g. if the connection query 'author' is changed to 'Bjarne Stroustrup', elements – 'Thinking in Java' and 'Modern Indian History' would disappear from main query results, and 'Bruce Eckel', 'Mohammad Tarique' and 'Tata Mc Graw Hill' would disappear from connection pane.

By changing the level slider, users would be able to discover more facts/knowledge related to main query results, which in turn would lead to enhanced user satisfaction.

7. RELATED & FUTURE WORK

To our knowledge, there isn't any open-source framework/library similar to Graphite, though there are few proprietary implementations such as Facebook's Graph Search [12] and Google's Knowledge Graph. Facebook's Graph Search, is similar to a Graphite application as it is also object-based, has a query advisor and outputs related objects. Google's Knowledge Graph shows related objects, provides user with results that were not explicitly asked by the user.

We have implemented the basic features required by a search application; issues related to security, scalability has not been analyzed. Future releases of Graphite would focus to address these issues.

Currently, our core focus and inspiration of Graphite was search. Hence we have only concentrated on 'nouns' and 'adjectives'; therefore the data model focuses only on that part of OOP principles. Also, the two common graph models - RDF graph and property graph considers only 'nouns' and 'adjectives'. It would be an interesting research foray to also model 'verbs' as it has a possibility to transform a SEARCH-ENGINE into a DO-ENGINE.

8. CONCLUSIONS

Graphite is an indexing framework, focused on providing users with a natural API for indexing classes, properties, objects and their relationships. In concept, it is similar to Lucene but with a focus on relations.

Traditionally, search applications were built to provide local search to local users i.e. they handled only application data and served only application users. Since users of the application had fair know-how of data provided by the application, users could formulate queries to suffice their needs. From the software developer perspective, maintaining a text-based inverted index using Lucene sufficed to cater such needs of user. Data deluge, increasing use of data by modern day applications (mashup [11]), dynamic nature of data and increasing sources of open data poses problems at the supply side as well as at the demand side.

At demand side, users do (can) not have (gain) fair know-how of data and hence can't formulate efficient queries for search engines to cater their demand efficiently. Graphite's QueryAdvisor helps the user at run-time to gain know-how about data as it can prompt user with valid queries. Graphite's ability to output related objects further enhances user's know-how about data.

At supply side, developers have to constantly make changes to adapt to the changing nature of application data. Graphite's ability to understand OOP principles, developers can write hooks to index searchable data from the core application itself and need not write/modify explicit code for search application.

With increasing sources of open data, users expect search applications to give results from such sources as well and not just the local data. Since Graphite data model embraces the principles of OOPs – a universally used principle to write web applications, graphite applications can collaborate with each other to provide users a rich search experience.

ACKNOWLEDGEMENTS

I would like to thank my managers – Aniket Shaligram and Manjusha Madabushi for providing access to a live product's data for conducting experiments and their valuable comments to improve the model.

REFERENCES

- [1] Michael MacCandless, Erik Hatcher, Otis Gospodnetić. (2010). Lucene in Action
- [2] Social Graph web page: http://en.wikipedia.org/wiki/Social_graph
- [3] Knowledge Graph Web page: http://en.wikipedia.org/wiki/Knowledge_Graph, <http://www.google.co.in/insidesearch/features/search/knowledge.html>
- [4] Interest Graph Webpage: http://en.wikipedia.org/wiki/Interest_graph
- [5] Linked Data Standards webpage: <http://www.w3.org/standards/semanticweb/data>
- [6] Jonas Partner, Aleksa Vukotic, and Nicki Watt. (2012). Neo4j in Action.
- [7] RDF Working Group, "Resource Description Framework (RDF): Concepts and Abstract Syntax", Klyne G., Carroll J. (Editors), W3C Recommendation 10 February 2004
- [8] Possessive Adjective definition: http://en.wiktionary.org/wiki/possessive_adjective
- [9] Noun Definition: <http://en.wiktionary.org/wiki/noun>
- [10] Cypher Query Language webpage: <http://docs.neo4j.org/chunked/stable/cypher-query-lang.html>
- [11] Mashup webpage: [http://en.wikipedia.org/wiki/Mashup_\(web_application_hybrid\)](http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid))
- [12] Facebook Graph Search webpage: <https://www.facebook.com/notes/facebook-engineering/under-the-hood-building-graph-search-beta/10151240856103920>

AUTHOR

Sushanta Pradhan has around 7 years of product development experience mostly in JAVA/J2EE technologies. His interests include semantic web, machine learning, parallel/distributed computing and JVM languages. At Talentica he has architected and built highly scalable and maintainable applications in the digital advertisement and marketing space. He has done his B.Tech from NIT Calicut in Electronics and Communication Engineering.

LOW POWER VLSI COMPRESSORS FOR BIOMEDICAL APPLICATIONS

Thottempudi Pardhu¹, S.Manusha² and K.Sirisha³

¹ Assistant.Professor, Department of E.C.E, Marri Laxman Reddy Institute of Technology & Management, Hyderabad, India

24.pardhu@gmail.com

^{2,3} Department of E.C.E, Marri Laxman Reddy Institute of Technology & Management, Hyderabad, India

sathvika444@gmail.com

smilysiri@gmail.com

ABSTRACT

We present a new design for a 1-bit full adder featuring hybrid-CMOS design style. Our approach achieves low-energy operations in 90nm technology. Hybrid-CMOS design style makes use of various CMOS logic style circuits to build new full adders with desired specifications. The new SERF- full adder (FA) circuit optimized for ultra low power operation is based on modified XOR gates with clock gating to minimize the power consumption. And also generates full-swing outputs simultaneously. The new full-adder circuit successfully operates at low voltages with excellent signal integrity. The new adder displayed better power and delay metrics as compared to the standard full adders. To evaluate the performance of the new full adder in a real circuit, we realized 4-2,5-2,5-3,7-2,11-2,15-4,31-5 compressors which are basically used in multiplier modules of DSP filters. Simulated results using 90nm standard CMOS technology are provided. The simulation results show a 5% - 20% reduction in power and delay for frequency 50MHz and supply voltages range of 1.1 v.

KEYWORDS

SERF Full adder, ultra low power

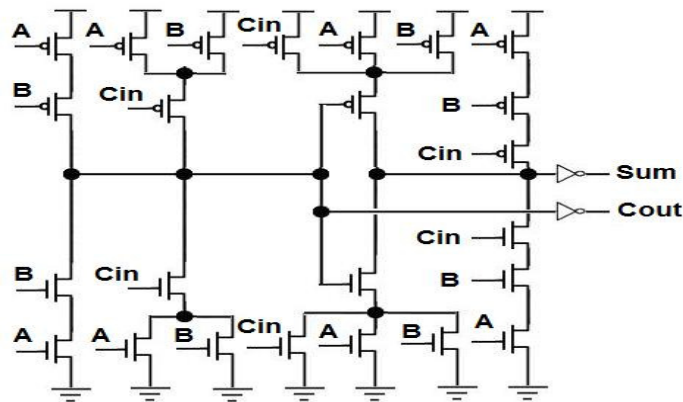
1. INTRODUCTION

The demand for mobile electronic devices of low-power and high-speed is driving designers to design for smaller silicon area, high speed, longer battery life and more reliability. Power dissipation is the limiting factor for hand held devices. as energy-efficiency is one of the most required features for high-performance and/or portable applications. The power-delay product (PDP) metric relates the amount of energy spent during the realization of a determined task, and stands as the more fair performance metric. For high performance design. microprocessors and digital signal processors rely on the efficient implementation of generic arithmetic logic units and floating point units to execute dedicated algorithms such as convolution and filtering [4].In most of these applications, multipliers have been the critical component and adder cells are in the

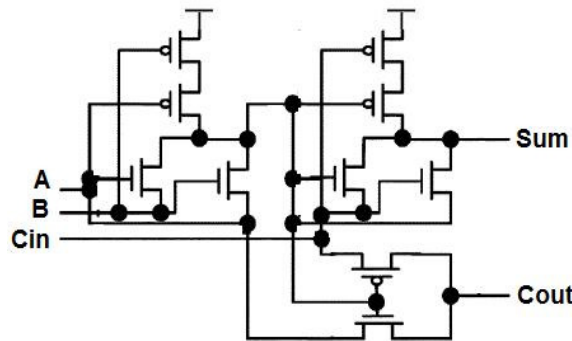
critical paths of these complex arithmetic modules. Therefore design of ultra low power circuits becomes critical for portable applications. Addition is a fundamental arithmetic used in application-specific digital signal processing (DSP) architectures and is the core of many arithmetic operations such as addition/subtraction, multiplication, division and address generation. Thus, the design of a full-adder having low-power consumption and low propagation delay results of great interest for the implementation of modern digital systems. At the circuit level, an optimized design is desired to avoid any degradation in the output voltage, consume less power, have less delay in critical path, and be reliable even at low supply voltage as we scale towards deep submicrometer. However, for ultra low power applications like implants and wireless sensor nodes, the most important design goal is to optimize for low power consumption. Digital hearing aids frequently employ the concept of filter banks whose complexity of computation requires more number of multiplications of higher power consumption. Therefore the construction of filter bank in Digital hearing aid with minimum number of multiplications is a desired design option. [11] Booth Wallace multiplier is used for implementing digital signal processing algorithms in hearing aids for low power consumption.

2. ADDER TOPOLOGIES

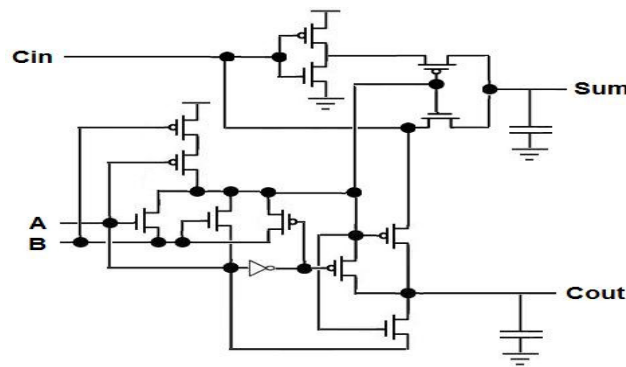
Most adder topologies are based on two XOR gates and one MUX. In [7] [10], [11.], different circuit topologies have been analyzed and simulated in different ranges of supply voltages. One of the most well known full adders is the standard static CMOS full adder that uses 28 transistors. In [4] the sense energy recovery full adder (SERF) is presented. The topology of this circuit is shown in Fig.c which requires only 10 transistors to implement it. In [12] different full adder topologies with a low number of transistors are presented.



(a) C-CMOS Full Adder [10]



(b) SERF Full Adder Architecture



(c) Modified SERF full adder

A. Multi-Operand Addition (> 2)

To avoid chaining of adders and to calculate the sum of multiple operands, two methods are in use. 1. Adder arrays 2. Adder trees. Adder arrays are constructed as a linear arrangement of either carry-propagate (CPA) or carry-save adders (CSA). In the latter case, a carry propagate adder is used to merge the carry and sum vectors. The fastest implementation is achieved by an array of CSAs followed by a fast CPA. Array adders have a more regular structure and lower interconnection, but lower performance when compared to adder trees. Adder trees are constructed by a tree arrangement of compressors followed again by a carry propagate adder. In this way carry propagation is only performed once and postponed until after the tree. Effectively, addition is performed in carry-save format and the final carry-propagate can be perceived as a conversion layer between the carry-save and the standard 2's complement number representation. An adder tree made of full-adders is commonly known as a wallace tree [1].

B. Compressors for high-speed arithmetic circuits

Multiplexor (MUX) is used extensively in the digital design, for the efficient design of arithmetic and logic circuits. The CMOS implementation of MUX [22], performs better in terms of power and delay compared to exclusive-OR (XOR). Suppose, X and Y are inputs to the XOR gate, the output is $XY + \overline{X}\overline{Y}$. The same XOR can be implemented using MUX with inputs $X; \overline{X}$ and select bit Y . Efficient compressors have been designed using MUX. In the proposed compressors, both output and its complement of these gates are used.

1) Description of Compressors

A $(p; 2)$ compressor has p inputs $X_1; X_2 : : X_{p-1}; X_p$ and two output bits (i.e., Sum bit and Carry bit) along with carry input bits and carry output bits. A full-adder is effectively a $(3,2)$ compressor that encodes three input bits to two. A $(4,2)$ achieves a higher compression rate and timing performance for the same area. For this reason it should be preferred for the construction of high fanin trees or the design of larger compressors. A $(5; 2)$ compressor takes 5 inputs and 2 carry inputs and a $(7; 2)$ compressor takes 7 inputs and 2 carry inputs. Block diagrams of these compressors are shown in Fig. 1.

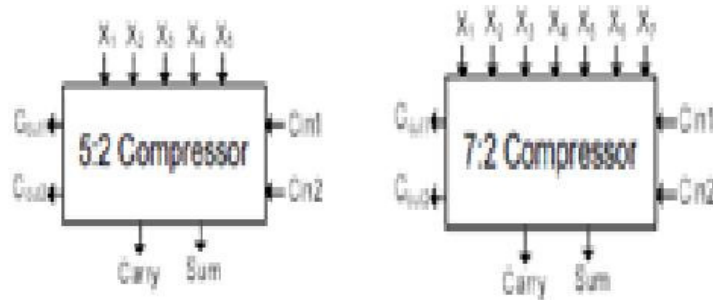


Fig. 1: Compressors (5:2,7:2)

Efficient designs of the existing XOR-based 5:2 and 7:2 compressors have critical path delays of $4\Delta(XOR)$ and $6\Delta(XOR)$ (delay denoted by Δ), respectively.

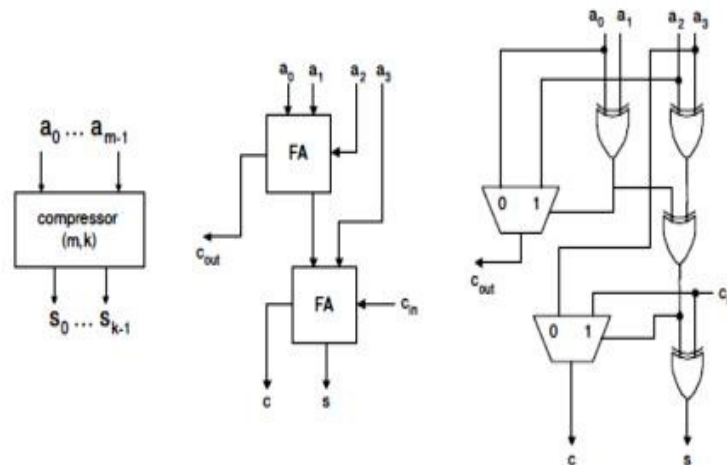


Fig. 2: 4:2 Compressor using FA

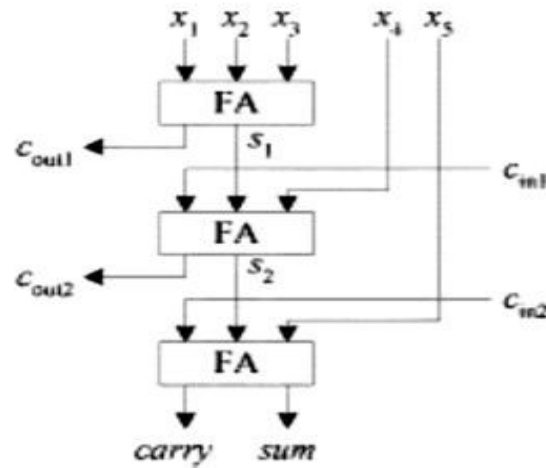


Fig. 3: 5:2 compressor using FA

C. 5-2 Compressor Architecture

The 5-2 compressor is another widely used building block for high precision and high speed multipliers. The block diagram of a 5-2 compressor is shown in, Fig 4 which has seven inputs and four outputs. Five of the inputs are the primary inputs x_1, x_2, x_3, x_4, x_5 and c_{in1}, c_{in2} the other two inputs, and receive their values from the neighboring compressor of one binary bit order lower in significance. All the seven inputs have the same weight. The 5-2 compressor generates an output SUM of the same weight as the inputs, and three outputs CARRY, COUT1, COUT2 weighted one binary bit order higher. The outputs COUT1, COUT2, are fed to the neighboring compressor of higher significance.

D. Multipliers

A fast array or tree multiplier is typically composed of three subcircuits: a Booth encoder for the generation of a reduced number of partial products; a carry save structured accumulator for a further reduction of the partial products' matrix to only the addition of two operands; and a fast carry propagation adder (CPA) [9] for the computation of the final binary result from its stored carry representation. Among these subcircuits, the second stage of partial product accumulation, often referred to as the carry save adder (CSA) tree [1], [13], occupies a high fraction of silicon area, contributes most to the overall delay, and consumes significant power. Therefore, speeding up the CSA circuit and lowering its power dissipation are crucial to sustain the performance of the multiplier to stay competitive. To lower the latency of the partial product accumulation stage, 4-2 and 5-2 compressors have been widely used for high speed multipliers. Replacing an adder array with a Wallace adder tree results in a Wallace multiplier and modified Wallace tree is shown in Fig 5.

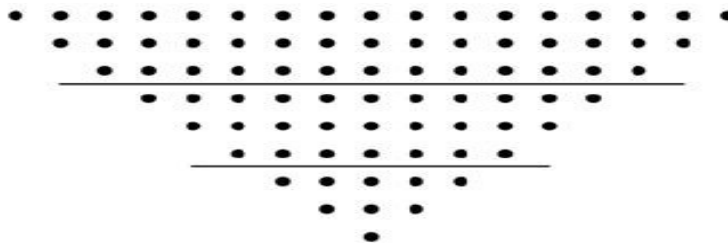


Fig. 4: Modified Wallace tree

3. PROPOSED SERF FULL ADDER

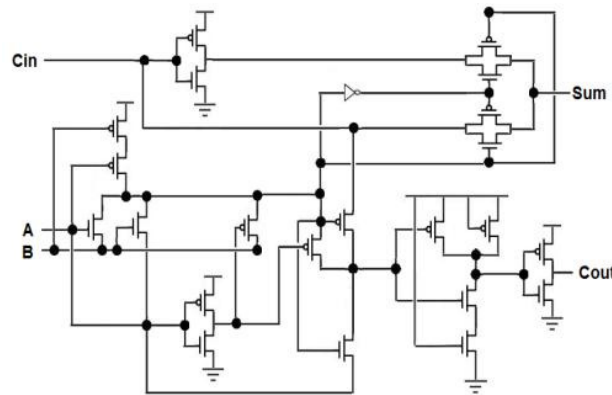


Fig. 5: Proposed FA based on SERF

The proposed Low power Full adder (Fig 5) uses the Novel XNOR circuit and the Transmission Gate based MUX along with level restoring clock gating technique for its operation. It is based on the Energy Recovery Concept [2]. By using Clock Gating Technique [14] the Carry out signal strength can be restored and also the use of AND gate which is one the most popular Low power Technique will reduce the power dissipation.

The gates of the pass transistors are connected to V_{DD} instead of $V_{DD}-V_{th}$, thus the power dissipating path is removed due to the completely turned-off PMOS transistors, which makes the output to rise till $V_{DD}-V_{th}$ from of $V_{DD}-2V_{th}$ [2], limiting the driving capability of the circuit. This reduction in output voltage (threshold voltage drop problem) makes cascading of pass transistor circuits difficult. And the use of such adder circuits in Compressors, Carry Propagation adders and Multiplier would lead to incorrect results. In the proposed design the Transmission Gate multiplexer is used so as to ensure full voltage swing at the sum output which will compensate for the loss of V_t . For input pulses of $A=200\text{MHz}$, $B=100\text{MHz}$ and $Cin=50\text{MHz}$ at $1.1V V_{DD}$, the modified serf FA (Fig 7) showed power dissipation of $58.65\mu\text{W}$ and a delay of 7.05nS , whereas the proposed FA recorded power dissipation of $23.99\mu\text{W}$ and a delay of 6.09nS

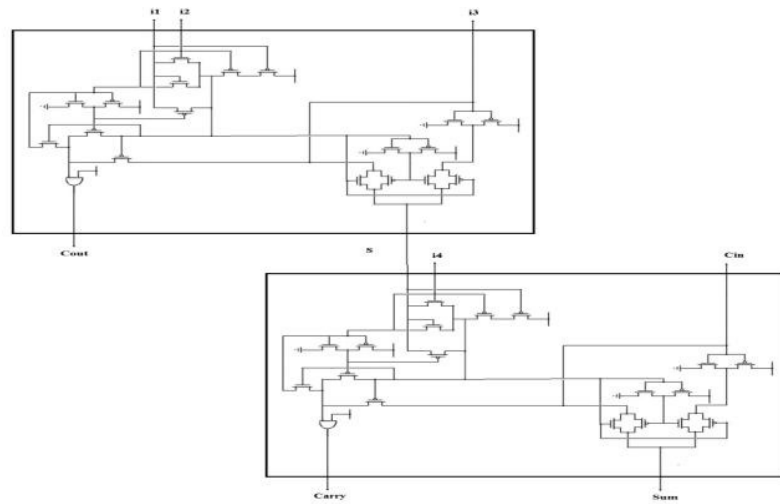


Fig. 6: 4:2 Compressor with proposed FA

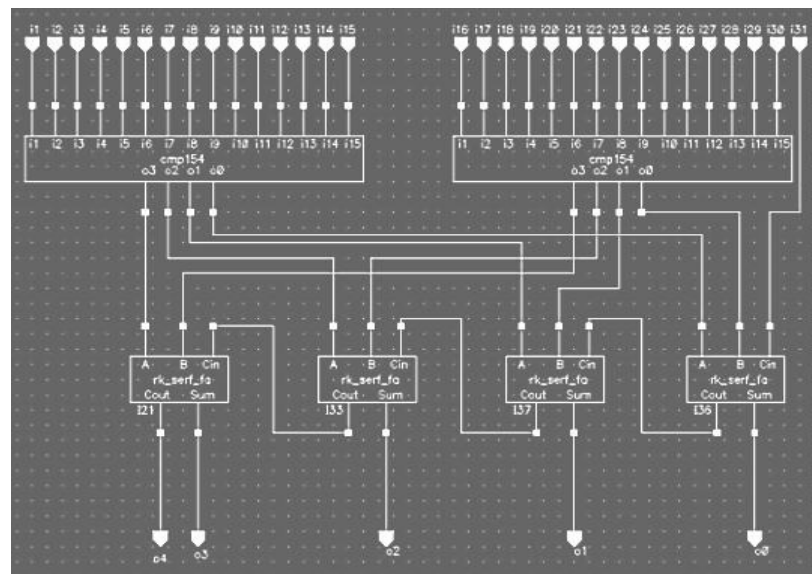


Fig. 7: 31: 5 compressor using lower level compressors

The compressors 11:2, 15:2, 31:2 which can be used in MAC units of DSP circuits are realized using the proposed full adder structure. The compressors speed up the addition process of partial products generate in a multiplier in turn reducing the delay and the adder used

Table I: PDP Full Adders (VDD =1.1V) CMOS 90nm Technology

FA Design	Power Delay Product (fJ)			
	5MHz	50MHz	100MHz	200MHz
C-CMOS	1403.100	221.8373	164.4129	144.8854
O-TGA	217.4269	174.5949	185.4085	212.7918
MBA-12T	05.60061	0.282211	0.25088	0.105492
ULPFA	263.4342	202.7862	217.5503	247.5988
Mfd SERF	6778.766	738.1514	398.0437	194.7882
Proposed FA	119.4373	048.9646	51.2313	054.7741

reduces the power dissipation due to AND operation used which is a kind of low power technique for power optimization.

4. SIMULATION RESULTS AND DISCUSSION

The Circuit Simulation works were carried out using numerous random input vectors with SPECTRE simulator in CADENCE VIRTUOSO Analog Design Environment under CADENCE CMOS 180nm and 90nm Technology respectively. The Power-delay product of the various full adder designs are listed in the Table I for comparison. The

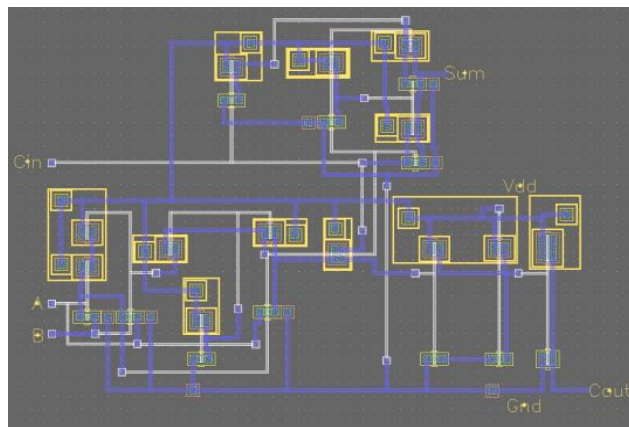


Fig. 8: Proposed adder layout

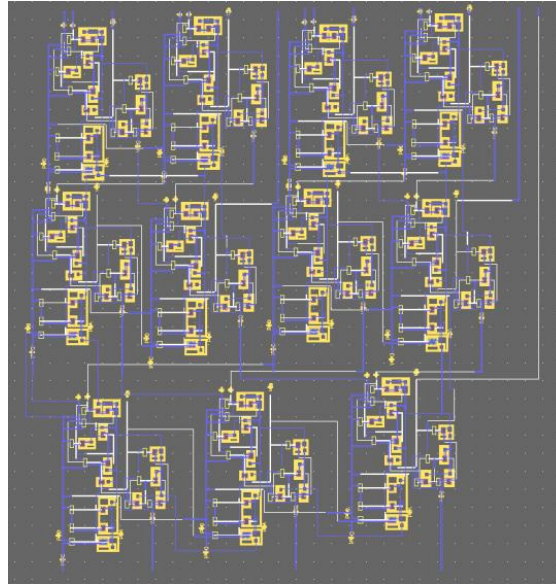


Fig. 9: 15:4 compressor Layout

layouts for all the adders and compressors were drawn using LAYOUT XL of CADENCE and DRC,LVS were run using ASSURA, RC parasites were extracted. In order to establish an impartial simulation environment, we preferred to give various input patterns as which covers every possible input combination of A, B, and C_{in} .

5. CONCLUSION

In this paper, we proposed a 24-transistor full-adder operating at 1.1 V power supply. The performance of various full-adders were compared and the simulation results proved that the proposed full-adder dissipate the lowest power consumption and has lowest PDP (Power Delay Product) and they can be used as a building block in compressor, multiplier and multiple and accumulate units that are used in hand held devices like Digital hearing aids which frequently employ the concept of filter banks. One of the major drawbacks of these techniques is the complexity of computation requiring more number of multiplications increasing the power consumption. Therefore the proposed multiplier architectures can be used as a new approach to speech enhancement for the hearing impaired and also the construction of filter bank in Digital hearing aid with minimum number of multiplications.

REFERENCES

- [1] N. Weste and K. Eshraghian, Principles of CMOS digital design. Reading, MA: Addison-Wesley, pp. 304–307.
- [2] R. Shalem R., E. John E., and L. K. John L. K.” A Novel Low Power Energy Recovery Full Adder Cell”. Proc. of the Great Lakes Symposium of VLSI, Feb. 1999, pp. 380-383.
- [3] R. Pedram R. and M. Pedram M. Low Power Design Methodologies. Kluwer Academic, US, 1996.
- [4] J. E. Gunn, K. S. Barron, and W. Ruczczyk, “A low-power DSP corebased software radio architecture,” IEEE J. Select. Areas Commun., vol. 17, no. 4, pp. 574–590, 1999.

- [5] N. Tzartzanis and W. C. Athas, "Design and Analysis", of a Low-Power Energy-Recovery Adder", IEEE Journal of Solid State design Proceedings of the IEEE Great Lakes Symposium on VLSI, 1995, pp. 66-69.
- [6] K. P. Parhi, "Fast Low-Energy VLSI Binary Addition", Proceedings of the Int Conference on Computer Design, 1997, pp. 676-684.
- [7] A. M. Shams and M. A. Bayoumi, "A novel high-performance CMOS 1-bit full-adder cell," IEEE Transactions on circuits and systems II: Analog and digital signal processing, vol. 47, no.5, pp. 478-481, May 2000.
- [8] A. M. Shams, T. K. Darwish, and M. Bayoumi, "Performance analysis of low-power 1-bit CMOS full adder cells," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 10, no. 1, pp. 20-29, Feb. 2002.
- [9] Yingtao Jiang, Abdulkarim Al-Sheraidah, Yuke Wang, Edwin Sha, and Jin-Gyun Chung, "A Novel Multiplexer-Based Low-Power Full Adder", IEEE Transactions on Circuits and systems-II, vol. 5, no. 7, July 2004.
- [10] Eng Sue Chew, Myint Wai Phyu, and Wang Ling Goh, "Ultra Low-Power Full-Adder for Biomedical Applications", IEEE International Conference of Electron Devices and Solid-State Circuits, 2009. pp 115 – 1182, 5-27 Dec. 2009
- [11] F.Vasefi and Z.Abid, "Low power n-bit adders and multiplier using lowest-number-of-transistor 1-bit adders, " IEEE Canadian Conference on Electrical and Computer Engineering, pp. 1731-1734, May 2005.
- [12] Jairam S, Madhusudan Rao, Jithendra Srinivas, Parimala Vishwanath, Udayakumar H, Jagdish Rao SoC Center of Excellence, Texas Instruments, India, Clock Gating for Power Optimization in ASIC Design Cycle: Theory & Practice, 2008

AUTHOR'S PROFILE

T.Pardhu was born in Luxettipet village in Adilabad District. He completed B.Tech in MLR Institute of Technology in the stream of Electronics and Communications Engineering in 2011. He has done his Master's degree (M.Tech) in Embedded Systems from Vignan University, Vadlamudi in 2013. He has done Project in Research Centre IMARAT, Hyderabad as Project Intern. He is Working as Assistant Professor in Marri Laxman Reddy Institute of Technology & Management. His interested fields are Digital signal processing, RADAR communications, Embedded systems, implementation of signal processing on applications in FPGA, Low power VLSI.



S.Manusha was born in Bhimavaram. She is studying B.Tech in Marri Laxman Reddy Institute of Technology & Management in the stream of E.C.E. Her interest fields are Digital Signal Processing, Communications, Low power VLSI, Digital Image Processing.



PERFORMANCE COMPARISON OF 4T, 3T AND 3T1D DRAM CELL DESIGN ON 32 NM TECHNOLOGY

Prateek Asthana, Sangeeta Mangesh

JSS Academy of Technical Education, Noida

prateekasthana1989@gmail.com, sangeetamangesh@jssaten.in

ABSTRACT

In this paper average power consumption of dram cell designs have been analyzed for the nano-meter scale memories. Many modern processors use dram for on chip data and program memory. The major contributor of power in dram is the off state leakage current. Improving the power efficiency of a dram cell is critical for the improvement in average power consumption of the overall system. 3T dram cell, 4T dram and 3T1D DRAM cells are designed with the schematic design technique and their average power consumption are compared using TANNER EDA tool .average power consumption, write access time, read access time and retention time of 4T, 3T dram and 3T1D DRAM cell are simulated and compared on 32 nm technology.

KEYWORDS

Low Power, DRAM, 3TDRAM, 4TDRAM, 3T1D DRAM

1. INTRODUCTION

Memories play an essential role in design of any electronics design where storage of data is required. Memories are used to store data and retrieve data when required. Read Only Memory (ROM) and Random Access Memory (RAM) are two types of memories used in modern day architectures. Random Access Memory is of two types Dynamic Random Access Memory (DRAM) and Static Random Access Memory (SRAM). SRAM is static in nature and faster as compared to DRAM. SRAM is expensive and consume less power. SRAM have more transistors per bit of memory. They are mostly used as cache memories. DRAMS on the other hand are dynamic in nature and slower as compared to SRAM. DRAM are expensive and consume more power, they require less transistor per bit of memory. They are mostly used as main memories. DRAM is widely used for main memories in personal and mainframe computers and engineering workstation. DRAM memory cell is used for read and write operation for single bit storage for circuits. A single DRAM cell is capable of storing 1 bit data in the capacitor in the form of charge. Charge of the capacitor decreases with time .Hence refresh signals are used to refresh the data in the capacitor. When a read signal reads the data it refreshes it as well. Many different cell designs exist for modern day DRAM cell. These designs are differentiated by the no. of transistors used in their designing. As the no. of transistors increase, power dissipation also increases. DRAM is one of the most common and cost efficient random access memory used as main memory for workstations. The charge stored in memory cell is time dependent. For high density memories DRAM cell with low power consumption and less area are preferred.

Many different cell designs exist for modern day DRAM cell. These designs are differentiated by the no. of transistors used in their designing. As the no. of transistors increase, power dissipation also increases. DRAM is most common and cost efficient random access memory used as main memory for workstations. The charge stored in memory cell is time dependent. First DRAM was proposed in 1971, having 1 kb capacity. This capacity has increased from 1kb to 1-10 GB level today.

2. LITREATURE SURVEY

2.1. Static random access memory (SRAM)

Structure: SRAM provide static random access memory implementation. Here 6 transistors are used to store bits of data. SRAM 6T form L1 data cache in microprocessor as they have short access time, they are able to retain data for 10's of microsecond. Current level of device miniaturizations makes it very difficult to model 6T SRAM memories with required level of reliability 6T SRAM also suffers from instability which results in performance reduction, which helps in gaining technology scaling. [3]. Process variation directly attaches the weakness of 6T SRAM producing transistors that deviate from their sizes , thereby causing device mismatches .Process variation directly attaches the weakness of 6T SRAM producing transistors that deviate from their sizes ,thereby causing device limits 6T performance scalability by causing variation in operating speed of individual cells and memory lines. [4]

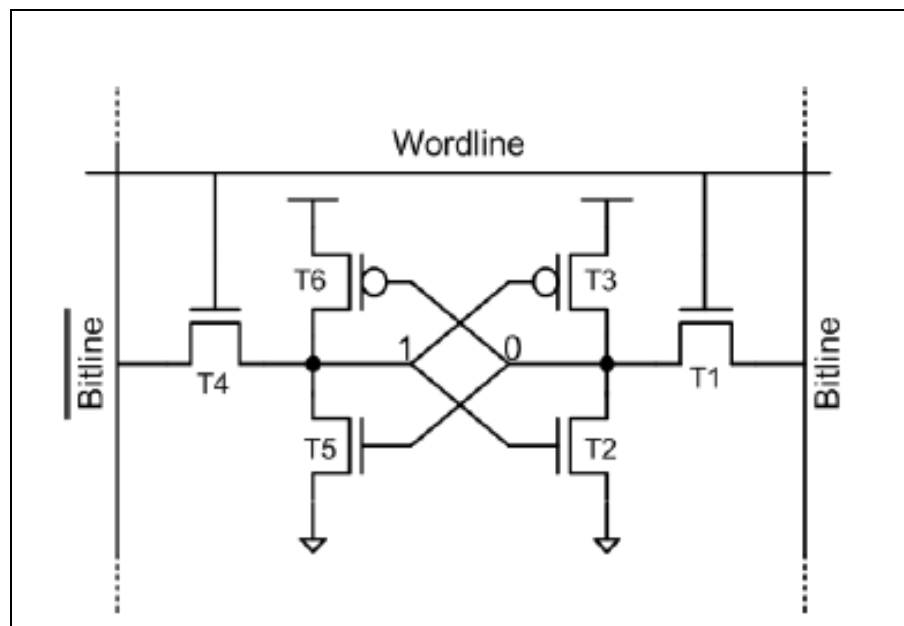


Fig.1 Schematic 6T SRAM

Process variation affects speed of 6T SRAM. Figure shows schematic of standard 6T cell. To perform read operation pre-charging of both bit-lines, strobing word line and seeing which bit-line discharges. If inverted then '1' is read if regular then '0'. Variation in gate length and threshold voltages of these transistors changes current driving capabilities. Process variation also attacks the stability of a 6T SRAM cell. For example, transistor T2 is designed to be very strong, transistor T1, moderately strong, and transistor T3, weak.

In reads, this allows T2 to quickly discharge the necessary bit-line while ensuring the intermediate node between T2 and T3 does not rise enough to store a 1 when it is supposed to store a 0. Any variation within the cell changes the strength of each transistor, and may lead to a weaker T2 that does not discharge the bit-line quickly enough. Such variation allows the value at the intermediate node to rise completely and flip the bit stored in the circuit, causing a pseudo-destructive read. The same analysis holds for transistors T4, T5, and T6. Variation also causes instability in writes. [5]

2.2. 1T1C DRAM Cell:

The information is stored as different charge levels at a capacitor in conventional 1T/1C DRAM. The advantage of using DRAM is that it is structural simple: only one transistor and capacitor are required for storing one bit, compared to six transistors required in SRAM. This allows DRAM to have a very high density. The DRAM industry has advanced over a period of time in packing more and more memory bits per unit area on a silicon die. But, the scaling for the conventional 1Transistor/1Capacitor (1T/1C) DRAM is becoming increasingly difficult, in particular due to a capacitor has become harder to scale, as device geometries shrink. Apart from the problems associated with the scaling of the capacitor, scaling also introduces yet another major problem for the DRAM manufacturers which is the leakage current.

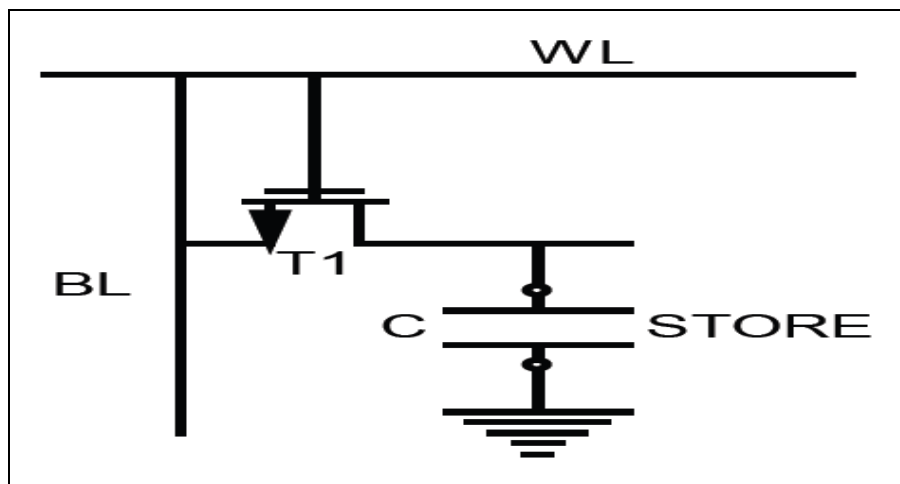


Fig. 2 Schematic 1T1C DRAM cell

2.3. 4T DRAM Cell:

The cell structure shown in fig. 3 is a 4T DRAM cell structure. This DRAM cell design consists of four transistors. One transistor is used as a write transistor, the other as a read transistor. Data in DRAM is stored in the form of charge at the capacitance attached with the transistor structure. There is no current path to the storage node for restoring the data; hence data is lost due to leakage with the period of time. Read operation for the 4T DRAM cell is non-destructive, as the voltage at the storage node is maintained.

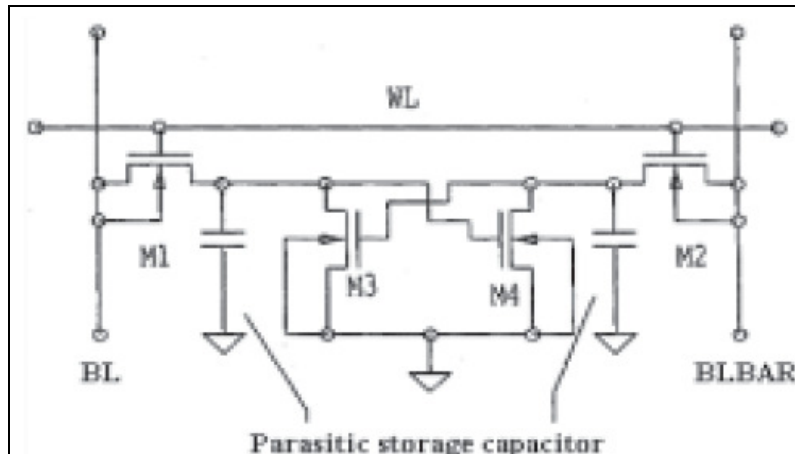


Fig. 3 4T DRAM CELL

2.4. 3T DRAM Cell:

The simplest DRAM cell is the 3T scheme. A 3T DRAM cell has a higher density than a SRAM cell; moreover in a 3T DRAM, there is no constraint on device ratios and the read operation is nondestructive. In this cell, the storage capacitance is the gate capacitance of the readout device, so making this scheme attractive for embedded memory applications; however, a 3T DRAM shows still limited performance and low retention time to severely limit its use in advanced integrated circuits. 3T DRAM utilizes gate of the transistor and a capacitance to store the data value. When data is to be written, write signal is enabled and the data from the bit line is fed into the cell. When data is to be read from the cell, read line is enabled and data is read through the bit line. 3T DRAM cell occupies less area compared to the 4T DRAM cell (fig. 4).

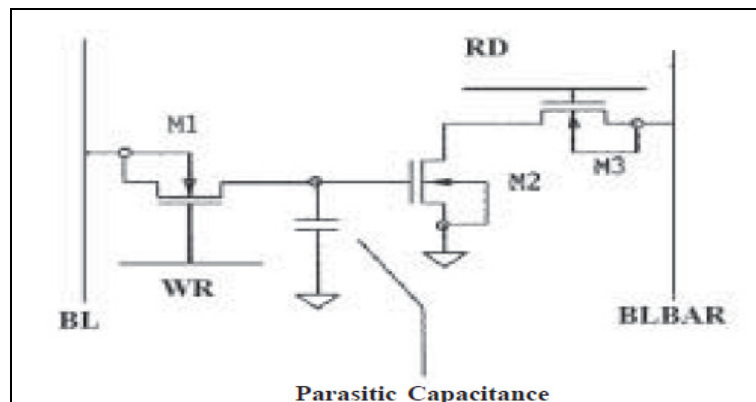


Fig. 4 3T DRAM CELL

2.5. 3T1D DRAM Cell:

This is a DRAM structure derived from 3T cell, like all DRAM it uses few transistor compared to static random access memory (SRAM). The 3T1D has an advantage over SRAM, is its resistance to process variation, this feature helps it to be used at low feature sizes. Another advantage of 3T1D DRAM is that it does not slow down as its size is scaled down. 3T1D DRAM uses the gated

diode instead of capacitor to store the data value. The absence of capacitor provides significance reduction in power consumption as compared to previous DRAM cell design.

In order to write the cell at the BL write line level, it is only required to activate T1 through the WL write line. Hence, the S node stores either a 0 or a $V_{DD}-V_{th}$ voltage depending on the logic value. This voltage results in the accumulation of charge at the gate of devices D1 and T2. [2]

The 3T1D cell in fig. 5 shows the scheme of the basic cell. The basis of the storage system is the charge placed in node S, written from BL write line when T1 is activated. Consequently, it has a DRAM cell nature, but it allows a non-destructive read process (a clear advantage over 1T1C memories) and high performance read and writes operation, comparable to 6T. With T1 and T3 transistors as accessing devices, the whole cell is composed by four transistors of similar size to the corresponding of 6T.

This implies a more compact cell structure. In order to write the cell at the BL write line level it is only required to activate T1 through the WL write line. Hence, the S node stores either a 0 or a $V_{DD}-V_{th}$ voltage depending on the logic value. This voltage results in the accumulation of charge at the gate of devices D1 and T2. A key aspect of the 3T1D memory cell is that the capacitance of the gated diode (D1) when V_{gs} is above V_{th} is significantly higher with respect to lower voltages, because there is a substantial amount of charge stored in the inversion layer.

In order to read the cell, the read bit line BL read has to be previously pre-charged at VDD level. Then T3 is activated from WL read line. If a high (1) level is stored in S, transistor T2 turns on and discharges the bit line. If a low (0) level is stored in S, transistor T2 does not reach enough conduction level. The objective of the gated diode D1 is to improve Read Access Time. When a high (1) level is stored in S, D1 connected to WL read line causes a boosting effect of the voltage level in node S. The voltage level reached at node S is close to Vdd voltage causing a fast discharge of the parasitic capacitance in BL read. If allow (0) level is stored, transistor T2 keeps turned off. [9]

Variability introduces a wide range of effects, especially on the performance of integrated circuits. Some of them appear during the manufacturing process, others during the working life, and all of them have as a consequence a decrease in the circuit reliability. In the case of 3T1D memory cell, all the possible variations can be lumped into a timing degradation. In this sense 3T1D tolerates higher levels of variability than 6T cell, which incurs into timing degradation as well as instability. [1]

Manufacturing process introduces variations in devices characteristic parameters, such as threshold voltage and physical dimensions (length and width) of transistors. These variations can be classified depending on their statistics as systematic (inter die and intra-die systematic) or random (intra-die random). To simulate these effects in a single cell we use a Gaussian distribution. Systematic variability is assumed to be the same for all the transistors in a single cell, while random variability is calculated independently for each transistor. Another key point is that this kind of variation remains static during the whole life of a circuit because it depends only on the manufacturing process. [2]

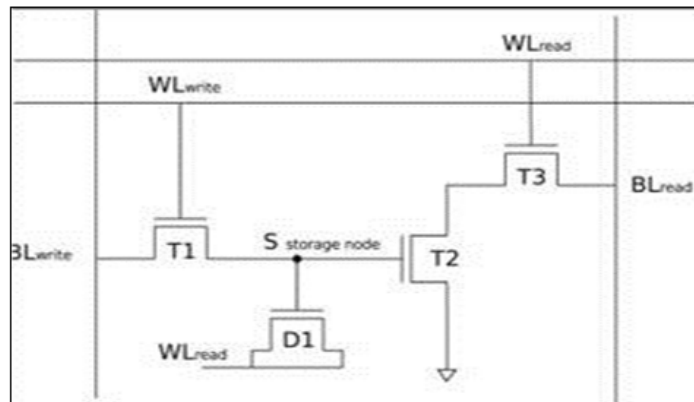


Fig. 5 3T1D DRAM CELL

3. SCHEMATICS OF CELLS

All simulation carried out on TANNER EDA 14.0 with model file of 32nm high performance taken from PTM. Tool used for circuit design is SEDIT and for simulation is TSPICE.

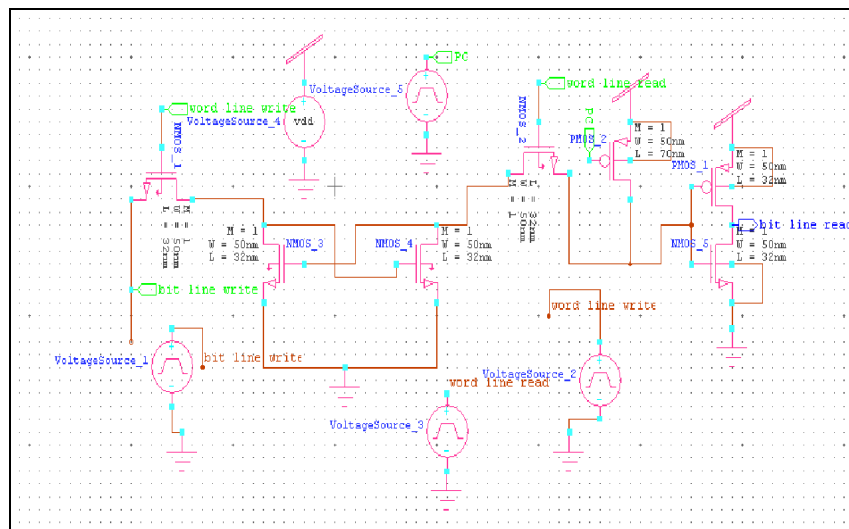


Fig. 6 Schematic of 4T DRAM

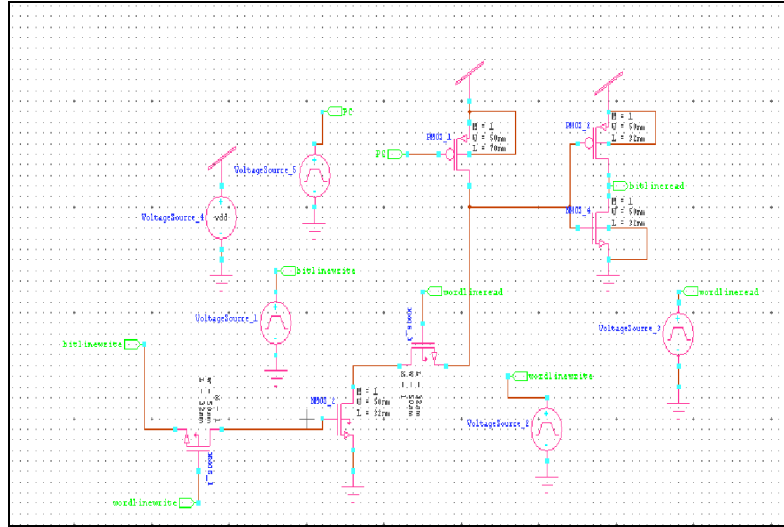


Fig.7 Schematic of 3T DRAM

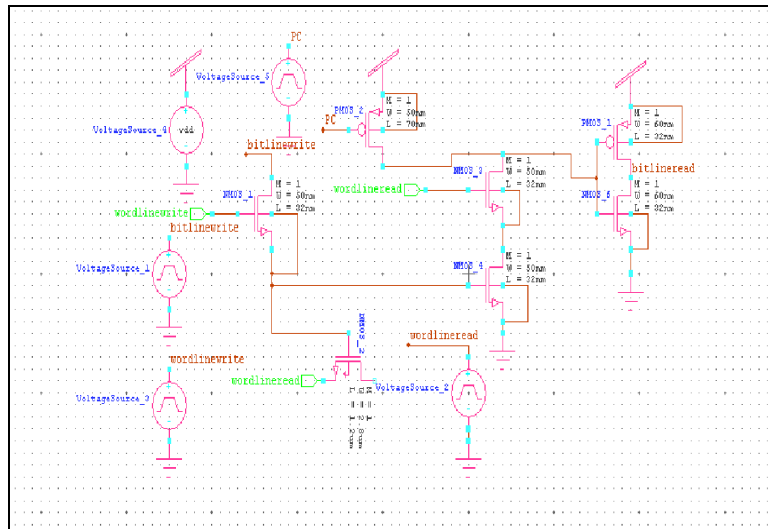


Fig.8 Schematic of 3T1D DRAM

4. SIMULATION RESULTS

Simulation for all five cells 4T, 3T,3T1D ,Gain 3T,Power modified 3T1D design are carried out from 0-10 ns. During this interval all the four process write '0', write '1', read '0' and read '1',are executed. Average power consumption is calculated for the full 0-10ns duration consisting of all four operations.

Table1. Operation of waveforms

Operation	Time Period
WRITE '1'	2-3ns
READ '1'	4-5ns
WRITE '0'	6-7ns
READ '0'	8-9ns

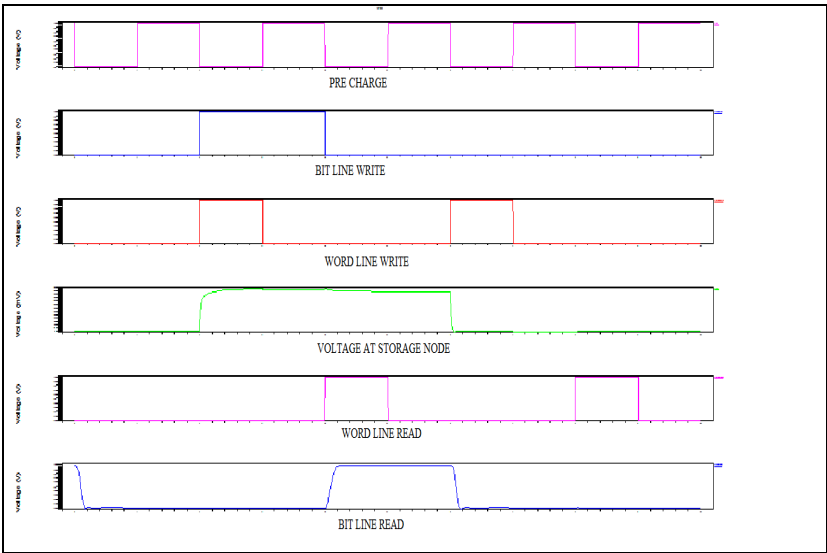


Fig.9 Read Write operation of 4T DRAM Cell

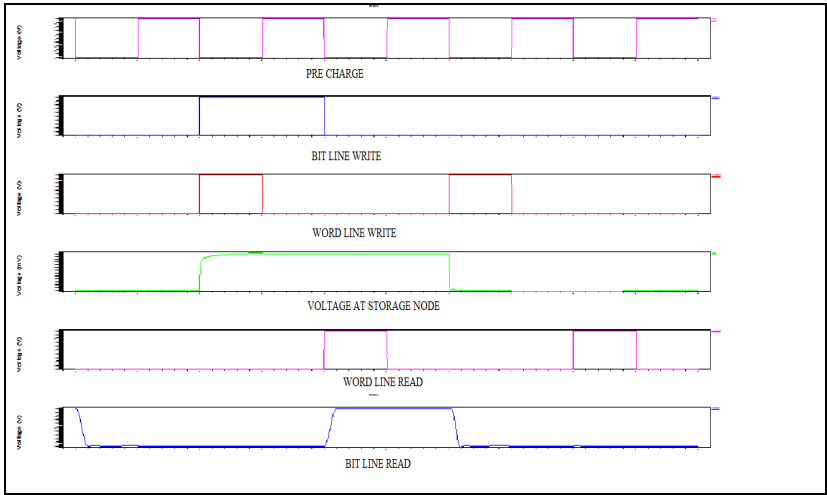


Fig.10 Read Write operation of 3T DRAM Cell

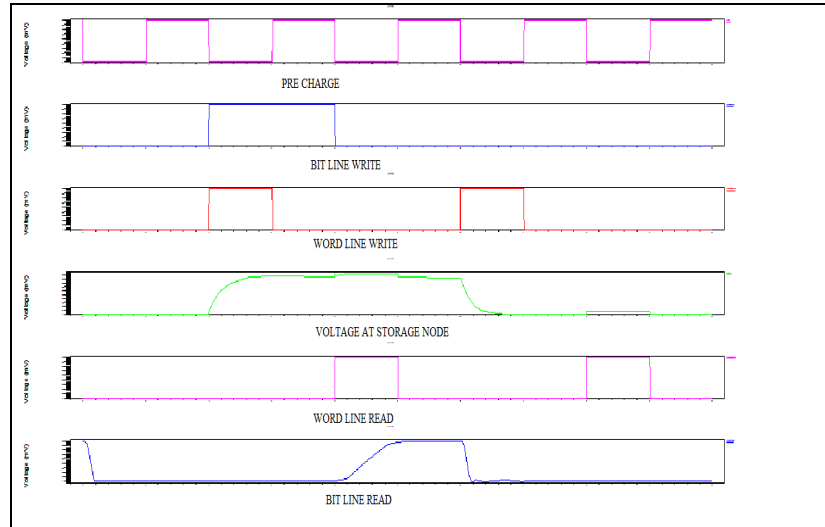


Fig.11 Read Write operation of 3T1D DRAM Cell

5. PERFORMANCE ANALYSIS

Average power consumption for 4T, 3T, 3T1D, and Power modified 3T1D is carried out. The average power consumption value is calculated with varying temperature from 20°C to 100°C. It is essential to perform power v/s temperature as it gives an idea about the average power consumption of the cell design when it is subjected to high temperature.

Table2. Average Power Consumption V/S Supply Voltage

SUPPLY VOLTAGE (volt)	AVERAGE POWER CONSUMPTION FOR 4T(u watt)	AVERAGE POWER CONSUMPTION FOR 3T(u watt)	AVERAGE POWER CONSUMPTION FOR 3T1D(u watt)
0.7	0.1656537	0.2507752	0.1299199
0.8	0.2625748	0.6496773	0.3846698
0.9	0.6618851	1.167894	1.255590
1	1.739659	1.662307	1.566112
1.1	2.384711	2.262632	2.170092

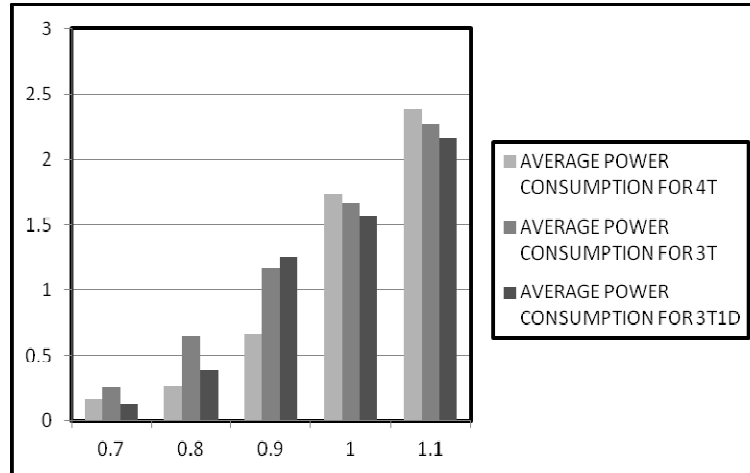


Fig. 12 Bar chart comparing power consumption of dram cell with supply voltage

Table3. Write Access Time V/S Supply Voltage

SUPPLY VOLTAGE (volt)	WRITE ACCESS TIME FOR 4T (p sec)	WRITE ACCESS TIME FOR 3T (p sec)	WRITE ACCESS TIME FOR 3T1D (p sec)
0.7	25.43	16.6	230.09
0.8	27.28	17.66	281.40
0.9	26.96	17.56	291.17
1	35.73	19.53	299.35
1.1	37.45	20.89	385.45

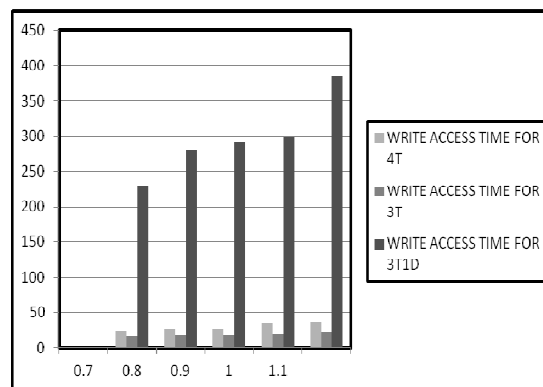


Fig. 13 Bar chart comparing write access time of DRAM cell with supply voltage

Table4. Read Access Time V/S Supply Voltage

SUPPLY VOLTAGE (volt)	READ ACCESS TIME FOR 4T (p sec)	READ ACCESS TIME FOR 3T (p sec)	READ ACCESS TIME FOR 3T1D (p sec)
0.7	----	----	----
0.8	----	----	----
0.9	----	181.86	506.43
1	100.87	87.72	103.40
1.1	71.19	70.7	68.65

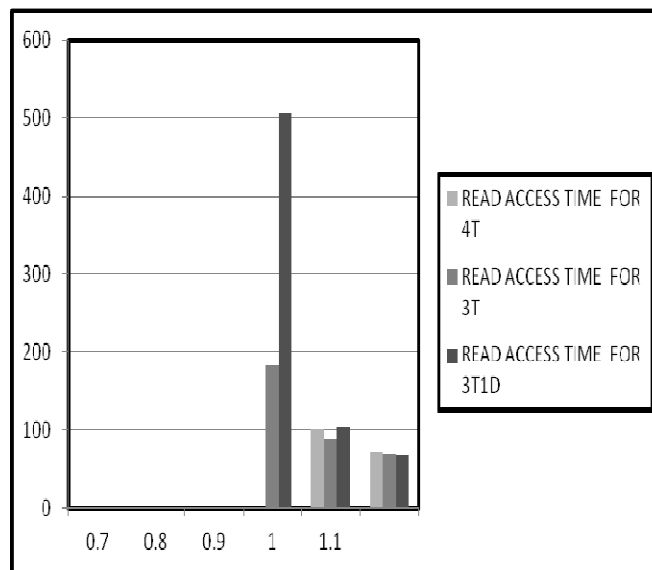


Fig. 14 Bar chart comparing read access time of DRAM cell with supply voltage

Table5. Retention Time V/S Supply Voltage

SUPPLY VOLTAGE (volt)	RETENTION TIME FOR 4T (u sec)	RETENTION TIME FOR 3T (u sec)	RETENTION TIME FOR 3T1D (u sec)
0.7	4.54378	5.31742	66.44272
0.8	4.02719	4.50587	59.30578
0.9	3.09407	3.44944	52.44869
1	2.61674	3.74946	45.55423
1.1	2.45621	3.49827	40.07223

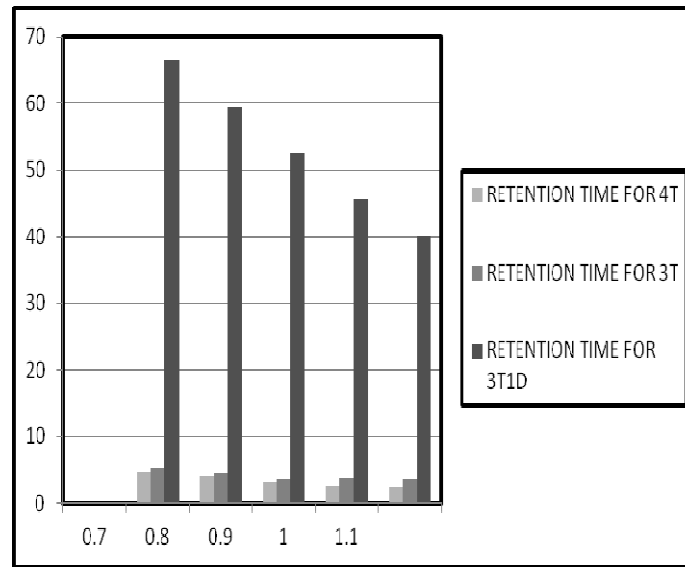


Fig. 15 Bar chart comparing retention time of dram cell with supply voltage

6. CONCLUSION

The study of 4T DRAM cell, 3T DRAM cell and 3T1D DRAM cell for average power consumption, write access time, read access time and retention time has been carried out. These parameters are studied in accordance with variation of supply voltage FROM 0.7, 0.8, 0.9, 1.0, 1.1V.

Analysis of Average Power Consumption shows that 3T1D DRAM cell has the least average power consumption compared to 4T DRAM and 3T DRAM cell. Average Power Consumption tends to increase as the supply voltage increases.

Analysis of Write Access Time shows that 3T DRAM cell has the least write access time compared to 4T DRAM and 3T1D DRAM cell. Write Access Time of 3T1D DRAM is significantly more than that of 3T and 4T DRAM cell. Write Access Time tends to increase as the supply voltage increases.

Analysis of Read Access Time shows that 3T DRAM cell has the least read access time compared to 4T DRAM and 3T1D DRAM cell. Read Access Time of 3T1D DRAM is significantly more than that of 3T and 4T DRAM cell. Read Access Time tends to decrease as the supply voltage increases.

The most significant parameter for DRAM cell is retention time. 4T DRAM cell has the least retention time among the three cells. The retention time for 3T1D DRAM cell is significantly more than that of 4T and 3T DRAM cell. Retention time tends to decrease as the supply voltage increases.

REFERENCES

- [1] M. S. B. S. Shyam Akashe, "Analysis of power in 3T DRAM and 4T DRAM Cell design for different Technology," IEEE, vol. 12, no. 978-1-4673-4805-8, pp. 18-21, 2012.
- [2] J. C. H. D. J. V. K. Wing k.luk, "A3-Transistor DRAM Cell with Gated Diode for enhanced Speed and Retention Time," Symposium on VLSI CircuitsDigest of Technical papers, vol. 06, no. 1-4244-0006-6, 2006 IEEE.
- [3] j.-W. C. a. Y. C. Weijie Cheng, "Design of logic compatible Embedded DRAM using Gain Memory Cell," ISOC, vol. 12, no. 978-1-4673-2990-3, pp. 196-199, 2012 IEEE.
- [4] H. K. L. H. H. S. J. a. Y. R. Myungjae Lee, "Analysis of Dynamic Retention Characteristics of Nwl Scheme in High Density DRAM," IPFA, vol. 13, no. 978-1-4799-0480-8, pp. 641-644, 2013 IEEE.
- [5] D. B. S. Z. Yong Sung Park, "Low Power High threshold LPDC decoder using non-refresh Embedded DRAM," JSSC, vol. 49, no. 0018-9200, pp. 1-12, 2014 IEEE.
- [6] P.-T. H. a. W. H. Mu-Tien Chang, "A 65nm Low Power 2T1D Embedded DRAM With leakage Current Reduction," IEEE, vol. 07, no. 978-1-4224-1593-9, pp. 207-210, 2007.
- [7] S. G. R. C. Nivard Asymerich, "Impact of Positive bias temperature instability on 3T1D DRAM cells," Integration , the VLSI Journal, vol. 45, no. 2011 Elsevier, pp. 246-252, 2011.
- [8] C. L. X. L. D. B. G.-Y. W. Kristen Lovin, "Emperical Performance Models for 3T1D Memories," IEEE, vol. 09, no. 978-1-4224-5028-2, pp. 398-403, 2009.
- [9] E. Amat, "Strategies to enhance 3T1D DRAM cell variability robustness," Microelectronics Journal, no. 2013 Elsevier, pp. 1-6, 2013.
- [10] Tutorial TANNER EDA LEDIT and TSPICE

AUTHORS

Mr. Prateek Asthana is currently pursuing M.tech in the field of advanced ECE with specialization in VLSI design. He has done his B.tech from AmityUniversity, Noida in the field of electronics and communication engineering in the year 2011. His research area includes VLSI design, low power VLSI design and Memories.



Mrs. Sangeeta Mangesh is assistant professor at JSS Academy of Technical Education, Noida. She is the guide and supervisor to the above research work. Her research area includes low power VLSI design, VLSI design and nanotechnology



INTENTIONAL BLANK

A HIGHLY ADAPTIVE OPERATIONAL AMPLIFIER WITH RECYCLING FOLDED CASCODE TOPOLOGY

Saumya Vij¹, Anu Gupta² and Alok Mittal³

^{1,2}Electrical and Electronics Engineering, BITS-Pilani, Pilani, Rajasthan, India

³High Speed Links, STMicroelectronics, Greater Noida

¹f2009587@pilani.bits-pilani.ac.in, ²anug@pilani.bits-pilani.ac.in

³alokkumar.mittal@st.com

ABSTRACT

This paper presents a highly adaptive operational amplifier with high gain, high bandwidth, high speed and low power consumption. By adopting the recycling folded cascode topology along with an adaptive-biasing circuit, this design achieves high performance in terms of gain-bandwidth product (GBW) and slew rate (SR). This single stage op-amp has been designed in 0.18 μ m technology with a power supply of 1.8V and a 5pF load. The simulation results show that the amplifier achieved a GBW of 335.5MHz, Unity Gain Bandwidth of 247.1MHz and a slew rate of 92.8V/ μ s.

KEYWORDS

Recycling Folded Cascode, Operational Amplifier, slew rate, Adaptive biasing, Transconductance

1. INTRODUCTION

In high performance analog integrated circuits, such as switch-capacitor filters, delta-sigma modulators and pipeline A/D converters, op amps with very high dc gain and high unity-gain frequency are needed to meet both accuracy and fast settling requirements of the systems. However, as CMOS design scales into low-power, low-voltage and short-channel CMOS process regime, satisfying both of these aspects leads to contradictory demands, and becomes more and more difficult, since the intrinsic gain of the devices is limited. [1]

In order to achieve high-gain, the folded cascode amplifier is often adopted as the first-stage of two-stage amplifiers. Actually, in the deep-submicron CMOS technology, high-gain amplifiers are difficult to be implemented because of the inherent low intrinsic gain of the standard threshold voltage MOS transistors. At the same time, because of the reliability reasons in the deep-submicron processes, the output swing of amplifier is severally restricted with the lower power supply voltage. [2]

To efficiently increase operational amplifier's gain and output swing, multi-stage fully-differential operational amplifier topology is appreciated. The operational amplifier with three or even more stages equipped with the Nested-Miller compensation or the Reversed Nested-Miller

compensation shows high efficiency in the gain enhancement, while they require additional large compensation capacitors compared to the traditional two-stage operational amplifier, which will lead to a larger die area and the limited slew rate. Besides, additional common mode feedback (CMFB) circuits would consume additional power. [3]

This paper presents a novel idea of implementing recycling folded cascode [4] along with an adaptive-biasing circuit [5] to achieve high gain, high bandwidth and high slew rate specifications. Section 2 describes the proposed design. Section 3 analyzes the design and working of the circuit. Implementation is discussed in section 4, simulations in section 5, followed by conclusion in section 6.

2. PROPOSED STRUCTURE

The proposed design presented in this paper employs the recycling folded cascode along with an adaptive bias current circuit. This single stage operational amplifier is capable of providing high gain of around 70dB along with a high bandwidth of 250 MHz and a slew rate of around 100V/ μ s which is approximately twice as that of the recycling folded cascode without the additional adaptive-biasing circuit.

Recycling folded cascode is basically a modified folded cascode where the load transistor also acts as a driving transistor, hence, enhancing the current carrying capability of the circuit. Recycling folded cascode is obtained by splitting the input transistors and the load transistors as given in figure 1. The cross-over connections of these current mirrors ensure that the small signal currents are added at the sources of M1, M2, M3 and M4 and are in phase.

This is called as recycling folded cascode (RFC), as it reuses/recycles the existing devices and currents to perform an additional task of increasing the current driving capability of the circuit. The proposed modification in the recycling folded cascode topology involves replacing the transistor M0 with an adaptive-biasing circuit (figure 1) [5] which further enhances the current driving capability of this circuit and hence the speed.

3. ANALYSIS AND DESIGN OF THE PROPOSED STRUCTURE

3.1 Low Frequency Gain

The open loop gain of an operational amplifier determines the precision of the feedback systems employing it. A high open loop gain is a necessity to suppress linearity [6]. The low frequency gain of OTAs is frequently expressed as the product of the small signal transconductance, G_m and the low frequency output impedance, R_o . The low frequency gain of the adaptive recycling folded cascode is almost the same as that of the recycling folded cascode topology, i.e.

$$R_{O_{ARFC}} \approx g_{m_{16}} r_{o_{16}} (r_{o_4} \parallel r_{o_{10}}) \parallel g_{m_{14}} r_{o_{14}} r_{o_{12}} \quad (1)$$

$$G_{m_{ARFC}} \approx G_{m_{RFC}} (=g_{m_1}(1+K)) \quad \text{where } K=3 \quad (2)$$

Both the RFC and adaptive RFC (ARFC) have similar noise injection gains from either supply. Although there is no discernable change in low frequency gain but extended bandwidth of the adaptive RFC ensures high GBW. Moreover, the extended GBW of the adaptive RFC extends the improved PSRR performance to higher frequencies than the RFC.

3.2 Phase Margin

The phase margin is often viewed as a good indicator to the transient response of an amplifier, and is determined by the poles and zeros of the amplifier transfer function. The adaptive RFC shares a dominant pole ω_{p1} , determined by the output impedance and capacitive load and a non-dominant pole ω_{p2} , determined by the parasitic at the source of M15/M16. It has a pole-zero pair, ω_{p3} and ω_{pz} ($= (K+1) \omega_{p3}$), associated with the current mirrors M7:M8 and M9:M10. However, this pole-zero pair is associated with NMOS devices, which puts it at a high frequency. In addition, adaptive RFC also have a pole due to adaptive current source, ω_{p4} . Due

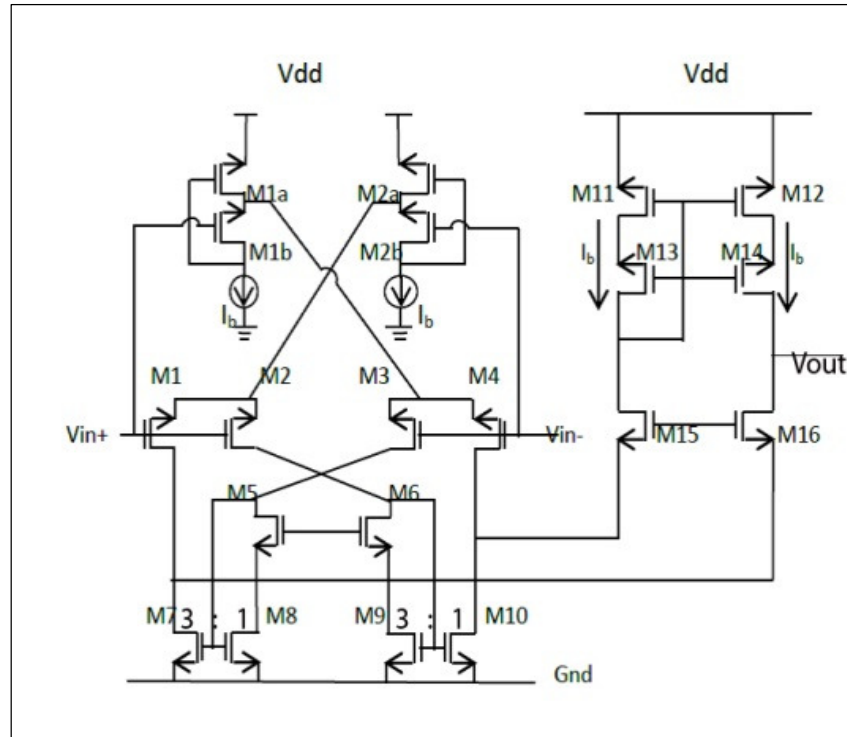


Figure1. Schematic of the proposed design

to low impedance at that node it is pushed to a high frequency. The pole-zero values from the PZ analysis in cadence virtuoso have been tabulated in Table I. Also, their positioning with respect to each other is shown in figure 2.

Table1. Pole/Zero Analysis

Pole/Zero	Real Value
ω_{p1}	-1.267e+05
ω_{p2}	-3.551e+08
ω_{p3}	-5.324e+08
ω_{p4}	-9.908e+08
ω_z	-21.296e+08

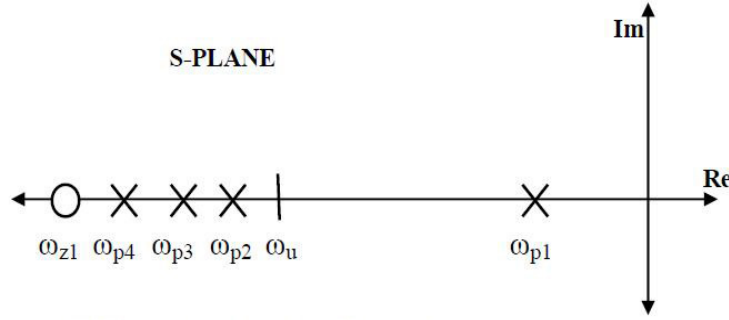


Figure 2. Pole-zero analysis of the proposed design

3.3 Slew Rate

Slew rate is one of the most critical design aspects especially for the kind of circuits where high speed is necessary. To achieve a high slew rate, adaptive biasing circuit plays a vital role. The upper part of the proposed design [5] that is the adaptive biasing circuit consists of four matched transistors M_1 , M_2 , M_3 and M_4 cross-coupled by two dc level shifters. Each level shifter is built using two transistors (M_{1a} , M_{2a} and M_{1b} , M_{2b}) and a current source. These level shifters are called Flipped Voltage Followers (FVFs). The dc level shifters must be able to source large currents when the circuit is charging or discharging a large load capacitance. Moreover, they should be simple due to noise, speed, and supply constraints.

Analysis of the proposed design shows that there is a significant improvement in its slew rate over the RFC topology. Suppose V_{in+} goes high, it follows that M_1 and M_2 turn off, which forces M_9 and M_{10} to turn off. Consequently, the drain voltage of M_9 rises and M_{16} is turned off whereas M_3 is driven into deep triode. This directs current I_d into M_4 and in turn is mirrored by a factor of 3(K) (M_7 , M_8) into M_{15} , and again by a factor of 1 into (M_{11} , M_{12}). For simplicity, if we ignore any parasitic capacitance at the sources of $M_{1,2,3,4}$ and follow the similar derivation steps but assuming V_{in+} goes low, the result is symmetric slew rate expressed in (3)

$$SR \text{ (adaptive)}_{RFC} = 6I_d/C_L \text{ [4]} \quad (3)$$

$$\text{We know that,} \quad I_d = I_D + i_d \quad (4)$$

Due to presence of the adaptive biasing circuit, this circuit changes current according to the input voltage and hence remains self-biased. It also causes minimal increase in power dissipation as the current only increase proportional to the voltage in one branch and correspondingly decreases in the other one.

Since the ac input signal is applied to both the gate and the source terminals of $M_{1,2}$ and $M_{3,4}$, the transconductance of this input stage is twice as that of a conventional differential pair.

The ac small-signal differential current of the input stage is

$$I_d = i_1 - i_2 \approx (1 + (g_{m2A,B} r_{oA,B} - 1)/(g_{m2A,B} r_{oA,B} + 1)) \quad (5)$$

Clearly ac small signal current is twice as that in the case of RFC without adaptive biasing circuit. Hence, Slew rate has improved in the proposed circuit.

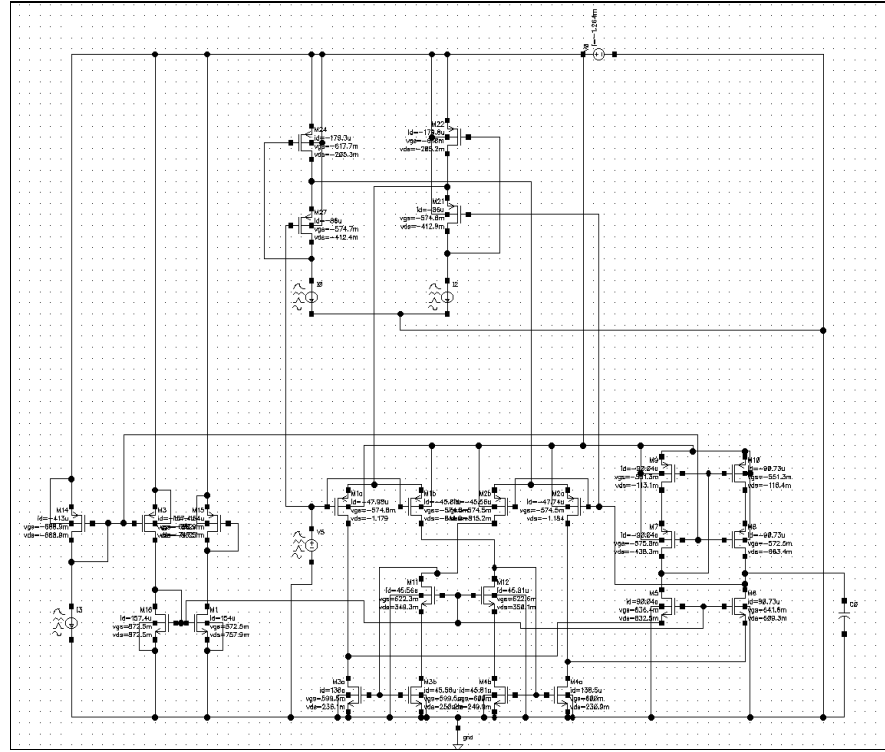


Figure 3. Snapshot from Virtuoso of Proposed Design Schematic

4. IMPLEMENTATION

To validate the theoretical results, we first implemented the recycling folded cascode topology as a benchmark for comparison with our proposed design. And then we simulated our own design and compared the results with our implementation of the RFC. Table II details the transistor sizes used in the implementation of the proposed structure as well as of our RFC implementation.

Table 2. Device sizes in implementation

Device	Proposed design	RFC
$M_{o[4]}$	-	$60\mu\text{m}/500\text{nm}$
M_{1a}, M_{1b}	$100\mu\text{m}/500\text{nm}$	-
M_{2a}, M_{2b}	$128\mu\text{m}/360\text{nm}$	-
M_1, M_2, M_3, M_4	$64\mu\text{m}/360\text{nm}$	$64\mu\text{m}/360\text{nm}$
M_{11}, M_{12}	$64\mu\text{m}/360\text{nm}$	$70\mu\text{m}/500\text{nm}$
M_{13}, M_{14}	$64\mu\text{m}/360\text{nm}$	$84\mu\text{m}/500\text{nm}$
M_5, M_6	$8\mu\text{m}/180\text{nm}$	$8\mu\text{m}/180\text{nm}$
M_{15}, M_{16}	$10\mu\text{m}/180\text{nm}$	$10\mu\text{m}/180\text{nm}$
M_7, M_{10}	$24\mu\text{m}/500\text{nm}$	$24\mu\text{m}/500\text{nm}$
M_8, M_9	$8\mu\text{m}/500\text{nm}$	$8\mu\text{m}/500\text{nm}$

5. SIMULATION RESULTS

All the simulations were done on cadence virtuoso with 0.18 μm technology using a VDD of 1.8V. The load capacitance was taken to be 5.6pF for all the simulations.

Here is the procedure for all the simulations. First of all DC analysis was done to ensure saturation for all transistors. After that, the AC analysis with differential input signal as 1VPP was done to measure the gain, GBW, UGB and Phase margin. After the AC analysis, a transient analysis was done to measure the slew rate and settling time (1%). For the transient analysis, the input signal was given as a square pulse (as shown in figure 9) of amplitude 1V at 5MHz. The results of the simulations are tabulated in Table III and Table IV. Table V details the bias currents in all the transistors of the proposed structure implementation.

Table 3. Results comparison with RFC Implementation

Parameters	Proposed structure (tt)	RFC simulation
DC Gain(dB)	68.48	71
UGB(MHz)	247.1	153
GBW(MHz)	335.5	172.26
Slew rate(V/ μs)	92.8	67.4
Settling time (1%)(ns)	12.39	21.93
Phase Margin	26.3°	58.1°
Power Dissipation(mW)	2.493	2.18
I(total) (mA)	1.385	1.215
Capacitive load	5.6 pF	5.6 pF
Technology	0.18 μm	0.18 μm

Table 4. Result of proposed design at extreme corners

Parameters	Tt	Ff	ss
DC Gain(dB)	68.48	63.83	66.3
UGB(MHz)	247.1	267.6	203.9
GBW(MHz)	335.5	352	280.27
Slew rate(V/ μs)	92.8	134.4	71.4
Settling time (1%)(ns)	12.39	8.9	17.25
Phase Margin	26.3°	34.9°	25.2°
Power Dissipation(mW)	2.493	3.334	2.049
I(total) (mA)	1.385	1.684	1.265
Capacitive load	5.6 pF	5.6 pF	5.6pF
Technology	0.18 μm	0.18 μm	0.18 μm

Table 5. Bias Current in Proposed Structure

Device	$I_{bias}(\mu A)$ (tt)
M_{1a}, M_{2a}	181.1
M_{1b}, M_{2b}	86
M_1, M_4	48.79
M_2, M_3	46.29
$M_{11}, M_{12}, M_{13}, M_{14}, M_{15}, M_{16}$	90.63
M_5, M_6	46.29
M_7, M_{10}	139.4
M_8, M_9	46.29

The UGB of the proposed design is 247.1MHz while for RFC it is 153MHz showing a significant increase in bandwidth as expected. The GBW has also increased from 172.26 MHz for RFC to 335.5 MHz for the proposed design. As proved theoretically, the slew rate has improved from 67.4V/ μ s to 92.8V/ μ s. Also, correspondingly, the settling time (1%) has decreased from 21.93 ns to 12.39 ns showing an increase in the speed of the circuit significantly. Although the phase margin has reduced but it can be dealt with by using a compensation capacitance when a second stage op Amp, which will cause serious issue. Hence RC compensation is a better choice, as it will allow moving the zero away or forcing it in LHP. The most impressive aspect of this design is the fact the increased speed and bandwidth is achieved with nearly the same power dissipation as the RFC. The circuit has been implemented on all corners with all transistors in the saturation state. Table III demonstrates the simulation results of the circuit in all corners i.e. tt, ss and ff.

Figure 4 shows the linear settling time response plotted during the transient analysis which was used for the slew rate and settling time calculations. The open loop AC response of the amplifier in tt, ff and ss corners is shown in figures 5, 6 and 7 respectively.

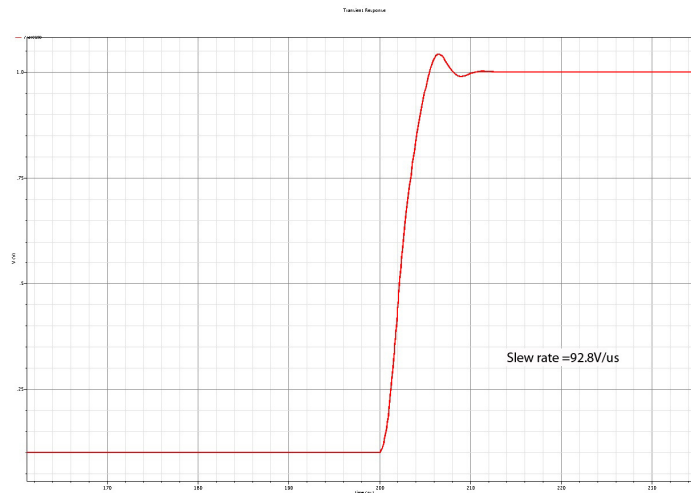


Figure 4. Graph for calculating rate slew

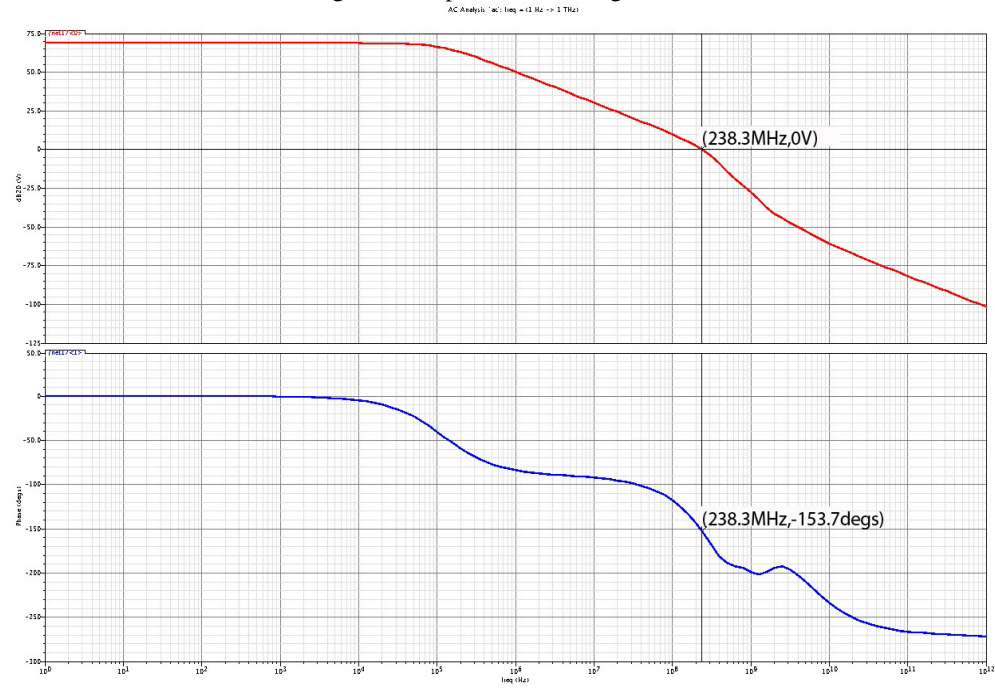


Figure 5. Gain & Phase plot for tt case

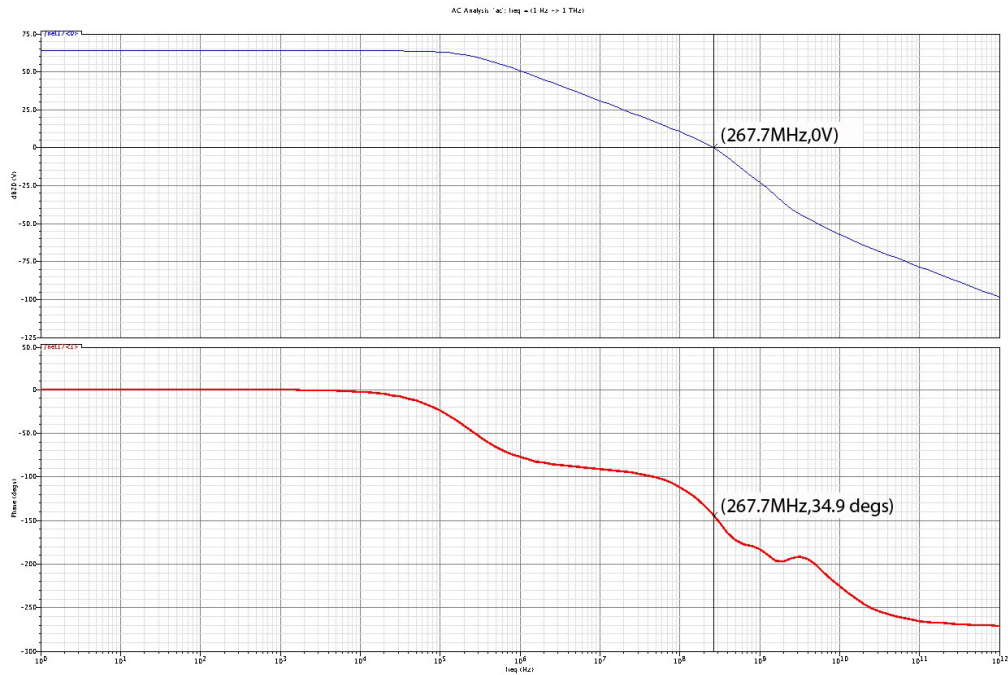


Figure 6. Gain & Phase plot for ff corner

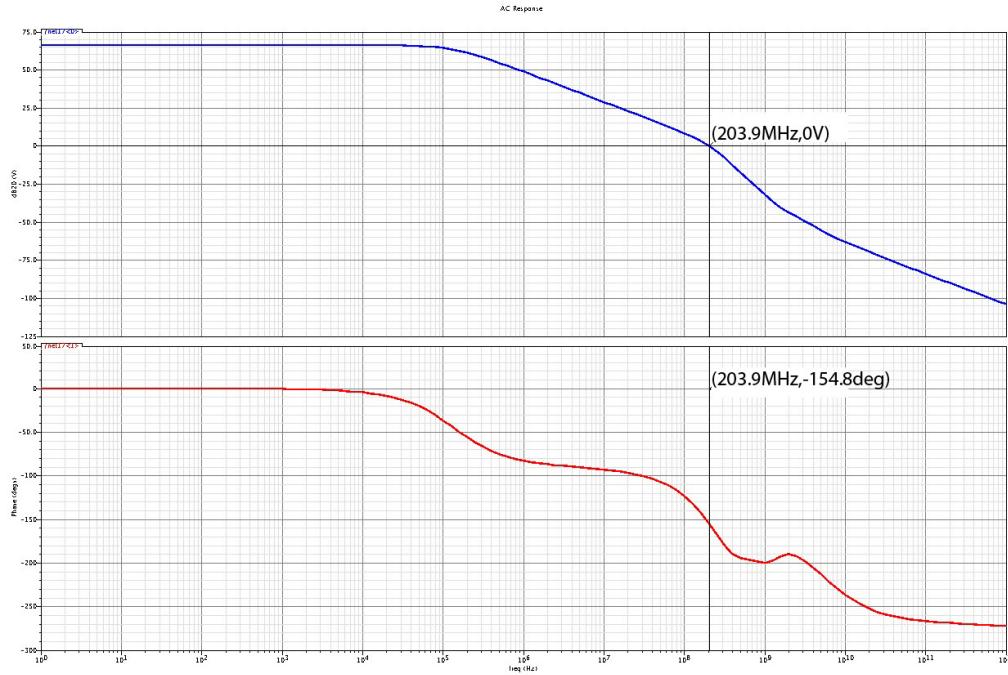


Figure 7. Gain & Phase plot for ss corner

5.1 Op-amp as a voltage follower

The proposed design was implemented with a negative feedback in a voltage follower configuration (shown in figure 8) to test the stability of the design. An input pulse of 1V was given at 5MHz to check its response and functioning. Figure 9 below shows the input and output pulses in a voltage follower configuration. It is evident from the output graph that the delay introduced by the voltage follower is very small. Also, a distortion less and non-sluggish output is achieved as a result of high slew rate and bandwidth provided by the ARFC.

Due to high slew rate and bandwidth characteristics, ARFC finds application in various other speed critical circuits such as switched capacitor circuits, comparators etc.

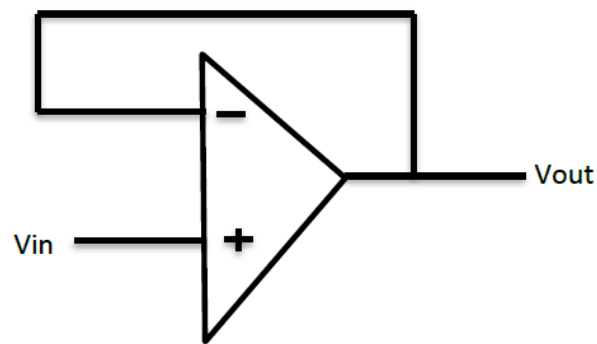


Figure 8. Voltage follower

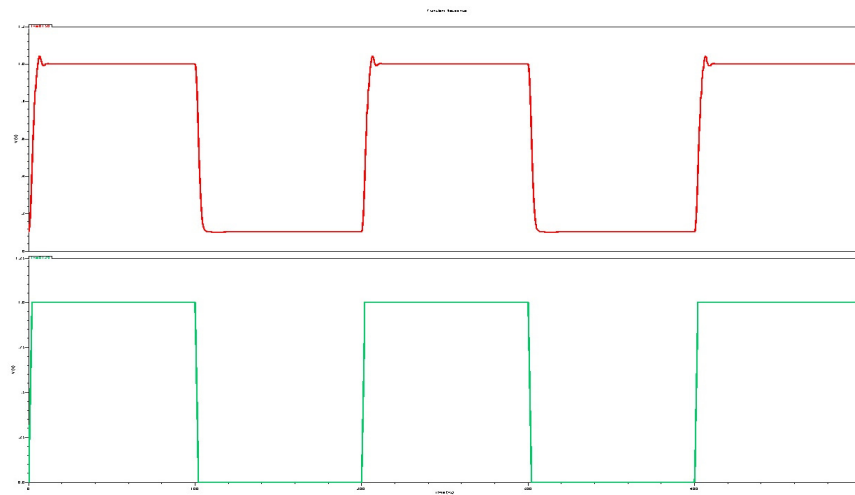


Figure 9. Transient Response in a voltage follower

6. CONCLUSION

It has been demonstrated that the proposed design shows a significant improvement over the conventional RFC in terms of UGB, GBW and slew rate with nearly the same power consumption. The additional adaptive biasing circuit added to the RFC, not only improves its speed and frequency response but also makes the circuit very adaptive to the changes in input voltage and noise fluctuations. With the RFC itself having an adaptive load, this addition of a self-adjusting current source makes it a very flexible, adaptive and self-biased circuit. This feature of the circuit also helps reducing the power consumption by changing currents corresponding to the changes in the input voltage. The theoretical results were confirmed with good agreement with the simulation data.

ACKNOWLEDGEMENT

The authors would like to take this opportunity to thank BITS Pilani, Pilani Campus Administration for providing them with the facilities and resources, which were required to conduct the research for this paper.

REFERENCES

- [1] SU Li QIU Yulin, "Design of a Fully Differential Gain-Boosted Folded-Cascode Op Amp with Settling Performance Optimization" IEEE Conference Electronic Devices and Solid-State Circuits, pp. 441 – 444, Dec 2005.
- [2] Zhou Qianneng', Li Hongjuan2, Duan Xiaozhong', and Yang Chong', "A Two-Stage Amplifier with the Recycling Folded Cascode Input-Stage and Feedforward Stage" Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC), vol. 2, , pp. 1557 – 1560, July 2011.
- [3] Hong Chen, Vladimir Milovanovic, Horst Zimmermann "A High Speed Two-Stage Dual-Path Operational Amplifier in 40nm Digital CMOS" Mixed Design of Integrated Circuits and Systems (MIXDES) conference, pp. 198-202, May 2012
- [4] Rida S. Assaad, Student Member, IEEE, and Jose Silva-Martinez, Senior Member, IEEE "The Recycling Folded Cascode: A General Enhancement of the Folded Cascode Amplifier" IEEE J. solid-state circuits, vol. 44, no. 9, pp. 2535 - 2542 September 2009.

- [5] Antonio J. López-Martín, Member, IEEE, Sushmita Baswa, Jaime Ramirez-Angulo, Fellow, IEEE, and Ramón González Carvajal, Senior Member, IEEE “Low-Voltage Super Class AB CMOS OTA Cells With Very High Slew Rate and Power Efficiency” IEEE J. solid-state circuits, vol. 40, no. 5, pp. 1068-1077, May 2009
- [6] B. Razavi, Design of Analog CMOS Integrated Circuit. New York: McGraw-Hill, pp. 291-333, 2001.
- [7] R. Assaad and J. Silva-Martinez, “Enhancing general performance of folded cascode amplifier by recycling current,” IEE Electron. Lett., vol. 43, no. 23, Nov. 2007.
- [8] P. E. Allen and D. R. Holberg, CMOS Analog Circuit Design., 2nd ed. Oxford, U.K.: 2002.
- [9] D. Johns and K. Martin, Analog Integrated Circuit Design. New York: Wiley, 1997, pp. 210–213.

AUTHORS

Saumya Vij

2014 Graduate in B.E.(Hons.) Electrical and Electronics Engineering and MSc(Hons.) Economics. Recently started working as an ASIC Design Engineer at NVidia Pvt. Ltd., Bangalore



Anu gupta

Presently working as Associate Professor in the Electrical and Electronics Engineering department of BITS, Pilani. Holds a post graduate degree in Physics from Delhi University, which was followed up with M.E in Microelectronics from BITS, Pilani. In March 2003, she obtained her PhD from BITS, Pilani, Rajasthan.



Alok Mittal

2013 Graduate in B.E(Hons.) Electrical and Electronics from BITS, Pilani. Currently working as Analog Front End design engineer at ST Microelectronics in High speed Links, NOIDA.



INTENTIONAL BLANK

TRANSFERRING OF INFORMATION IN WIRELESS ADHOC SENSOR NETWORK USING SHORTEST PATH ALGORITHM

N. Pushpalatha¹, Dr.B.Anuradha²

¹Assistant Professor, Department of ECE, AITS, Tirupathi
pushpalatha_nainaru@rediffmail.com

²Associate Professor, Department of ECE,S.V. University College of Engineering,Tirupathi.
anubhuma@yahoo.com

ABSTRACT

Wireless sensor networks discover probable in military, environments, health and commercial applications. The process of transferring of information from a remote sensor node to other nodes in a network holds importance for such applications. Various constraints such as limited computation, storage and power makes the process of transferring of information routing interesting and has opened new arenas for researchers. The fundamental problem in sensor networks states the significance and routing of information through a real path as path length decides some basic performance parameters for sensor networks. This paper strongly focuses on a shortest path algorithm for wireless adhoc networks. The simulations are performed on NS2 and the results obtained discuss the role of transferring of information through a shortest path.

KEYWORDS

Sensor Node, Shortest Path algorithm, WSNs (Wireless Sensor Networks), Radio Range.

1. INTRODUCTION

Wireless adhoc sensors are being used in various applications and have gained curiosity as well as importance during the last decades. Wireless adhoc sensor network consists of a number of sensors increase across a geographical area, each sensor has wireless communication ability and some level of intelligence for signal processing and networking of the information. Some examples of wireless ad hoc sensor networks also includes military sensor networks (MSN) and wireless observation sensor networks (WSSN). Specific applications like object tracking, vehicle monitoring and forest fire detection rely totally on adhoc networks, since there use, design and exploitation is fixed requiring great amount of stability. Therefore two ways to classify wireless adhoc sensor networks are, (a) whether or not the nodes are independently addressable (b) whether the information in the network is aggregated. The sensor nodes in a parking lot network should be independently addressable, so that one can trace the entire object. In some applications broadcasting of a message is required within all the nodes. Therefore each node in the network is liable and its priority of node placement also becomes important [10]. The above cited theory reflects an important requirement for adhoc networks to ensure that the required data is scattered

to proper end users through a genuine and shortest path. The work proposed in this paper shows that adhoc networks can be easily managed and configured for specific use if the routing path is shortest. Sensors can schedule their role more accurately and in time if the connecting path is shortest. The reduced path length also improves localization and power consumption for self powered sensor nodes within adhoc networks.

Wireless Sensor Networks (WSNs) consists of mobile wireless nodes communicating without the support of any pre-existing fixed communications. Such great WSNs offer vast application perspectives. Sensors are small devices with hardware constraints (low memory storage and low computational resources) that rely on battery. Sensor Networks thus require energy efficient algorithms to make them work properly in a way that their hardware features and application requirements. A low power sensor node has limited transmission power and this can communicate only to limited number of nodes, called its neighbourhood. Multi-hop communications are used to route data from source to destination [8].

2. RELATED WORKS

The current interests in sensor networks has led to a number of routing schemes that use limited resources available for sensor nodes to effectively find and resolve to a shortest path for power optimization and an efficient information forwarding scheme. Some of the existing shortest path algorithms are discussed as follows. Many research and performance studies have been made on evaluation and energy utilization. However, there has been only modest research on how the network topology impacts WSN performances. Most of the research on the topic rather focuses on how to efficiently place node on a field to achieve the best performances for given algorithm [11].

2.1 DV (Distance Vector) Hop localization algorithm

In multihop propagation the distance between two or more than two hops is calculated using conventional DV-Hop algorithm [2]. In a sensor network each node whether it is a beacon node or an anchor node as a hop count. The information is processed from one node to another through a hop path, if higher is the hop count of a sink node the information becomes unusable more early, therefore only a minimum level of hop count should be maintained within all useful nodes. This algorithm relies on averaging of hops and is performed to calculate approximately the size of a single hop, upon receiving average size of the hop, left over node multiply the size of the hop with the total number of hop count to calculate the actual distance between two hops as shown in equation 1.

$$Hopsize_i = \frac{\sum \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}}{\sum h_j} \quad (1)$$

Where (x_i, y_i) , (x_j, y_j) are the coordinates of anchor node i and anchor node j, h_j is the hop between beacon node i and beacon node j. The technique facilitates the unknown nodes to receive hop size information, and save time, they transmit the hop size to their neighboring nodes and could assure that the majority of nodes receive the hop-size from a beacon node which has the least hop between them. Lastly unknown nodes compute the distances of the beacon nodes based on hop length.

2.2 LEACH (Low Energy Adaptive Clustering Hierarchy)

LEACH [3] is a cluster based routing protocol in which a cluster head collects information from sensor node belongs to a cluster and sends the information to the sink node after the collection procedure. To make all sensor nodes in this network consume their node energy equally and develop the life time of the network, this algorithm at random changes the cluster head, which in turn uses more energy compare to other nodes. To reduce the communication costs, the cluster head does information aggregation and then sends the information to the sink node. The theory is explained through a mathematical relation in equation 2.

$$T(n) = \begin{cases} \frac{Pt}{1 - Pt \cdot (r \bmod \frac{1}{Pt})} & \text{if } (n \in G) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Where Pt is the desired percentage of cluster heads, r is the current position number, G is the set of nodes that have not been cluster-heads in the last $1/Pt$ positions. It consists of two phases; a set-up phase and a steady state phase. This algorithm has three postulates that are specifically predecided namely cluster set, cluster node and cluster head, and can be seen in fig 1. The cluster head send the aggregated data to the sink node, called its base station. To reduce the slide of the cluster head, many positions of information frame transfer are performed followed by a repeat of the cluster reconfiguration procedure. Since LEACH uses a possibility in selection of cluster heads, its advantage is that all nodes have an opportunity of becoming a cluster head within a network, hence maintaining uniformity.

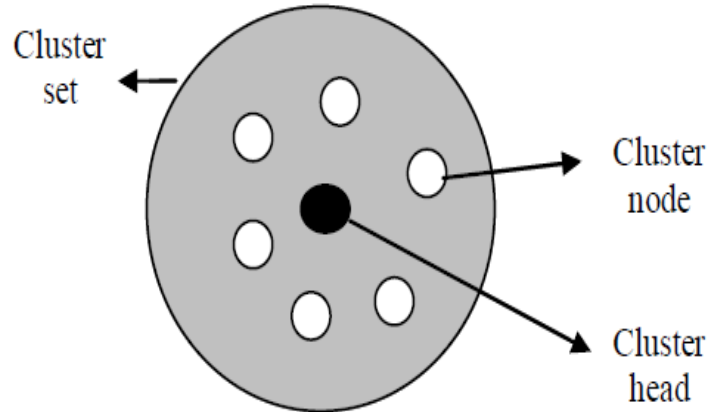


Figure 1: Clusters of LEACH Algorithm

2.3 Greedy Algorithms

In this approach whenever a node decides the transmission path based on the position of its neighbors, the source compares the localization of the destination with the coordinates of its neighbors and propagates the information to the neighbor which is closest to the final destination. The process is repeated until the packet reaches the deliberate destination. Several metrics related to the concept of closeness have been proposed in this area, among them, the most popular metrics is the Euclidean distance and the projected line joining the relaying node and the destination. In this scheme the unreliable neighbors are not taken into account for the

retransmissions. Another geographic protocol for information is discussed SPEED (Stateless Protocol for End-to-End Delay) to calculate approximately the delay of the transmitted packets [4][8]. The major limitation of the greedy algorithms is that the transmission may fail when the current holder of the message has no neighbors closer to the destination except itself. This could occur even when there is a possible path between the two extremes, for instance, when an obstacle comes into existence. The setup is shown in Figure 2.

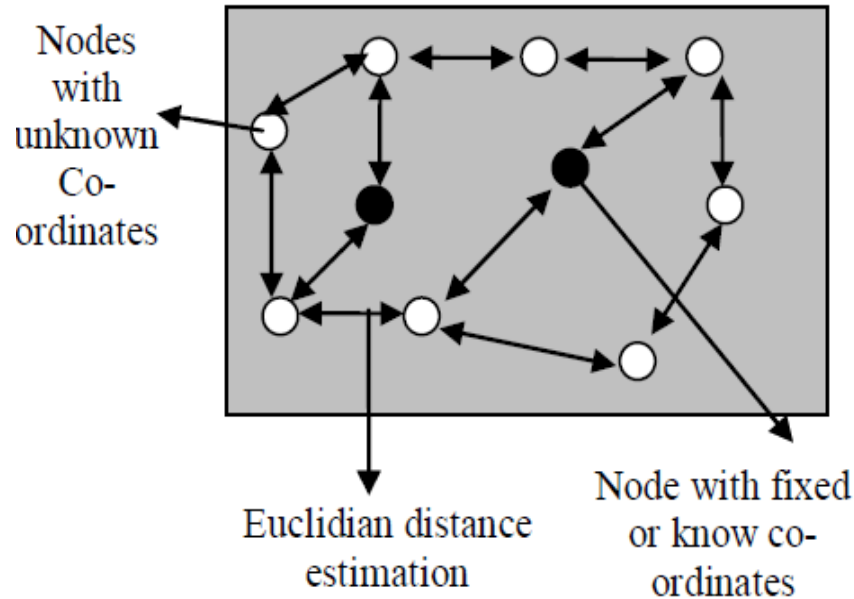


Figure 2: Distance Estimation of Greedy Algorithm

2.4 SPIN (Sensor Protocols for Information via Negotiation)

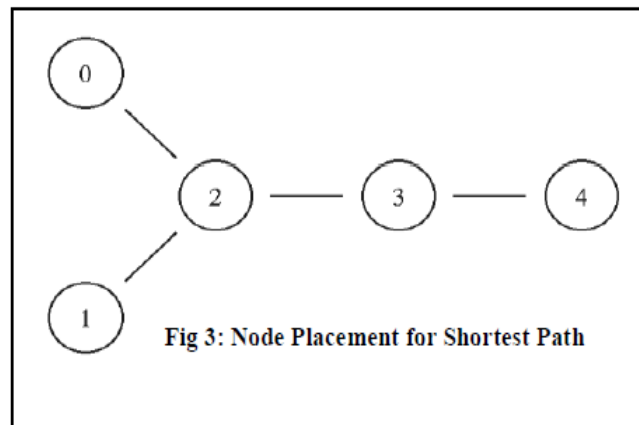
A family of adaptive protocols, called SPIN [5] is suggested efficiently to broadcast information among sensors in an energy-constrained wireless sensor network. Nodes organization a SPIN communication protocol name their information using high-level information descriptors, called as meta-information. They use meta-information discussions to eliminate the transmission of redundant information throughout the network. The SPIN nodes can base their communication decisions both upon application-specific knowledge of information and on knowledge of the resources that are available to the information. This allows the sensors to efficiently distribute information given a limited energy supply. Four specific SPIN protocols have been analyzed they are, SPIN-PP and SPIN-EC, which are optimized for a point-to-point set of connections, and SPIN-BC and SPIN-RL, which are optimized for a distribution network. In point-to-point networks, the sender announces that it has new information with an advertisement message to each neighbor. When the neighbor receives the information, the node checks the meta-information to know if it already has stored the information item. If the neighbor is interested in the information, it responds with a request memo, upon receiving it, the sender retransmits the information in a information message. The neighbor that receives the message informs about the availability to its own neighbors with an advertisement message. Taking into account the broadcast transmission, the node also responds with just one information message even when it has received multiple request messages. SPIN incorporates some consistency functionalities to keep track of the messages that it receives and its position of origin. This algorithm is also very successful in energy starved WSNs.

2.5 Data Centric Routing Protocol

The information centric routing protocol is the first group of routing protocols and discusses some conventional aspects. The SPIN which is a source-initiated protocol [6] does not apply a three stage handshake interface for disseminating information. The source and destination might transmit alternately as follows, request to send, ready to receive, send message, message received [7][9]. Meta-information is used to discuss with each other before transmitting information to avoid transmitting unnecessary information in the network. This protocol can be implemented for real time sensor networks.

2.6 A Basic Approach towards Problem Formulation

The route mapping problem requires a wide area of research, along with an algorithm, pertaining to different cases. There has been a great research on existing algorithms and suggested approaches for designing a network based on successful path for minimizing energy consumption. The basic problem in this environment relies totally on managing such a path that extensively overcomes and out performs the existing approaches. The algorithm suggested often finds use in applications based on a reasonable and pedagogic approach. The current trend mainly focuses on a probabilistic approach to transfer information from or within nodes deployed to form a self-sustainable wireless sensor network. The work highlighted in the paper shows an approach for transmission of information within randomly placed sensor nodes. It presents a basic technique for analyzing the information transfer between nodes that are deployed to form a WSN.



Initially five nodes are positioned in an open environment as shown in fig.3 for sensing the mechanism to route information depends totally on the path and its supportive algorithms. The direction of information transfer is calculated using the position of the node, protocols and the topology. Simulations are performed on NS2 and verification of results are generally discussed in sections to follow. The algorithm is as follows.

Step1: Label five nodes

Step2: Check information flow between nodes

Step3: Check route between node 0 to node 4

Step4: Check information transfer between nodes

Step5: Check route between node 1 to node 4

Step6: Check overlap between node 2 and node 3

Step7: If path breaks between node 0 and node 2 find novel path,

Step8: Novel path between node 1 to node 2 , If novel path fails,

Step9: Ensure again a novel path between node 2 to node 4

Step 10: Ensure information flow between node 2 to node 4, check information flow again

Step11: The shortest path is verified for information flow

2.7 Results

The information transfer mechanism can achieve between the nodes taking different routes in amount. The prophecy of route is performed over Nam (Network Animator). Nam provides clear prophecy of packet follow between nodes deployed to from a network. Initially the route followed is from node 0 to node 4 as shown in fig. 4 having some amount of time t which equals to 0.5 ns. During second mode of packet forwarding the route starts from node 1 to node 4 as shown in fig. 5 having the previous amount of delay time. It is noted that at the same instance of time t collisions of information packet occurs between nodes 2 and node 3 as shown in fig 6, as between node. The route gets break and information flow is intermittent. The key point to be noted is, upon route breaking between node 0 and node 2 and the path breaking between node 1 and node 2. The information flow starts from node 2 to node 4 adaptive in shortest path to promote the information flow in a continuous manner, this can be visualized in fig 7. The simulated results are captured in a trace file of nam as shown in the graphs highlighted in Figure 8.

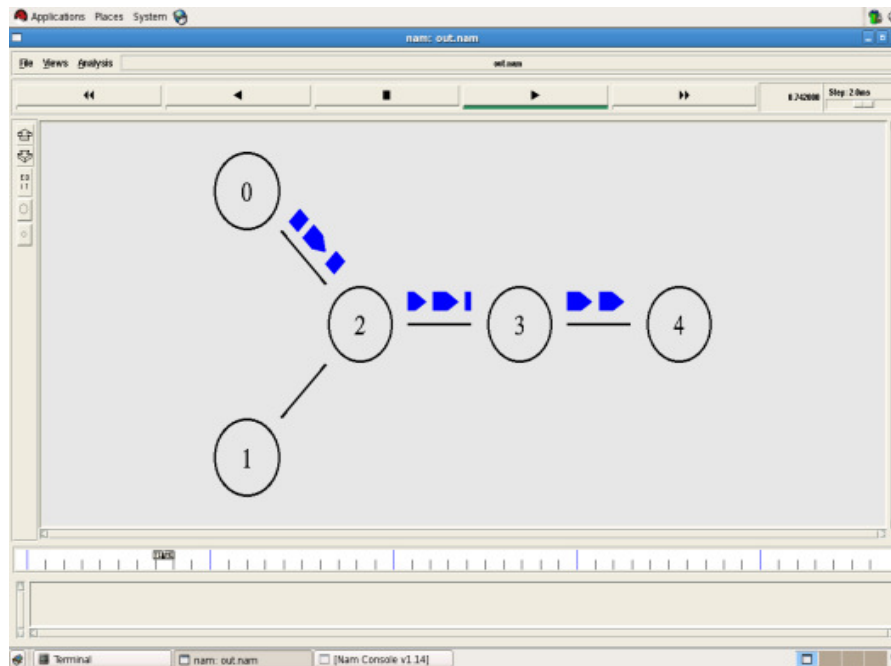


Figure 4: Information Flow from Node 0 to Node 4

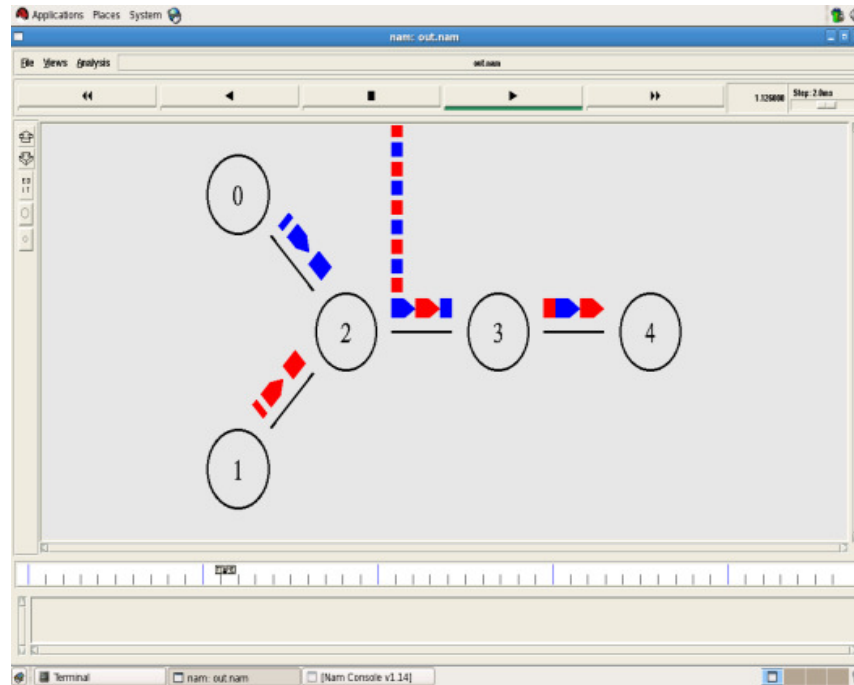


Figure 5: Information Flow from Node 1 to Node 4

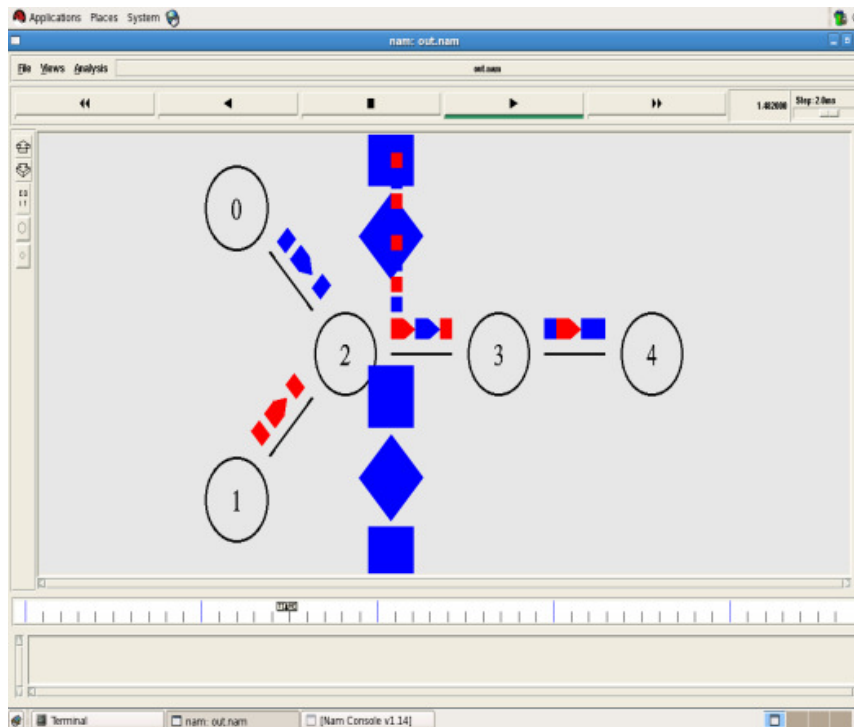


Figure6: Collisions of Information Packet

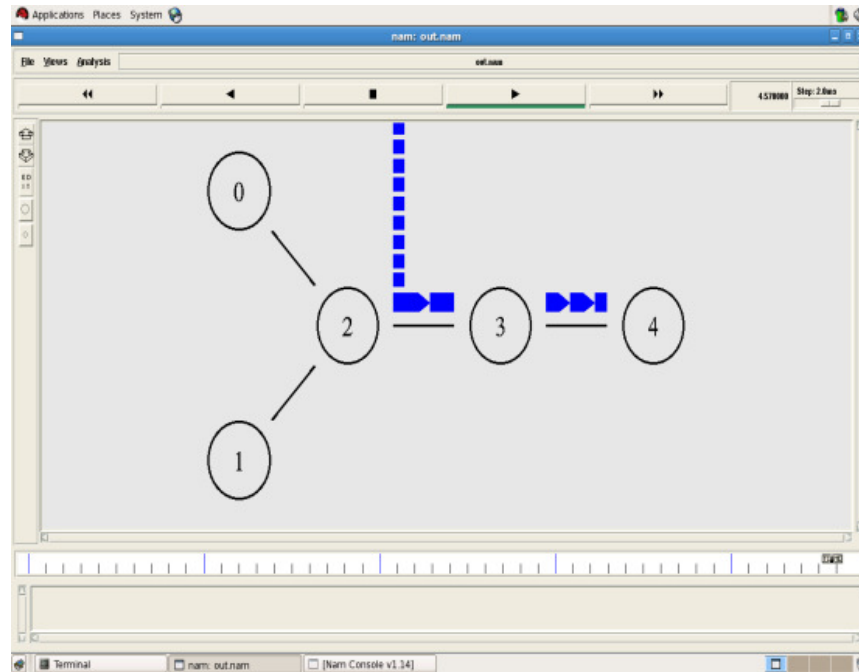


Figure 7: Path Break Between One Node to Another Node

3. CONCLUSIONS

The results show major trade off for the values plotted in this graph. The number of nodes and their proximity can be seen for a shortest path used to transfer information between nodes in an adhoc sensor network. The proposed scheme needs a particular justification and testing before applying to attain practical results pertaining to such types of deployments. In the graph blue line show the flow of information.

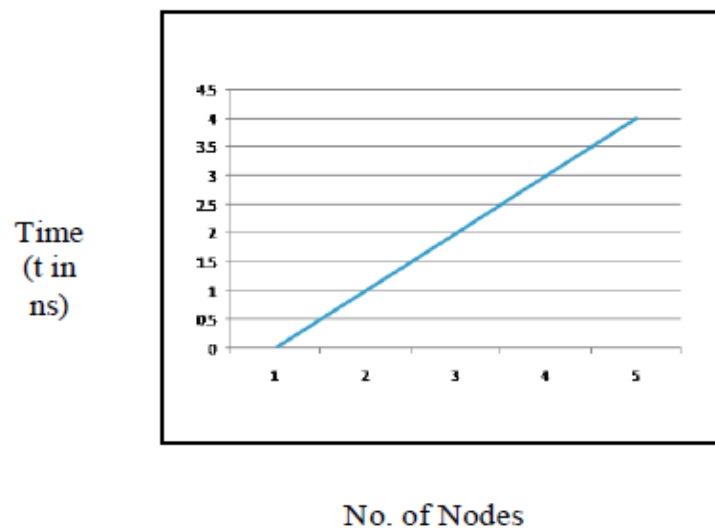


Figure 8: Simulation Results of Nodes Vs Time Plot

REFERENCES

- [1] R. Jurdak. "Modeling and Optimization of Ad Hoc and Sensor Networks," Bren School of Information and Computer Science, University of California Irvine. Ph.D. Dissertation. September, 2005.
- [2] T. He, C. Huang, B.M. Blum, J.A. Stankovic and T. Abdelzaher, "Range Free Localization Schemes for Large Scale Sensor Networks", ACM International Conference on Mobile Computing and Networking, pp.81-95,2003.
- [3] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks," proc. Hawaii International Conference on System Sciences, Vol. 8, pp. 1-10, Jan. 2000.
- [4] Kulik, J.; Heinzelman, W.; Balakrishnan, H. "Negotiation-based Protocols for Disseminating Information in Wireless Sensor Networks". 8, 169–185. Wirel. Netw.2002.
- [5] He, T.; Stankovic, J.A.; Lu, C.; Abdelzaher, T.F. SPEED: "A Stateless Protocol for Real-Time Communication in Sensor Networks". In Proceedings of the 23rd International Conference on Distributed Computing Systems (ICDCS), Providence, RI, USA, pp. 46–55. May, 2003.
- [6] Akkaya, K.; Younis, M. "A Survey on Routing Protocols for Wireless Sensor Networks". 3, 325–349, Ad Hoc Netw. 2005.
- [7] Wireless Sensor Networks by F. L. Lewis, [http://arri.uta.edu/acs/networks/Wireless Sensor Net Chap04.pdf](http://arri.uta.edu/acs/networks/Wireless%20Sensor%20Net%20Chap04.pdf), 2005.
- [8] Xing Guoliang, Lu Chenyang, Pless R, Huang Qingfeng. Im-pact of sensing coverage on greedy geographic routing algo-rithms. IEEE Trans. Parallel Distrib. Syst., 2006, 17(4):348{360.
- [9] Šimek, M.; Komosný, D.; Burget, R.; Morávek, P.; Sá Silva, J.; Silva, R. Data Gathering Model for Wireless Sensor Networks Based on the Hierarchical Aggregation Algorithms for IP Networks. International Journal of Computer Science and Network Security. 2008. 8(11). p.200-208. ISSN 1738-7906.
- [10] V. Vassiliou and C. Sergiou. Performance study of node placement for congestion control in wireless sensor networks. In Inter. Conf. on New Technologies, Mobility and Security, (NTMS), 2009.
- [11] Tony Ducrocq, Michael Hauspie, Nathalie Mitton and Sara Pizzi, On the Impact of Network Topology on Wireless Sensor Networks Performances, International Workshop on the Performance Analysis and Enhancement of Wireless Networks(PAEWN)(2014)

ACKNOWLEDGEMENT

The author gratefully acknowledges to Annamacharya Institute of Technology and Sciences for the Environment for funding this work. The author also acknowledges S.V. University College of Engineering for supporting our WSN deployment at the facility. Finally, a thank you goes to the reviewers for their insightful suggestions to improve the quality of this paper.

AUTHORS

N.Pushpalatha completed her B.Tech at JNTU, Hyderabad in 2004 and M.Tech at A.I.T.S., Rajampet in 2007. Presently she is working as Assistant Professor of ECE, Annamacharya Institute of Technology and Sciences Tirupati since 2006. She has guided many B.Tech projects and M.Tech Projects. Her Research area includes Data Communications and Ad-hoc Wireless Sensor Networks.



Dr.B.Anuradha is working as Professor in the Department of ECE, at Sri Venkateswara University College of Engineering since 1992. She has guided many B.Tech and M.Tech projects. At present Five Scholars are working for PhD. She has published a good number of papers in journals and conferences



INTENTIONAL BLANK

DUTY CYCLED MULTI CHANNEL MAC FOR WIRELESS SENSOR NETWORKS

M. Ramakrishnan

Department of Electrical and Electronics Engineering, Vel Tech Dr.RR and Dr.
SR Technical University, Chennai, India
dramakrishnan@veltechuniv.edu.in

ABSTRACT

In this work, Duty Cycled Sensor Multi Channel (DC-SMC) Medium Access Control (MAC) has been proposed for wireless sensor networks. The DC-SMC MAC uses a dedicated control channel and multiple data channels. The effective solution for the multi channel hidden terminal problem and missing receiver problem has been proposed in this work. The performance of the DC-SMC MAC has been compared with that of the single channel duty cycled CSMA/CA MAC by taking the throughput and latency as performance metrics. It has been shown that the duty cycled multi channel MAC gives high throughput and less latency even with lower duty cycles.

KEYWORDS

Multi Channel Medium Access Control, Duty Cycling, Wireless Sensor Networks, Multi channel hidden terminal problem

1. INTRODUCTION

A Wireless Sensor Network is the network of tiny devices, which has both sensing and communication capabilities. Nowadays, many sensor network hardware platforms like MICA2, Telos, etc, have an RF transceiver which is capable of communicating in different channels which can be dynamically selected from the firmware. The multi channel capability gives another degree of freedom for medium access in wireless sensor networks. Still multi channel medium access control inherently has some issues which have to be dealt with carefully, while doing the MAC design to improve the network performance. We classify the multi channel MAC protocols according to the channel assignment methods: *fixed assignment*, *semi-dynamic assignment* and *dynamic assignment*. In fixed assignment approaches, the radios are assigned channels for permanent use. Although the assignment of the channels can be renewed, for instance due to changing interference conditions, radios do not change the operating frequency during communication. In semi-dynamic approaches, the radios are assigned constant channels, either for receiving or transmitting, but it is possible to change the channel for communicating with the radios that are assigned different channels. In dynamic approaches, nodes are not assigned static channels and can dynamically switch their interfaces from one channel to another between successive data transmissions.

Natarajan Meghanathan et al. (Eds) : ICCSEA, SPPR, VLSI, WiMoA, SCAI, CNSA, WeST - 2014

pp. 157–174, 2014. © CS & IT-CSCP 2014

DOI : 10.5121/csit.2014.4723

In dynamic channel assignment approaches, every data transmission takes place after a channel selection. The channel selection can be, measurement based or status based. In measurement based approaches, the communicating parties measure the SINR values on a channel before transmitting. In status-based approaches, the nodes keep track of the status of the channels, such as busy or idle, according to the received control packets. In the dynamic multi channel MAC design, there are three types of implementations. They are, the split phase, the dedicated control channel and the channel hopping. In the Dedicated Control Channel based multi channel MAC, the nodes synchronize by exchanging control packets on the dedicated control channel and negotiate for the channel to be used for data exchanges. Examples of the dedicated control channel approaches are presented [1][2]. In the split phase approach, such as the MMAC [3] and MAP [4] time is divided into two phases—the appointed phase and the data transmission phase. In the appointed phase the nodes negotiate and select their channel for communication through the exchange of control packets in a common channel. In the data transmission phase, the data packets are transmitted on selected channels. This split phase multi channel MAC requires time synchronization among nodes. In the frequency-hopping approaches, nodes switch, or in other words hop, between different channels.

In the literature, the multi channel MAC has been proposed for systems with multiple transceivers and a single transceiver. To keep the cost and power consumption low, sensor nodes are equipped with a single transceiver. In this work a multi channel MAC has been proposed for wireless sensor nodes with a single half duplex transceiver, which uses a dedicated control channel for channel negotiation. A multi channel protocol performs better in the one-to-one topology, rather than in a star topology or topologies in which multiple source nodes communicate with a single sink node. But in sensor networks, convergecast communication is often used where multiple sensor nodes report their sensor data to the base station either in a single hop or multi hop fashion. So, the advantages of using multiple parallel links for communication shrinks, when the packet converges towards the base station in multi hop networks. To avoid this, multiple radios can be used at the Base Station to exploit the power of multi channel communication in Convergecast communication also. In general, in an ad hoc wireless sensor network, the multi channel MAC protocol improves the throughput and latency performance, as it allows concurrent transmissions in different orthogonal channels. This kind of multi channel MAC is inherently suitable for the Wireless Network Control System (WNCS), where the Multi channel MAC makes many wireless control loops co exist with each other. This significantly improves the network delay, which is the major influencing factor in the system performance in WNCS.

The paper is organized as follows. Section 2 describes the multi channel MAC proposed in the literature and section 3 discusses the proposed sensor multi channel MAC and its simulation results. Section 4 discusses the MATLAB based discrete event simulation of the proposed Sensor Multi Channel (SMC) MAC in a multi hop scenario, and section 5 reports the proposal of the duty cycled Sensor multi channel MAC (DC-SMC).

2. RELATED WORK

In this work, the usage of the multi channel MAC is to eliminate the interference to give better performance in terms of throughput and latency. Also the multi channel capability is exploited to give better energy efficiency in wireless sensor networks. This is achieved by introducing the multi channel feature in duty cycled MAC protocols to keep the throughput and latency constant even in low duty cycle conditions. In some works, the multi channel capability is used to avoid

jamming attacks [5][6][7]. Reference [8] has proposed a TDMA based multi channel MAC, YMAC for wireless sensor networks. It requires time synchronization among nodes. Reference [9] has proposed HyMAC, a hybrid TDMA/FDMA Medium Access Control for wireless sensor networks. It schedules the medium access for the nodes while using the multiple frequencies available in the commercial sensor node hardware platforms. In a reference [10], a TDMA based multi channel MAC for wireless sensor networks, TFMAC, has been proposed. The TFMAC requires time synchronization, and it uses single half duplex transceiver. This protocol divides each channel into time slots and the slot scheduling has been done in the medium access. The frame has been divided into a contention access period where the slot scheduling and channel allocation has been done and a contention-free period where the data transfer has been done. In the literature, a Multiple frequency Medium access control for wireless Sensor Networks (MMSN) [11] has been proposed, which divides the protocol into two functionalities. They are frequency assignment and medium access. In the frequency assignment, four different techniques are proposed. They are, 1. Exclusive frequency assignment 2. Even selection 3. Eavesdropping 4. Implicit consensus and the medium access is done by dividing the frame into broadcast contention period (T_{bc}) and transmission period (T_{Tran}). The node contends for the channel, for both broadcast and unicast with a non uniform back off. The paper assumes that the nodes are stationary, time synchronized and the frequency assignment has been done once. The Time synchronization overhead becomes higher than the RTS/CTS control packet overhead during low traffic conditions. Moreover, maintaining a tight time synchronization in the ad hoc multi hop wireless sensor is difficult. Hence, in this work, a dedicated control channel based multi channel MAC has been proposed for the wireless sensor network. In reference [12] asynchronous multi channel protocol (AMCP) has been proposed. The two issues, information asymmetry and flow in the middle which happens while using CSMA/CA in multihop environment has been stated. The multi channel MAC issues such as multi channel hidden terminal problem and missing receiver problem also has been stated. And the bottleneck analysis of dedicated control channel also reported and the theoretical upper bound for the number of data channels for a given channel capacity has been given. Though SMC MAC design has been inspired from the AMCP [12], the following differences exist.

1. In the channel negotiation of the AMCP, the transmitter selects a free channel and sends RTS with the selected channel. If the channel is not available in the receiver, it sends the negative CTS (nCTS) along with its channel status. Then the transmitter selects a common free channel as its preferred channel. This channel selection procedure decreases the throughput and latency performance, if the occurrences of transmission of nCTS are many. To avoid this, in the proposed SMC MAC, the transmitter sends its entire channel status to the receiver and the receiver selects a free common channel.
2. The AMCP does not explicitly specify the overhead associated with channel negotiation, whereas in the SMC MAC only an 8 bit field (Channel Status) has been included with the RTS and CTS control packet. This reduces the control packet overhead in wireless sensor networks where the data packet size is small.
3. The AMCP handles the Multi Channel Hidden Terminal (MCHT) problem by making the transmitting node wait after transmission for a specific time to avoid collision due to loss of channel information. This increases the latency. In the SMC MAC, the transmitting node senses the other channels to regain the information about other data channels. Hence the latency performance can be retained.

4. In the SMC MAC, the control channel is used as a broadcast channel which is required for many routing protocols in the network layer.

In a reference [13] the performance evaluation of the multi channel extension of 802.11 MAC has been done. It is stated that the channel assignment can be done by the measurement based method and the status based method. In our proposed multi channel MAC protocol, we use the status based method for channel assignment, and the measurement based method for avoiding loss of channel information problem, which is prevalent in the dedicated control channel multi channel medium access control protocols. In a reference [14], a cooperative multi channel MAC (CAM – MAC) has been proposed, in which the loss of channel information problem can be solved by getting the channel information from the cooperating neighbouring node to select a collision free channel for the communicating nodes. In another reference[15], the On Demand Channel Switching (ODC) has been proposed for multi channel medium access control. In this protocol, each node will stay in a channel as long as its traffic share in that channel does not go below a threshold value. If the traffic share of the node in a channel goes below the threshold, then the node will switch to a different channel after broadcasting the switching event. In a reference [16], the signal strength measurement based channel selection has been done in the proposed multi channel CSMA MAC. In reference[3], So et al have proposed the MMAC, which is a split phase multi channel medium access control protocol. At the starting of the beacon interval in the ATIM window, the node which has packets to transmit will negotiate for the channel, and if the channel is acquired, then the communicating nodes switch to that data channel and do the data transfer.

3. SENSOR MULTI CHANNEL MEDIUM ACCESS CONTROL FOR WIRELESS SENSOR NETWORKS

3.1. SMC MAC Algorithm

The proposed Multi channel MAC, Sensor Multi Channel MAC, has been described in the following section. The SMC MAC uses a single dedicated control channel and eight data channels. The Multi Channel MAC has been designed by taking the following points into consideration.

All the nodes are equipped with a single half duplex transceiver, which has the capability to switch from one channel to another channel dynamically. The switching can be done via software control. There are eight data channels and one control channel, and all the channels have equal capacity. The channel switching time is assumed to be negligible and all the channels are orthogonal and non-overlapping.

As a single channel is dedicated to the control packet flow, it creates a bottleneck. It poses a constraint on the number of data channels that can be used in the multi channel MAC. The number of data channels that can be used in a dedicated control channel MAC, is given by the following expression by neglecting the back off time.

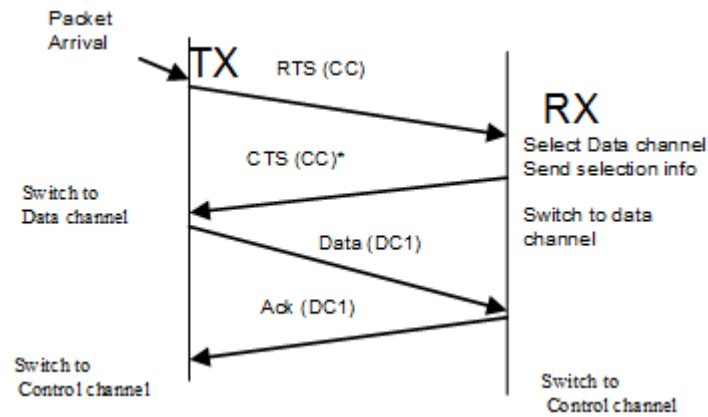
$$M = \frac{(T_D + T_R + T_C)}{T_R + T_C} \quad (1)$$

where	T_D	-	DATA and ACK transfer period
	T_R	-	RTS packet transfer period
	T_C	-	CTS packet transfer period
	M	-	Maximum Number of data channels

If the packet transfer period is quantified in bytes for convenience, then the T_D , T_R and T_C in the proposed SMC MAC protocol are 107(100+7), 7 and 7 bytes respectively. By substituting this we get $M = 8$.

The SMC MAC is described as follows:

1. Initially all the nodes stay in the control channel. The channel negotiation is done via the RTS/CTS control packets
2. When a packet arrives in a node, it sends the RTS with its channel status. The channel status is an eight bit field in which 0 indicates a free channel and 1 indicates a busy channel.
3. When the node to which the RTS has been transmitted, receives this packet, it selects the first common free channel for both the transmitter and the receiver and intimates the selection by setting the corresponding bit in the channel status field. Then the CTS is transmitted through this channel status field.
4. After the transmission of the CTS packet, the transmitting node switches to the selected channel
5. When the CTS packet is received by the intended node, it switches to the selected data channel. When the CTS is received by unintended nodes, the selected channel is marked as busy in the channel status table. The channel is marked as an idle channel, after the DATA+ACK transfer period.
6. The DATA and ACK are transmitted in the data channel. After the transmission and reception of the ACK packet, the node measures the RSSI for each channel and updates the channel status table. Thus, the loss of channel information problem is countered. Then, the node switches to the control channel.



(a)

CHANNEL NEGOTIATION – Algorithm

```

1:   Node.Ch_Status = [0, 0, 0, 0, 0, 0, 0, 0]
2:   IF Unicast Packet Arrival THEN
3:       RTS.Ch_Status = Node.Ch_Status
4:       Send Request To Send (RTS) Packet
5:   END IF
6:   IF RTS Received THEN
7:       flag=0;
8:       FOR  $k = 1:8$ 
9:           IF RTS.Ch_Status[ $k$ ] == 0 AND
                                   Node.ch_status[ $k$ ] == 0 THEN
10:              Node.Ch_status[ $k$ ] = 1;
11:              CTS.Ch_status[ $k$ ] = 1;
12:              CTS.Active_Channel =  $k$ ;
13:              Node.Active_Channel =  $k$ ;
14:              flag=1;
15:              break;
16:           END IF
17:       END FOR
18:       IF flag == 0 THEN
19:           --No Common free channel is available
20:           Do Nothing;
  
```

```

21:      ELSE
22:          Send CTS Packet;
23:      END IF
24:  END IF
25:  Switch the Transceiver to  $k^{\text{th}}$  Channel after CTS Transmission
26:  IF CTS Received THEN
27:       $k = \text{CTS.Active\_Channel}$ ;
28:       $\text{Node.Ch\_Status}[k] = 1$ ;
29:      Switch the Transceiver to  $k^{\text{th}}$  Channel
30:      Send DATA Packet;
31:  END IF
32:  IF DATA_Received THEN
33:      Send ACK Packet
34:  END IF
35:       $k = \text{Node.Active\_Channel}$ ;
36:       $\text{Node.Ch\_Status}[k] = 0$ ;
37:  Switch the Transceiver to the Control Channel after DATA Transmission
38:  IF ACK_Received THEN
39:       $k = \text{Node.Active\_Channel}$ ;
40:       $\text{Node.Ch\_Status}[k] = 0$ ;
41:      Switch the Transceiver to the Control Channel
42:  END IF

```

(b)

Figure 1.(a) Unicast packet flow in SMC MAC (b) Channel Negotiation Algorithm

The Control channel is used as the broadcast channel to support the broadcast which is required for the route discovery process of some routing protocols. Figure 1 illustrates the channel assignment and data transfer in SMC MAC.

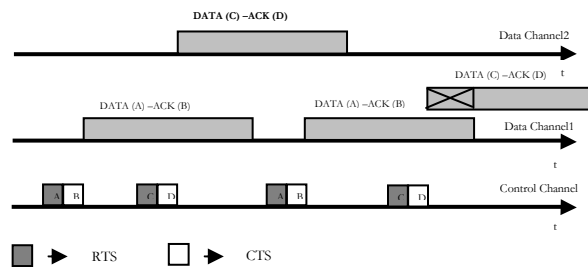


Figure 2. Multi Channel Hidden terminal problem

In the scenario shown in Figure 2, during the CD pair's data communication at channel 2, the AB pair has selected channel 1 for its data communication through the RTS/CTS packet (3rd control packet in Figure 2). Now as the CD pair could not overhear the channel negotiation of the AB pair, it has lost the channel 1 information. When the next unicast packet arrives at node C, the node selects data channel 1 which is already in use by the AB pair. This causes a collision. This is called the Multi Channel Hidden Terminal Problem. This happens due to the loss of channel information for the nodes C and D.

The SMC MAC solves this problem by sensing all the channels after the data transfer as shown in Figure 3. In the missing terminal problem shown in Figure 4, the node A tries to communicate with node C by sending the RTS packet, while C is busy in data transfer at channel 2. This problem can be alleviated in the SMC MAC by increasing the RTS timeout value, when the transmitter has detected that some data channel has gone into the busy state, while it was doing its previous data transfer. The RTS timeout happens in the wireless environment, due to a low SNR for the RTS in the receiver. In the multi channel environment, as the data transfer is offloaded from the control channel, the probability of getting high noise (low SNR) in the control channel is low.

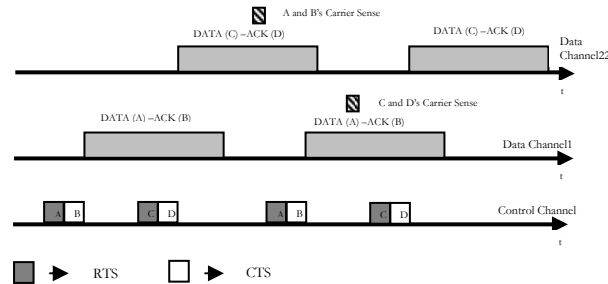


Figure 3. SMC MAC solution for multi channel hidden terminal problem

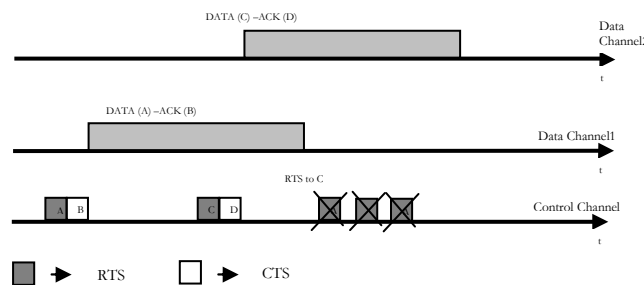


Figure 4. Missing Terminal Problem

3.2. Simulation of the SMC MAC:

A discrete event simulation has been done in MATLAB to analyze the performance of the proposed multi channel MAC. The simulation has been done for the single hop topology. In a single hop environment nodes are placed at random, and the one-to-one traffic is given. The simulation is repeated for different packet inter arrival times. Through simulation, the

performance metrics, the throughput and latency of the single channel MAC have been compared with those of the SMC MAC.

Table 1. Channel Status Table

	Ch1	Ch2	Ch3	Ch4	Ch5	Ch6	Ch7	Ch8
Status	0	1	0	0	0	0	1	1
Tx	-	2	-	-	-	-	4	1
Rx	-	3	-	-	-	-	5	10

0 – IDLE channel; 1 –BUSY channel

In the physical layer, a log-shadowing radio model is used. CSMA/CA with the RTS/CTS and random back off mechanism has been used in the MAC layer. Each node has a channel status table, which has a structure, shown in Table 1. The simulation parameters are summarized in Table 2.

Table 2. Simulation Parameters

S.No	Simulation Parameters	Value
1.	Number of Data Channels	8
2.	Radio Model	Log-Shadowing Model
3.	MAC Layer	CSMA/CA with RTS/CTS and Multi Channel extension
4.	Data rate	115 kbps
5.	Max. Power	+13dBm
6.	Area	30x30m
7.	Topology	Single hop Random Topology
8.	SNR _{threshold}	+30dBm
9.	Size of packets : RTS/CTS/DATA/ACK	7/7/100/7 Bytes

3.3.Performance of the Proposed Multi Channel MAC:

Simulation has been repeated for different values of the packet inter arrival time and the network throughput and latency have been observed for various traffic loads. From Figure 5 it is observed, that the throughput of the proposed multi channel MAC is higher than that of the single channel MAC during high traffic conditions. When there is a light traffic load, the performance of the multi channel and single channel MAC are similar. It is observed from Figure 6 that the latency in high traffic conditions for the SMC MAC is lower than that of single channel MAC. In this protocol, a random back off scheme is implemented for the access of the control channel. The contention in the control channel limits the latency performance of the SMC protocol in high traffic.

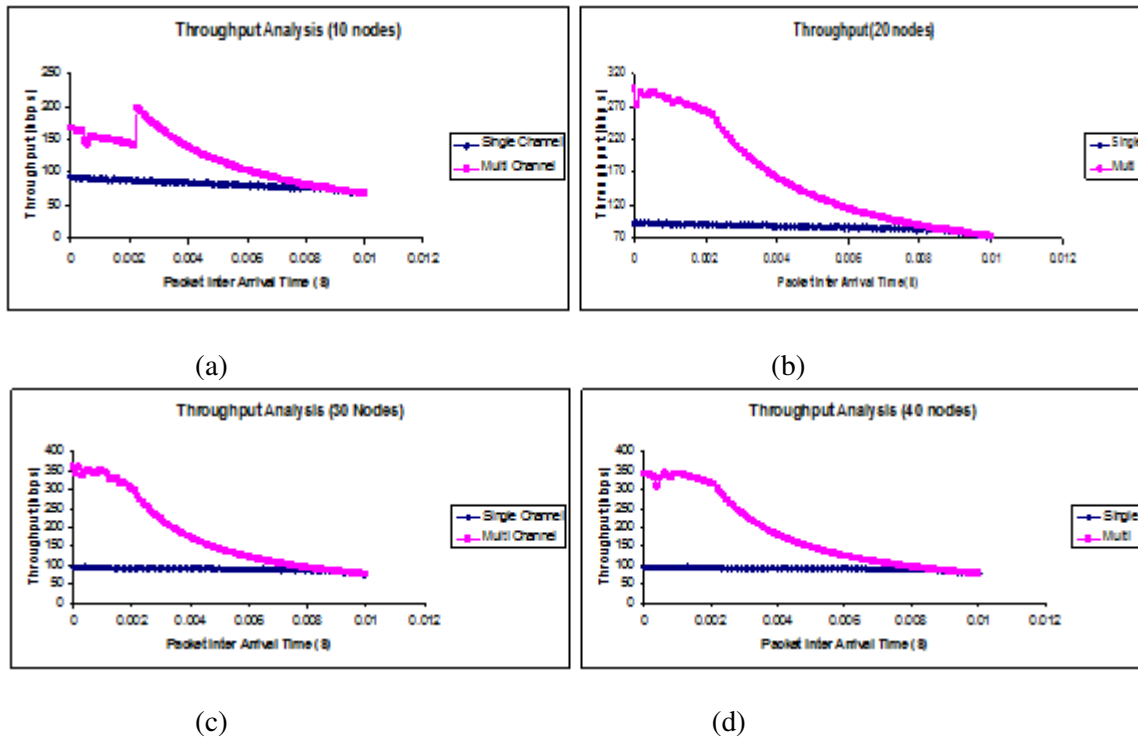
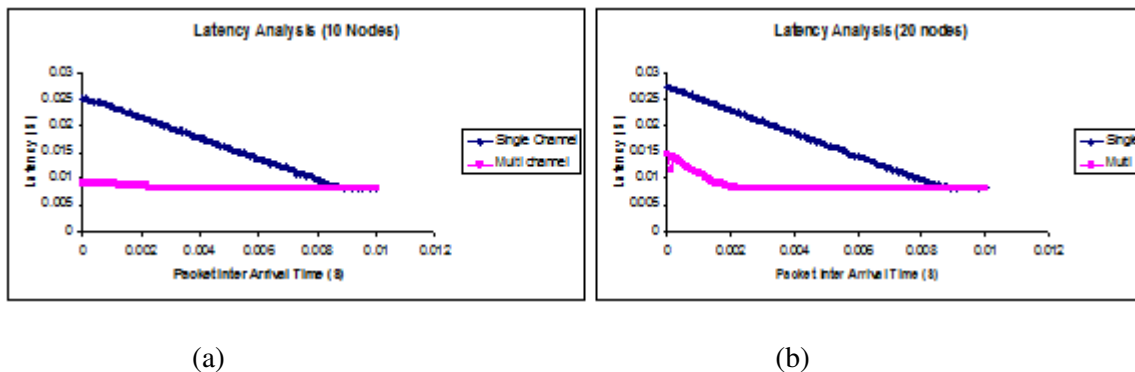


Figure 5 Throughput Analysis for Various Node densities (Random Topology) (a) 10 nodes (b) 20 nodes (c) 30 nodes (d) 40 nodes



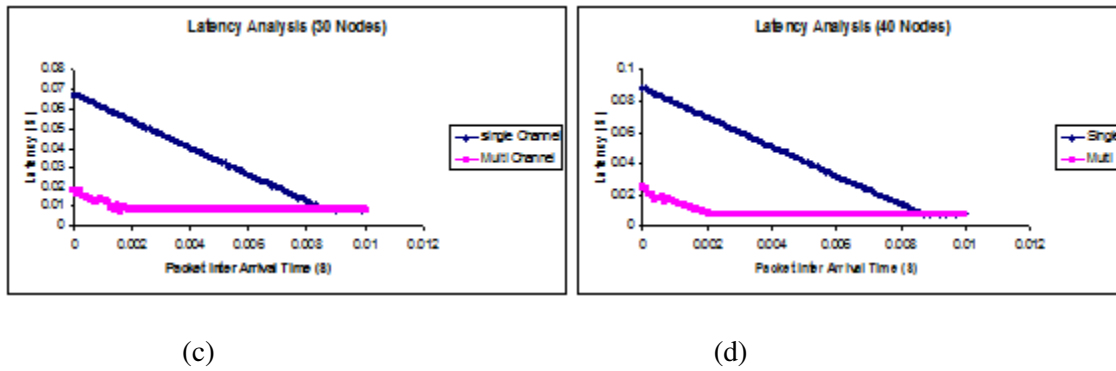


Figure 6. Latency Analysis for various Node Densities (Random Topologies) (a) 10 nodes (b) 20 nodes (c) 30 nodes (d) 40 nodes

Figure 7 shows the effect of the Multi Channel Hidden Terminal (MCHT) Problem on the throughput. Due to the loss of channel information, nodes select the data channel which is busy. This causes the collision of data packets, which decreases the throughput significantly. From Figure 7, it is observed that carrier sensing to retrieve channel information helps avoiding the Multi channel Hidden Terminal Problem in the proposed Multi Channel MAC.

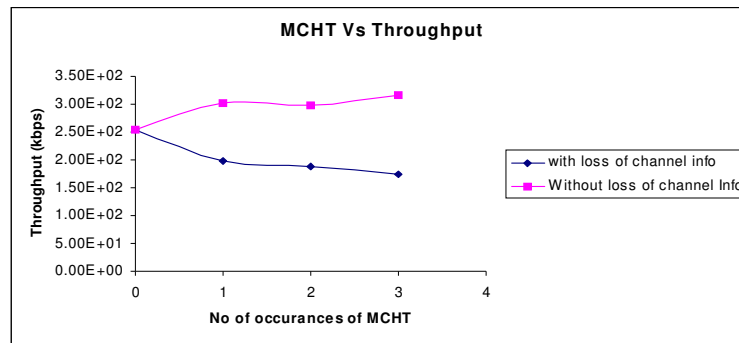


Figure 7. Effect of the Multi Channel Hidden terminal Problem on Throughput

3.4. Comparison of the SMC MAC with the AMCP

In the AMCP protocol channel negotiation is done as mentioned below.

The transmitter node, for the unicast packet should select a particular free channel as the preferred channel. And the RTS is transmitted with the preferred channel field. Now the receiver checks whether the preferred data channel is free. If it is free, then the CTS is sent, and it changes to the preferred data channel. After getting the CTS, the transmitter changes the frequency channel to the preferred channel, and transmits the DATA packet. If the preferred channel is busy in the receiver, then the receiver transmits a negative CTS (nCTS) packet along with its channel status. After receiving the nCTS the transmitter once again selects the common free channel, as the preferred channel and the RTS is transmitted again. When the preferred channel is not available in receiver, two extra control packets (nCTS and RTS) have to be transmitted. This affects the throughput and latency performance of the AMCP. To avoid the multi channel hidden terminal problem, after the data transmission, the AMCP marks the status of all the channels except the

current data channel as busy for the transmitter and receiver. The simulation has been done for the sensor network with 40 nodes, which are arranged in a single hop random topology to compare the performance of the SMC and the AMCP MAC in terms of throughput and latency. Figures 8 shows the throughput and latency comparison of the SMC and the AMCP MAC.

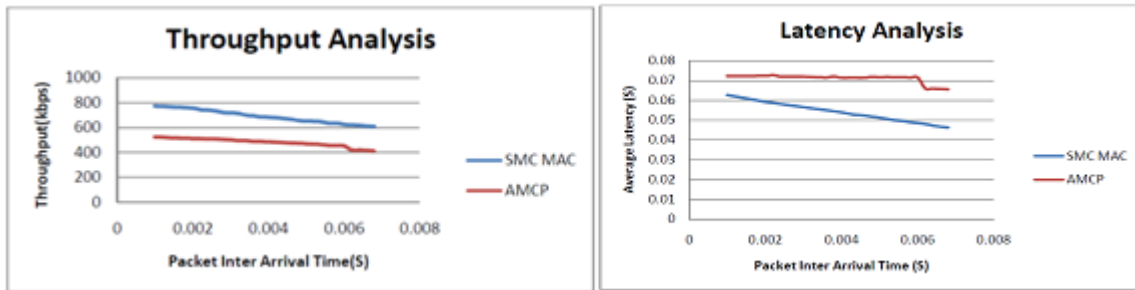


Figure 8. Throughput and Latency of SMC and AMCP MAC

From Figure 8, it is shown that the throughput of the SMC MAC is 27% to 32% higher than that of the AMCP. From the results it is shown that the throughput of the SMC MAC is 13% to 32% lower than the latency of the AMCP. The reason for this performance difference is that the AMCP inhibits the use of the data channels after the data transmission and reception. This leads to a negative CTS flow, and a consecutive RTS flow. This causes the drop in latency and throughput performance, whereas in SMC MAC, the RTS is transmitted with the channel status, so the receiver finds the common free channel and selects that for the data communication. To solve the Multi channel hidden terminal problem the AMCP proposes a method of inhibiting the nodes from using the channels other than the current data channel used Figure 10 shows the effect of Multi Channel Hidden Terminal Problem(MCHT) occurrences in throughput performance of SMC and AMCP MAC. The throughput of the multi channel MAC without MCHT avoidance is also shown.

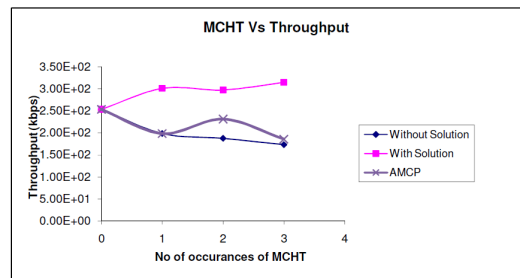


Figure 10. Impact of MCHT on Throughput of SMC and AMCP MAC

4. SMC-MAC SIMULATION IN MULTIHOP SCENARIO:

The low power radio in a wireless sensor node limits the communication range in single hop topologies. So it is preferable to have multi hop communication in wireless sensor networks.

Table 3. Simulation Parameters (Multi hop Scenario)

S.No	Simulation Parameters	Value
1.	Number of Data Channels	8
2.	Radio Model	Log-Shadowing Model
3.	MAC Layer	CSMA/CA with RTS/CTS and Multi Channel extension
4.	Data rate	115 kbps
5.	Max. Power	+13dBm
6.	Area	50x50m
7.	Topology	Multi hop Random Topology
8.	SNR _{threshold}	+30dBm
9.	Size of packets : RTS/CTS/DATA/ACK	7/7/100/7 Bytes
10.	Routing Protocol	DSR

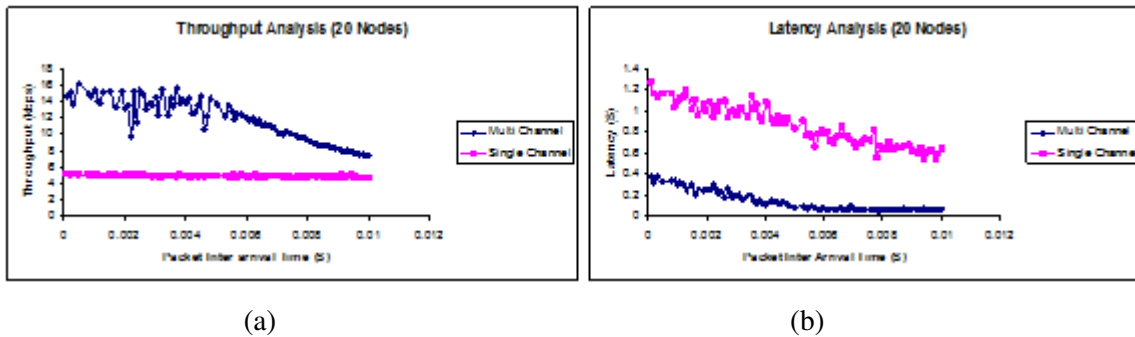


Figure 11. Performance Comparison of SMC MAC with Single Channel MAC in Multi hop environment (a) Throughput Analysis (b) Latency Analysis

The proposed SMC MAC has been simulated in MATLAB with multi hop random topologies. Dynamic Source Routing (DSR) is used as the routing protocol. The throughput and latency have been taken as performance metrics and simulation has been done for multiple node densities. The simulation parameters are summarized in Table 3. DSR is a reactive routing protocol. It will search for the route by broadcasting the Route Request (RREQ) packet. The propagation of the RREQ packets diminish the throughput and latency performance of the multi hop network more than that of single hop network topologies. From Figure 11(a) it is observed that the throughput performance of the SMC MAC is up to 70% higher than that of Single Channel CSMA/CA MAC. From Figure 11(b) it is observed that the latency performance of SMC MAC is up to 91% lower than its single channel counterpart. In the multi hop environment, the RREQ packet

propagation as broadcast in the network inhibits the greater throughput performance of a multi channel MAC. The higher throughput performance of the multi channel MAC is attributed to the parallel unicast packet flow in high traffic conditions.

5. DUTY CYCLED MULTI CHANNEL MAC

5.1 DC-SMC MAC

In a typical Medium Access Control for Wireless Sensor Network, duty cycling has been adapted for the energy conservation in a battery powered sensor node. The node periodically goes to sleep and to the listen state. The energy consumption of the 802.11 MAC is 2-6 times that of the SMC. Still the energy conservation due to the periodic sleep/wakeup schedule comes at the cost of increased latency and decreased throughput. In an Event driven sensor network, though the network is idle most of the time, when the event occurs the traffic will be significantly more. In this case, the duty cycled MAC exhibits increased latency and decreased throughput. In this work to reduce the latency, with the fixed duty cycle MAC, multi channel capability is combined, which significantly improves the latency performance even if the sleep time is more.

The periodic sleep/wakeup schedule has been included with the sensor Multi Channel MAC. The duration of the period ($T_{ON} + T_{OFF}$) is set as 6s and for various duty cycles 4%, 6%, 8%, 10% and 12%, the network throughput and average packet latency have been derived from the MATLAB based discrete event simulation. In simulation the one-to-one traffic is taken to show the performance of the Duty Cycled Multi Channel MAC. The data rate has been set as 9.6kbps and the size of the data packet is 100 bytes.

The simulation has been repeated by varying the packet inter arrival time, which is varied till half the period duration (3s). When the packet arrives at the MAC layer, two conditions are checked to process it. The conditions are

1. Whether the node is in the wakeup state? If the node is in the sleep state, then the packet has to be buffered for the transmission at the next wakeup schedule. If the node is in the wakeup state then the next condition is checked.
2. Whether the duration of the transmission of the packet is within the wakeup period.

If both the conditions are satisfied then the packet will be transmitted otherwise the packet is buffered for transmission at the next wakeup schedule.

5.2 RESULTS AND DISCUSSION

An analysis of the network throughput and average packet latency has been done for various duty cycle values (4%, 6%, 8%, 10% and 12%). As the number of packets transmitted is constant, the performance of the DC-SMC MAC becomes similar to that of the single channel duty cycled MAC, when the duty cycle increases. Figures 12 and 13 show the throughput and latency performance of the DC-SMC MAC and the single channel duty cycled MAC. From Figure 14, it is observed that the throughput of the DC-SMC MAC outperforms that of the single channel duty cycled MAC, during the lower duty cycle conditions which is necessary for the energy conservation. When the duty cycle is increased to 12% (Figure 12d.) the throughput performance

of DC-SMC MAC is similar to that of single channel MAC as there is no accumulation of packets to send as a burst at the starting of the wakeup period. Similarly, Figure 13 shows the variation of the latency for the different packets inter arrival time. It is also observed that the latency performance of the DC-SMC MAC is almost unaltered by the duty cycle variation for a particular traffic. Figure 14 shows the variation of the throughput and average packet latency for different duty cycle values. From Figure 14 it is observed that the throughput and latency of the DC-SMC MAC shows almost no variation for different duty cycle values, whereas the variation of the throughput and latency of the duty cycled single channel MAC is significant for different duty cycle values. As the duty cycle is directly proportional to the energy consumption, with the DC-SMC MAC, a higher throughput and lower latency can be achieved with minimal energy consumption.

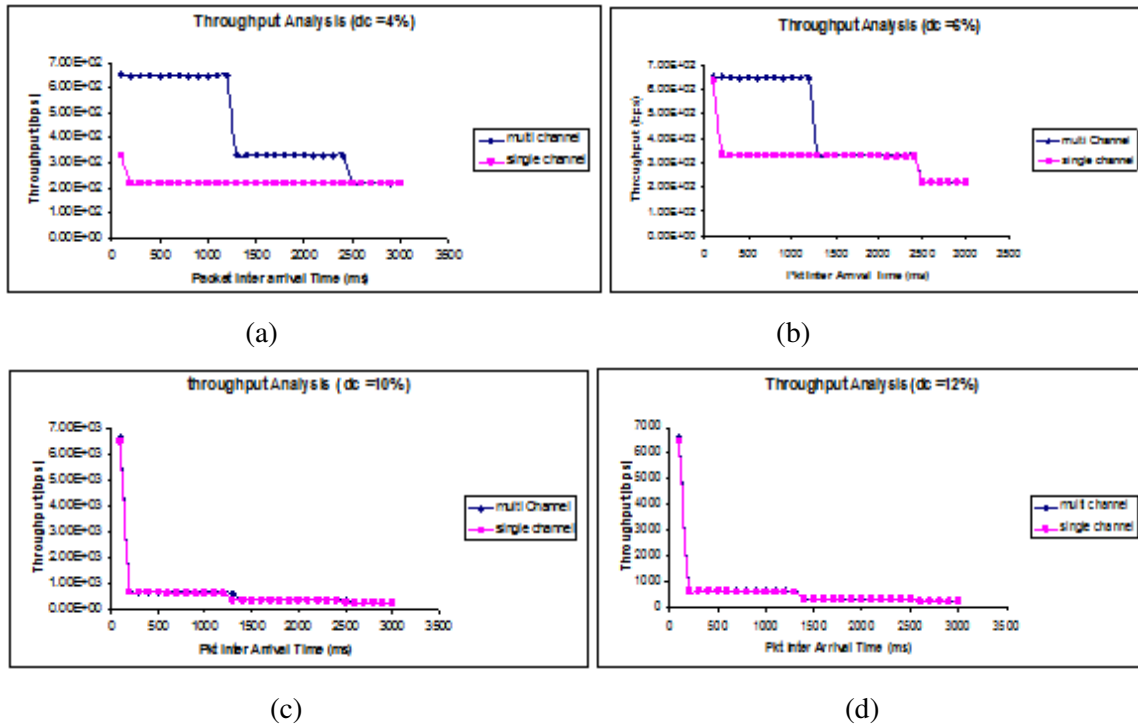
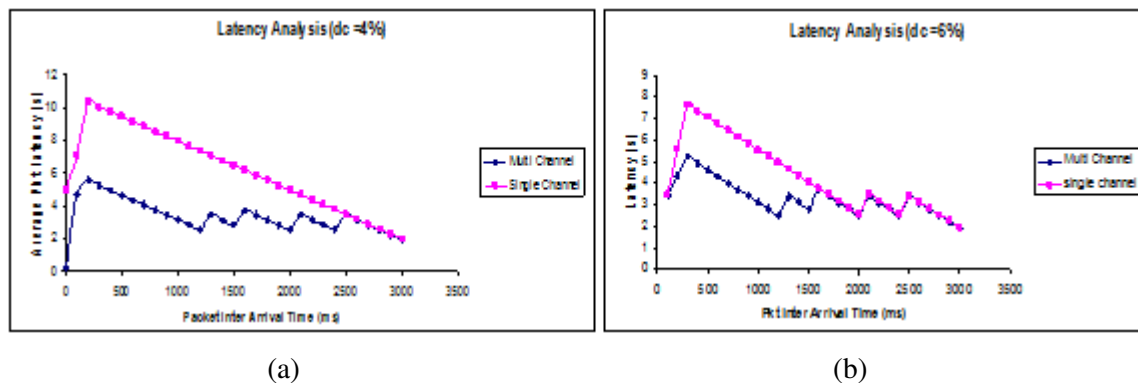


Figure 12. Duty Cycled Multi Channel MAC - Throughput Analysis (a) Duty Cycle 4% (b) Duty Cycle 6% (c) Duty Cycle 10% (d) Duty Cycle 12%



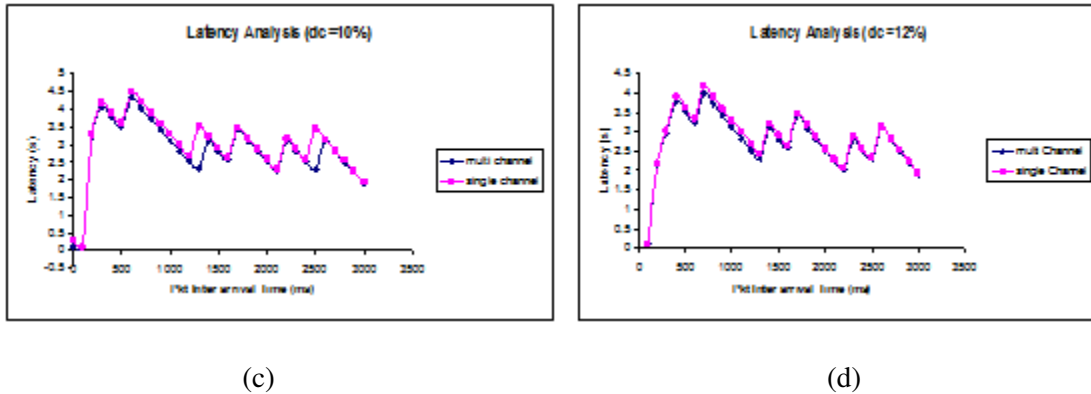


Figure 13. Duty Cycled SMC MAC – Latency Analysis
(a) Duty Cycle 4% (b) Duty Cycle 6% (c) Duty Cycle 10% (d) Duty Cycle 12%

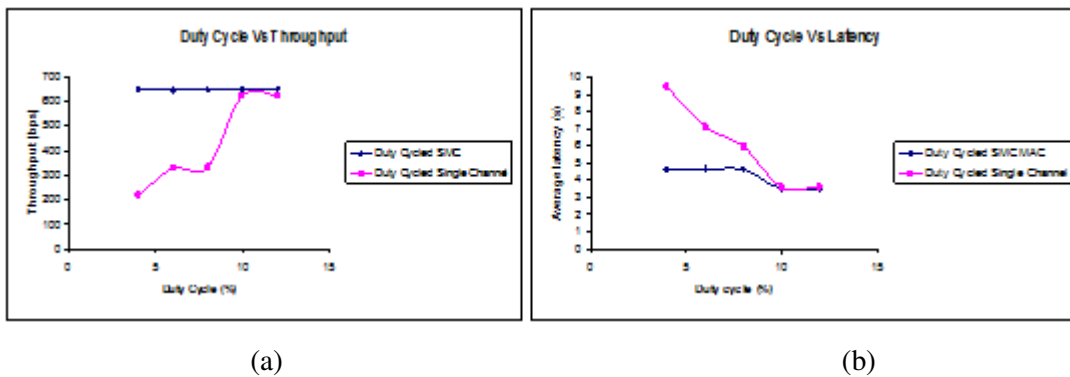


Figure 14. Effect of Duty Cycle in Protocol Performance (a) Duty Cycle Vs Throughput
(a) Duty Cycle Vs Average latency

5. CONCLUSION

In this paper Duty Cycled Sensor Multi Channel (DC-SMC) MAC has been proposed for wireless sensor network. The proposed multi channel protocol uses a dedicated control channel and eight data channels. The contribution of the paper is it combines the status based channel assignment and measurement based channel information retrieval. The paper proposes the scheme to alleviate the issues in multi channel MAC like multi channel hidden terminal problem and the missing receiver problem. By taking the throughput and latency as the performance metric, the performance of DC-SMC MAC has been compared with that of the single channel CSMA/CA with RTS/CTS for various traffic loads. It is observed that the performance of DC-SMC MAC outperforms the single channel MAC in the high traffic conditions. It has been shown that, with the DC-SMC MAC, a higher throughput and lower latency can be achieved with minimal energy consumption.

REFERENCES

- [1] Jain N., Das S. and Nasipuri A. (2001), 'A multichannel CSMA MAC protocol with receiver-based channel selection for multihop wireless networks', in Proceedings of the 10th International Conference on Computer Communications and Networks, Phoenix, pp. 432-439.
- [2] Li J., Haas Z.J., Sheng M. and Chen Y. (2003), 'Performance evaluation of modified IEEE 802.11 MAC for multi-channel multi-hop ad hoc network', in AINA 2003: Proceedings of the 17th International Conference on Advanced Information Networking and Applications, China, pp. 312-317.
- [3] So J. and Vaidya N.H. (2004), 'Multi-channel mac for ad hoc networks: handling multi-channel hidden terminals using a single transceiver', in MobiHoc '04: Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing, ACM Press, New York, NY, USA, pp. 222-233.
- [4] Chen J., Sheu S. and Yang C. (2003), 'A new multi channel access protocol for IEEE 802.11 ad hoc wireless LANs', in PIMRC 2003: The Proceedings of the 14th IEEE Personal, Indoor and Mobile Radio Communications Symposium, Vol. 3, pp. 2291-2296.
- [5] Alnifie G. and Simon R. (2007), 'A multi-channel defense against jamming attacks in wireless sensor networks', in Q2SWinet '07: Proceedings of the 3rd ACM workshop on QoS and security for wireless and mobile networks, New York, NY, USA, ACM, ISBN 978-1-59593-806-0, pp. 95-104.
- [6] Wood A.D., Stankovic J.A. and Zhou G. (2007), 'Deejam: Defeating energy-efficient jamming in IEEE 802.15.4-based wireless networks', in Proceedings of the 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON '07), pp. 60-69.
- [7] Xu W., Trappe W. and Zhang Y. (2007), 'Channel surfing: defending wireless sensor networks from interference', in IPSN '07: Proceedings of the 6th international conference on Information processing in sensor networks, New York, NY, USA, pp. 499-508.
- [8] Kim Y., Shin H. and Cha H. (2008), 'Y-MAC: An Energy- Efficient Multi-channel MAC Protocol for Dense Wireless Sensor Networks', in Proceedings of International Conference on Information Processing in Sensor Networks (IPSN'08), Missouri, USA, pp. 53-63.
- [9] Mastooreh S., Hamed S. and Antonis K. (2007), 'HYMAC: Hybrid TDMA/FDMA Medium Access Control Protocol for Wireless Sensor networks', in PIMRC 2007: The proceedings of the 18th IEEE Personal, Indoor and Mobile Radio Communications Symposium, Athens, Greece, pp. 1-5.
- [10] Milica Jovanovic D. and Goran Lj. Djordjevic (2006), 'TFMAC: Multi-channel MAC Protocol for Wireless Sensor Networks', in Proceedings of 8th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services (TELSIKS 2007), pp-23-26.
- [11] Gang Zhou, Chengdu Huang, Ting Yan, Tian He, John A. Stankovic and Tarek F. Abdelzaher (2006), 'MMSN: Multi-frequency media access control for wireless sensor networks', in Proceedings of IEEE International Conference on Computer Communications, Barcelona, Spain, pp.1-13
- [12] Jingpu Shi, Theodoros Salonidis and Edward W. Knightly (2006), 'Starvation mitigation through multichannel coordination in CSMA multi-hop wireless networks', Proceedings of the 7th ACM International Symposium on Mobile Ad hoc Networking and Computing, pp. 214-225.
- [13] Jiandong Li, Zygmunt J. Haas, Min Sheng and Yanhui Chen (2003), 'Performance Evaluation of Modified IEEE 802.11 MAC for Multi-Channel Multi-Hop Ad Hoc Network', in Proceedings of the International Conference on Advanced Information Networking and Applications (AINA 2003), pp. 312-317.
- [14] Luo T., Motani M. and Srinivasan V. (2006), 'CAMMAC: A Cooperative Asynchronous Multi-Channel MAC Protocol for Ad Hoc Networks', in Proceedings of IEEE Third International Conference. Broadband Communications, Networks and Systems (BROADNETS '06), San Jose, CA, pp.1-10.
- [15] Priyank Porwal and Maria Papadopouli (2004), 'On demand channel switching for multi-channel wireless MAC protocols', Proceedings of 12th European Wireless Conference, Athens, Greece, [online] <http://ics.forth.gr/netlab/mobile/publications/ew06.pdf>.

- [16] Asis Nasipuri and Samir R Das (2000), 'Multi Channel CSMA with signal power based channel selection for multihop wireless networks', Proceedings of Vehicular Technology Conference, Vol.1, pp. 211-218.

AUTHOR

M. Ramakrishnan was born on 9th May 1980 in Thirunelveli District, Tamil Nadu, India. He has completed his Bachelor's degree in Electrical and Electronics Engineering from the University of Madras in 2001 and has completed his Master of Engineering degree from Faculty of Electrical Engineering, Anna University, Chennai in 2005. He pursued his PhD studies in Wireless Sensor Networks as a UGC Research Fellow since 2007. After his PhD, he has worked as Chief Technology Officer in Reindeer Technologies Private Ltd, Chennai for 4 years and currently he is working as an Associate Professor in Department of Electrical and Electronics Engineering of Vel Tech Dr.Rangarajan and Dr.Shakunthala Technical University. He has 2 years of teaching experience. His research interests are protocol development for wireless sensor networks, embedded systems and signal processing.



AN EFFICIENT AND MORE SECURE ID-BASED MUTUAL AUTHENTICATION SCHEME BASED ON ECC FOR MOBILE DEVICES

Shubhangi N. Burde and Hemlata Dakhore

Department of Computer Science and Engineering,
RTMNU University, Nagpur
smathnikar@yahoo.com
hemlata.dakhore@raisoni.net

ABSTRACT

Mobile services are spread throughout the wireless network and are one of the crucial components needed for various applications and services. However, the security of mobile communication has topped the list of concerns for mobile phone users. Confidentiality, Authentication, Integrity and Non-repudiation are required security services for mobile communication. Currently available network security mechanisms are inadequate; hence there is a greater demand to provide a more flexible, reconfigurable, and scalable security mechanism. Traditionally, the security services have been provided by cryptography. Recently, techniques based on elliptic curve cryptography (ECC) have demonstrated the feasibility of providing computer security services efficiently on mobile platforms. Islam and Biswas have proposed a more efficient and secure ID-based system for mobile devices on ECC to enhance security for authentication with key agreement system. They claimed that their system truly is more secure than previous ones and it can resist various attacks. However, it is true because their system is vulnerable to known session-specific temporary information attack, and the other system is denial of service resulting from leaking server's database. Thus, the paper presents an improvement to their system in order to isolate such problems.

KEYWORDS

Authentication, Dynamic ID, Elliptic curve cryptosystem, Session key.

1. INTRODUCTION

Elliptic Curve (EC) systems as applied to cryptography were first proposed in 1985 independently by Neal Koblitz and Victor Miller. The discrete logarithm problem on elliptic curve groups is believed to be more difficult than the corresponding problem in the underlying finite field. The technology can be used, such as Diffie-Hellman and RSA with most public key encryption methods. Elliptic curve cryptography (ECC) is an approach to public key cryptography (PKC) based on the algebraic structure of elliptic curves over finite fields. According to some researchers, Elliptic curve cryptography (ECC) can have high level of security with a 164-bit key than other systems require a 1,024-bit key because ECC helps to establish equivalent security with lower computing power and battery resource usage. It is widely used for mobile

applications. Elliptic Curve Cryptosystem (ECC) based remote authentication system has been use for mobile devices. Mobile phones are most common way of communication and accessing Internet based services. However, the security of mobile communication has topped the list of concerns for mobile phone users. In 2009, Yang [6] proposed a system combining elliptic curve and identity-based cryptosystems to enhance security. They claimed that their system's secure against various attacks, such as replay attack, impersonation attack. But in the same year, Yoon [7] pointed out that Yang's system can't withstand impersonation attack. Furthermore, it doesn't achieve perfect forward secrecy property, which is a very important security in evaluating a strong authentication and key agreement protocol. Then, Yoon proposed another system to fix such problems. In 2010, Chen [5] proposed remote mutual authentication system for mobile devices to improve Yang's system. And they also claimed that their system's more secured to authenticate users and remote servers for mobile devices. However, Islam and Biswas [4] in 2011 have provided a security for mobile devices on elliptic curve cryptosystem. Then, they claimed that their system's truly efficient and usable for mobile users in many internet applications or wireless networks. Nevertheless, in this paper, we prove that the Islam's system can't resist known session-specific temporary information and denial of service resulting from leaking server's database attacks. Afterward, we propose an improvement of their system to overcome such entanglements. Besides, our system possesses low power consumption and computation cost than previous systems. Our main ideas aren't using point addition operation between a random point and user's authentication key and not letting random value is stored into server's database to fix recommended problems of Islam's system [4].

2. RELATED WORKS

This paper reviews the basic concepts of elliptic curve cryptosystem.

2.1 Elliptic Curve Cryptosystem

An elliptic curve's a cubic equation of the form

$$y^2 + a_1xy + a_2y = x^3 + a_3x^2 + a_4x + a_5, \quad (1)$$

where a_1, a_2, a_3, a_4, a_5 are real numbers. Elliptic curves over $GF(p)$ are of the form

$$E_p(a, b): y^2 \pmod{p} = x^3 + ax + b \pmod{p} \quad (2)$$

Where $a, b \in F_p$ and $(4a^3 + 27b^2) \pmod{p} \neq 0$. Given an integer $s \in F_p^*$ and a point $P \in E_p(a, b)$, the point multiplication $s \cdot P$ over $E_p(a, b)$ can be defined as

$$s \cdot P = P + P + \dots + P \text{ --- } s \text{ times} \quad (3)$$

After generating Elliptic curve, any number is entered and checked for prime number. If it is not prime number then lower number which is prime is selected as prime number.

2.2 Finding points on the curve:

The following algorithm gives the points on the curve $E_p(a, b)$ [1].

Algorithm elliptic points (p, a, b)

```
{
x=0
While(x<p)
{
```

```

w=(x3+ax+b) mod p
If (w is a perfect square in Zp)
Output ((x, √w), (x,-√w)) x=x+1
}
}

```

3. REVIEW & CRYPTANALYSIS OF ISLAM & BISWAS'S SCHEME

In this section, the paper “A more efficient and secure ID-based remote mutual authentication with key agreement scheme for mobile devices [4]” is review & show that their scheme's vulnerable to known session-specific temporary information attack and denial of service resulting from leaking server's database.

3.1 Review of Islam and Biswas's Scheme

This scheme includes four phases: system initialization phase, user registration phase, mutual authentication with key session agreement phase & leaked key revocation phase.

Some important notations in this scheme are listed as follows:

- S: The server.
- U: The user.
- IDU: Identity of U.
- AIDU: U's authentication key.
- qS: The private key of server S.
- rU: A secret number chosen by U.
- rS: A secret number chosen by S.
- H (.): A one way secure hash function.
- Kdf: A one way key derivation function.
- OR: OR operation.
- ||: Message concatenation operation.

3.1.1 System Initialization Phase:

The system initialization phase of Islam includes four steps:

Step 1: S selects a base point P with order n & K-bit prime number from the elliptic curve group G_p.

Step 2: S chooses random number qS (master key of the S) from [1, n - 1] and computes the public key QS = qS.P.

Step 3: S chooses a two one-way secure hash function

$$H1: \{0, 1\} \rightarrow G_p \quad (4)$$

$$H2: G_p \times G_p \rightarrow Z^*_p \quad (5)$$

And a one way key derivation functions:

$$Kdf: \{0, 1\}^* G_p \times G_p \rightarrow \{0, 1\}^k \quad (6)$$

Step 4: S publishes (Ep (a, b), P, QS, H1, H2, kdf)

3.1.2 User Registration Phase:

The user registration phase is proposed only once when the user wants to take part in the system. Islam's scheme involves three steps:

Step 1: U chooses identity $IDU = \{0, 1\}^p$ and submits it to S with some personal secret information via a secure channel.

Step 2: S checks U's IDU. If IDU already exists in the server database, S asks user U for different ID. Thereafter details of registration will be checked by S and computes the authentication key

$$AIDU = qS. H1(IDU \parallel X), \quad (7)$$

where $X \in Z^*_p$ is a random number chosen by S. S stores the information (IDU, X, status bit) about U to the secure database. S sets the status bit to 1 if the user's logged in, otherwise sets to zero.

Step 3: S returns AIDU to U via secure channel.

In this phase, Islam's scheme stores random value X into server's database. And if information of database is leak, then attackers can modify these random values of many users. Therefore, these users can not login to S at authentication phase & we'll fix this problem in this scheme.

3.1.3 Mutual Authentication with Key Session Agreement Phase:

In this phase, authors assume the message communication in this phase is over an open channel.

Step 1. U keys identity IDU and AIDU into the mobile device & randomly chooses a number rU from $[1, n - 1]$, and computes

$$N = R + AIDU \quad (8)$$

$$M = rU \cdot QS \quad (9)$$

Where $R = rU \cdot P$. U computes the dynamic identity

$$CIDU = IDU \oplus H2(R \parallel AIDU) \quad (10)$$

and sends the message (CIDU, N, M) to S.

Step 2. On receiving (CIDU, N, M), S computes

$$R^* = q^{-1}S \cdot M \quad (11)$$

$$AIDU = N - R^* \quad (12)$$

Then, S extracts the user's identity by computing

$$IDU = CIDU \oplus H2(R^* \parallel AIDU) \quad (13)$$

And checks the validity of IDU. If IDU is valid, S continues to next step, otherwise rejects U's login request.

Step 3 Furthermore, S computes

$$AID^*U = qS. H1 (IDU \parallel X) \quad (14)$$

(IDU and X are taken from server's Database) and checks $AID^*U = AIDU$. If it doesn't hold, the server S rejects U 's login request, otherwise chooses a random number rS from $[1, n - 1]$, then computes

$$T = R^* + S \quad (15)$$

$$HS = H2 (S \parallel AID^*U) \quad (16)$$

Where $S = rS$. P . Now S sends the message (T, HS) to U .

Step 4. On receiving (T, HS) , U performs $S^* = T - R$ and

$$H^* S = H2(S \parallel AIDU) \quad (17)$$

And checks $H^* S = HS$. If it holds, U authenticates S and sends (H_{RS}) , where $H_{RS} = H2(R \parallel S^*)$. U computes the session key

$$SK = \text{kdf} (IDU \parallel AIDU \parallel K) \quad (18)$$

Where $K = rS$. $R = rS$. rU . P .

Step 5. On receiving (HRS) , S computes $H^*_{RS} = H (R^* \parallel S)$ and compares it with H_{RS} . If it holds, S authenticates U and computes the session key

$$SK = \text{kdf} (IDU \parallel AIDU \parallel K) \quad (19)$$

Where $K = rS$. $R = rS$. rU . P . In this phase, the Islam's scheme performs point addition operation between random point R and $AIDU$. It's very dangerous because if information of any past session's random point R or S is revealed, $AIDU$ will be known by attackers.

3.1.4 Leaked Key Revocation Phase:

In this phase, authors assume that $AIDU$ is leaked to an adversary, so user U makes a request to server S for fresh authentication key. U submits the old authentication key $AIDU$, the identity IDU and personal secret information to S . Now S first checks the validity of U . After validating user's credential, server S selects another random number $X \in \mathbb{Z}^*_p$ and issues the fresh authentication key $AIDU = qS. H1 (IDU \parallel X)$ with old identity IDU . It's to be noted that the revocation of authentication key doesn't need new identity, only X will be changed in each revocation. S returns the new authentication key $AIDU$ to user U via secure channel. S keeps the database same except that X is replaced by X .

In their leaked key revocation phase, the information of user U is vulnerable to attacks because it's transmitted through open channel. So, the secure channel should be used to protect user U 's information when it's submitted in this phase.

3.2 Cryptanalysis of Islam and Biswas's Scheme

In this subsection, the paper shows that their scheme's vulnerable to known session-specific temporary information attack & denial of service resulting from leaking server's database.

3.2.1 Known Session-Specific Temporary Information Attack:

In paper, the authors mentioned that our scheme can resist known session-specific temporary information attack. In their opinion, when another adversary has the session ephemeral secrets rU and rS , he or she still can't compute session key SK because of lacking of $AIDU$'s information. However, it isn't true because with rU and rS , we'll prove that adversary still can know $AIDU$'s information of user U . For example, adversary A has rU , rS and past package ($CIDU$, N , M) of another user U , he or she'll perform following steps to obtain SK .

Step 1: Computes $R = rU \cdot P$ and $S = rS \cdot P$.

Step 2: Computes $AIDU = N - R$.

Step 3: Computes $IDU = CIDU \oplus H_2(R \parallel AIDU)$.

Step 4: Computes $SK = kdf(IDU \parallel AIDU \parallel K)$, where $K = rU \cdot rS \cdot P$.

In Islam's authentication phase, the authors performed point addition operation between a random point R and authentication key $AIDU$. This is a mistake because if R 's information is leaked, user U 's $AIDU$ will be easily computed.

3.2.2 Denial of Service Resulting From Leaking Server's Database:

In the user registration phase of Islam's scheme, we see that server S store (IDU , X , status-bit) of user U . This is dangerous because if information of server's database is leaked, another adversary can modify $X(s)$'s value(s). This causes many users not to login to the server S later. Following is the demonstration of this problem.

Step 1: User U sends login message ($CIDU$, N , and M) to server S .

Step 2: On receiving ($CIDU$, N , and M) from U , S computes

$$R^* = q^{-1} S \cdot M \quad (20)$$

$$AIDU = N - R^* \quad (21)$$

$$IDU = CIDU \oplus H(R^* \parallel AIDU) \quad (22)$$

$$AID^*U = qS \cdot H_1(IDU \parallel X') \quad (23)$$

Where X' is a modified random value of another adversary.

Step 3: S checks if $AIDU^* = AIDU$. Clearly it doesn't hold due to X' . So, S rejects user U . Hence, Islam's scheme's vulnerable to denial of service resulting from leaking server's database. In this scheme, we don't store random value to database to resist this kind of attack.

4. PROPOSED AUTHENTICATION SYSTEM

The proposed system will result more efficient enhancements for security on mobile devices using ECC. The proposed system not only inherits the advantages of their system, it also enhances the security. In registration phase, the main goal is achieving $AIDU$. Random value X helps to resist reregistration of attackers, with the same identity but various authentication keys at different time. In authentication phases, we use two random value rU and rS for server & user to challenge each other. Furthermore, we don't store random value X into database & don't perform point addition operation for $AIDU$. This system's divided into the four phases of system

initialization, user registration, and mutual authentication with key agreement & leaked key revocation phase.



Figure 1: System Design Model

4.1 System Initialization Phase

In this phase, three one-way hash functions are used. The system initialization phase includes four steps:

Step 1: S chooses k-bit prime number p & base point P with order n from the elliptic curve group G_p .

Step 2: S chooses random number q_S from $[1, n - 1]$

Step 3: S chooses three one-way hash functions

$$H1: \{0, 1\}^* \rightarrow G_p \quad (24)$$

$$H2: G_p \times G_p \rightarrow \{0, 1\}^k \quad (25)$$

$$H3: G_p \rightarrow \{0, 1\}^k \quad (26)$$

Step 4: The server publishes $(E_p(a, b), P, H1, H2, H3)$ as system parameters & keeps the master key q_S secret.

4.2 User Registration Phase

There are 3 requirements for a registration phase: secrecy for information transmitted between user & server, difference between keys provided for each time of registration by server & server mustn't store user's information which can be a hazardous risk. Easily, Islam's system achieved first two requirements but not the last. So, to recover this point accomplishes a good registration phase. This system consists of 3 steps illustrates these ones.

Step 1: U chooses identity $IDU = \{0, 1\}^k$ and Submits it to S with some personal information via secure channel.

Step 2: S checks U's IDU. If IDU already exists in the server's database, S asks U for different identity. Otherwise, S chooses a random value $X \in \mathbb{Z}_p^*$. Then, S computes

$$AIDU = q_S \cdot H1(IDU \parallel X) \quad (27)$$

Finally, S stores $(IDU, \text{status-bit})$ of that user U into database.

Step 3: S returns AIDU to U via a secure channel.

ECC User Authentication

System Initialization

Maximum Range Of elliptic curve Assumed (n) 4000

Random Prime Number Generated Within Range 1000 - 4000 (d)
1049

Public Key(m= d*p) 1049 * (9, 5)

User Registration Computations

IDU(unique user id) 12

Hash IDU 23a63310b8e95a3702t

Public Key(m= d*p) 1049 * (9, 5)

Server Computaion

Public Key(m= d*p)
1049 * (9, 5)

AIDU
Public Key(m= d*p) * (X || hash(IDU))

Store this IDU and send this to user for further process

Figure 2: ECC User Authentication

4.3 Mutual Authentications & Session Key Agreement Phase

Similarly, this phase also proposes 3 requirements that help authentication be more secure: firstly, user & server must use random values to challenge each other. Secondly, user & server share a secret session key. Finally, temporary information mustn't affect negatively to important information such as authentication key. In Islam's system, both user & server use random values to challenge each other. However, their system's easy to leak authentication key AIDU if any random point's known. Thus, this phase will fix this weak point. In this phase, S and U will have the same session key SK.

Step 1: At first, U keys identity IDU & the authentication key AIDU into the mobile device & randomly choose a number r_U from $[1, n - 1]$. Then, mobile device computes

$$R = r_U \cdot H1(IDU \parallel X) \quad (28)$$

$$R' = r_U \cdot AIDU \quad (29)$$

$$M = H2(R \parallel AIDU) \quad (30)$$

$$CIDU = IDU \text{ OR } H2(R) \quad (31)$$

Mobile device sends $(X, CIDU, M, R)$ to S.

Step 2: On receiving $(X, CIDU, M, \text{ and } R)$ from U, S computes $R^* = qS \cdot R$. Then, S extracts user's identity by doing

$$IDU = CIDU \text{ OR } H2(R^*) \quad (32)$$

and then checks the validity of the identity IDU. If IDU is valid, S continue to go next step, otherwise rejects U's login message request.

Step 3: U computes session key

$$SK = H3(X \cdot R^*) \quad (33)$$

Step 5: S authenticates U and computes session key

$$SK = H3(X \cdot R^*) \quad (34)$$

ECC User Authentication

[Previous](#)

Mutual Authentication With Key Session Agreement
Choose 3 one way hash function H1, H2, H3

IDU
IDU(unique user id)

AIDU
Public Key($m = d \cdot p$) * ($X \parallel \text{hash}(\text{IDU})$)
Public Key($m = d \cdot p$)

Key Calculation
Consider Random Number R_u
Compute ($R = R_u \cdot H1(\text{IDU} \parallel X)$)
Compute ($\text{CIDu} = \text{IDU} \text{ OR } H2(R)$)

Server Computaion Extract User Identity
Server First Compute $R^* = (\text{public key} \cdot R)$

Then server Extract users Identity by doing
 $\text{IDU} = \text{CIDu} \text{ OR } H2(R^*)$
This way server computes the identity of user
To calculate session key :
 $S = H3(X \cdot R^*)$

Figure 3: ECC User Mutual Authentication with session Key

4.4 Leaked Key Revocation Phase

This phase's similar to Islam's system. However, this phase use a secure channel in two ways to protect secret information of user. And Islam's system doesn't mention secure channel in this phase.

Research Work in the proposed Scheme:

The research work is to provide the secure channel integration. Each and every message and its response are passed through secure channel. By secure channel, the request is encoded using encryption method at source end and the request is again decrypted at the destination end by destination's private key. Then the computation of ECC starts. When the response is built, then response creator becomes the source and again encrypts the response. The destination again decrypts the response and then process the response. Base64 encoding schemes are used when there is a need to encode binary data that needs to be stored and transferred over media that are designed to deal with textual data.

The proposed scheme needs less computational amount than previous schemes. The performance of the proposed scheme evaluates in terms computation cost with other schemes. The proposed scheme is more efficient ID-based client authentication scheme for mobile client-server environments.

5. SECURITY AND EFFICIENCY ANALYSIS

This section discusses the 2 aspects i.e. security & efficiency of the proposed system.

5.1 Security Analysis

Here, various security properties must be considered for the mutual authentication and session key agreement scheme like replay attacks, impersonation attacks, stolen-verifier attacks, mutual

authentication, session key security and perfect forward secrecy, must be considered for the proposed scheme.

5.1.1 Replay attack

In the proposed scheme, the freshness of the messages transmitted in the mutual authentication with key agreement phase is provided by the random points RU and RS, and the shared session key k. Only U and S, who can get the session key k and the shared authentication key AIDU can embed the X and the k in the hashed messages generated by U and S respectively. Therefore, the proposed scheme can resist replay attacks.

5.1.2 Known Session-Specific Temporary Information Attack

Proposed scheme can resist this kind of attack like Islam's scheme. We assume that another adversary A knows random number of user and server of another past session. However, adversary still can't know session key and user authentication ID. So, adversary can't compute random point to know session key.

5.1.3 Stolen-verifier attack

The proposed scheme can withstand stolen-verifier attacks, because S doesn't store any table with information related to U. Server S generates a random value X for each user. Therefore, when authenticating with S, U only needs to send X to S and S uses master key qS to re-construct AIDU of that user. So, S doesn't need to keep U's password in the storage space when a new user's added in the system.

5.1.4 Mutual authentication

Mutual authentication means that both the user and server are authenticated to each other within the same protocol. In the proposed scheme, the goal of mutual authentication is to generate an agreed session key k between U and S for particular session. After S receiving the message from U(X, CIDU, M, R) to S. Afterward, S checks $M = H_2(R' \parallel AIDU)$ and U and S authenticate with each other. Therefore, the proposed scheme provides the mutual authentication.

5.1.5 Session key security

Session key security means that at the end of the key exchange, the session key is not known by anyone but two communication entities. In proposed scheme, after finishing mutual authentication successfully, both user & server share a session key SK to encrypt message later. So, proposed scheme not only satisfies mutual authentication but also provides session key to partners.

5.1.6 Perfect forward secrecy

Perfect forward secrecy means that if a long-term private key (e.g., user password/secret key or server private key) is compromised, this does not compromise any earlier session keys.

5.1.7 Lost/Stolen mobile device attack

In proposed scheme, client stores the information into his mobile device, which can help both the client and the server S for mutual authentication. Suppose an adversary steals mobile device, extracts information from the device and then try to get login S by using the extracted information. However, from the adversary cannot extract due to the difficulties of ECDLP problem. Therefore, adversary cannot get any valuable information from the stolen/lost mobile device that can help him to impersonate the client. Thus, the lost/stolen mobile device attack is infeasible to the proposed scheme.

5.2 Efficiency Analysis

To analyze computational complexity, compare efficiency between proposed system & the previous systems. That is, let H be the hash function operation, PM be the elliptic curve scalar point multiplication, and PA be the elliptic curve scalar point addition or subtraction. Furthermore, slight difference with Islam's system, the proposed system ignores XOR because it requires very few computations. Clearly, proposed system needs less computational amount than previous systems.

TABLE 1

Schemes/Computation Type	Yoon[3]	Islam[2]	TRUONG[1]	Proposed Scheme
Registration phase	1PM+1H	1PM+1H	1PM+1H	1PM+1H
Mutual authentication phase	7PM+4PA+12H	7PM+4PA+6H	7PM + 2PA + 10H	4PM+2PA+6H
Total computation cost	8PM+4PA+13H	8PM+4PA+7H	8PM + 2PA + 11H	5PM+2PA+7H
PM: Elliptic curve scalar point multiplication; H: hash operation; PA: Point Addition Operation				

Figure 3: Efficiency Comparison

6. COMPARISON WITH OTHER SCHEME

This section evaluates the performance of the proposed scheme in terms computation cost of proposed scheme with other schemes [2, 3, 4, 5, 6]. To estimate the computation cost of proposed scheme, the following notations are defined: PM is the time complexity to execute elliptic curve scalar point multiplication, H is the time complexity to execute hash operation and PA is the time complexity to execute elliptic curve scalar point addition. It is to be noted that the XOR operation needs very few computations; it is usually neglected considering its computational cost. The computation cost of a scheme is defined by the time spent by the client and the server for registration phase and mutual authentication with session key agreement phase. Besides, proposed scheme avoids the problem of clock synchronization, stolen verifier attack, denial of service attack and achieves users' anonymity as well, which requires two extra OR operations. In addition, the proposed scheme can offer resilience against various attacks such as many logged-in users' attack, lost/stolen mobile device attack, impersonation attack, known session-specific temporary information attack, privileged-insider attack, replay attack, etc. We summarize the computation cost of proposed scheme and carried out a comparison with other schemes in above Table, which shows that proposed scheme is more efficient ID-based client authentication scheme for mobile client-server environments.

7. CONCLUSIONS

With the continuous growth of wireless networks, such as GSM, CDPD, 3G and 4G, remote authentication systems play an important role in communicating between parties. After examining the security, implementation and performance of ECC applications on various mobile devices, we can conclude that ECC is the most suitable PKC system for use in a constrained environment. The efficiency and security makes it an attractive alternative to conventional cryptosystems. Consequently, we propose an improved system to eliminate some problems. Also provide the actual implementation of ECC based on the proposed paper. Compared with related systems, the proposed system has the following main advantages: It needs less computational cost. It provides secure user's anonymity. It doesn't hold any verification table. It provides mutual authentication

with session key agreement. As a result, the proposed system's able to provide greater security & be practical in wireless communication system.

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their helpful comments in improving our manuscript.

REFERENCES

- [1] Toan-Thinh TRUONG, Minh-Triet TRAN & Anh-Duc DUONG, "Improvement of the more efficient and secure ID-based remote mutual authentication with key agreement system for mobile devices on ECC", IEEE 26th International Conference on Advanced Information Networking and Applications Workshops, 2012.
- [2] S. H. Islam and G. P. Biswas, "A more efficient and secure id-based remote mutual authentication with key agreement system for mobile devices on elliptic curve cryptosystem", Journal of Systems and Software, vol. 84, no.11, 2011
- [3] Eun-jun yoon, Sung-bae choi and Kee-young yoo, "A secure and efficiency id-based authenticated key agreement system based on elliptic curve cryptosystem for mobile devices", International journal of innovative computing, information and control, April 2012.
- [4] J.H. Yang, C.C. Chang, "An ID-based remote mutual authentication with key agreement scheme for mobile devices on elliptic curve cryptosystem", Computers & Security 28, 2011, 138-143.
- [5] H. Debiao, C. Jianhua, H. Jin, "An ID-based client authentication with key agreement protocol for mobile client-server environment on ECC with provable security", Information Fusion, 2011,
- [6] T. H. Chen, Y.C. Chen and W.K. Shih, "An advanced ecc id-based remote mutual authentication system for mobile devices", Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, pp. 116-120, 2010
- [7] J. L. Tsai, T.S. Wu, H.Y. Lin and J.E. Lee, "Efficient convertible multi-authenticated encryption system without message redundancy or one-way hash functions", International Journal of Innovative Computing, Information and Control, 2010.
- [8] H. Debiao, C. Jianhua, H. Jin, "An ID-based client authentication with key agreement protocol for mobile client-server environment on ECC with provable security", Information Fusion, 2011.
- [9] J.H. Yang, C.C. Chang, "An ID-based remote mutual authentication with key agreement scheme for mobile devices on elliptic curve cryptosystem", Computers & Security 28, 2011, 138-143.
- [10] E.J. Yoon and K.Y. Yoo, "Robust id-based remote mutual authentication with key agreement system for mobile devices on ecc", IEEE International Conference on Computational Science and Engineering, vol. 2, pp. 2009, 633-640.
- [12] M.L. Das, A. Saxena, V. P. Gulati, "A dynamic ID-based remote user authentication system", IEEE Transactions on Consumer Electronics, 2009, 629-631.

AUTHOR

Shubhangi N. Burde

Department of computer science and engineering, G. H. Raisoni Institute of Engineering and technology for Women Nagpur, India smathnikar@yahoo.com



HUMAN GAIT ANALYSIS AND RECOGNITION USING SUPPORT VECTOR MACHINES

Deepjoy Das and Dr. Sarat Saharia

Department of Computer Engineering, Tezpur University, Assam

deepjoy2002@gmail.com

sarat@tezu.ernet.in

ABSTRACT

Human gait reveals feelings, intentions and identity which is perceived by most human beings. To understand this perceptual ability, Swedish psychologist Gunnar Johansson (1973), devised a technique known as PL (Point Light) animation of biological motion. In his work, the activity of a human is portrayed by the relative motions of a small number of markers positioned on the head and the joints of the body. This paper explores the basic concept of PL animation along with machine vision and machine learning techniques to analyze and classify gait patterns. Basically, frames of each video are background subtracted, the silhouette noise found were salt noise and noise connected in large blobs which are detected and removed based on morphological operations and area of connected components respectively. Image is then segmented and body points such as hand, knee, foot, neck, head, waist along with the speed, height, width, area of person are determined by an algorithm. We then fit sticks connecting pairs of points. The magnitude and direction of these stick along with other features forming a 24-dimensional feature vector for each frame of a video are classified using SVM Modelling using LIBSVM Toolkit. The maximum recognition accuracy found during testing by cross validation with parameters of LIBSVM was 93.5 %.

KEYWORDS

Human gait analysis, Recognition, body point identification, stick view, support vector machines.

1. INTRODUCTION

Biometric is the field of study that uses physiological or behavioural traits to identify a person. The study of human gait recognition is related to analysing human's distinctive way of walking and extract patterns used for recognition. Human gait can be used in identification of actions, gender, mood, emotions and intentions of a person. Gait requires no subject contact compared to other biometric technologies like fingerprint, iris detection, face recognition which requires the subject to be in close proximity with the sensor, therefore gait can be far more superior compared to other biometric technologies for example Geisheimer et al. [1] used continuous wave radar developed to record human gait signatures and extracted various gait parameters using signal processing techniques like short time Fourier transform (STFT).

Accuracy of gait analysis and recognition depends on two factors 1) feature extraction techniques followed to extract features from the walking subject e.g., a) models can be a shape based model for example, Lee & Grimson [2] used shape based features by dividing the silhouette into 7 regions and by extracting centroid, aspect ratio of major and minor axis of the ellipse and the orientation of major axis of the ellipse from each of the region , b) motion based model for example, Johansson [3] devised the technique known as PL animation, by which an observer was able to recognize walking subject affixed with lights in the major joints of the body or c) a combination of shape & motion based model. Feature extraction techniques should closely model the information captured by human brain 2) Classification model chosen to classify the gait patterns e.g., supervised techniques like Artificial Neural Network, Support vector machines etc. or unsupervised techniques like clustering etc.

2. RELATED WORK

In 1973 Johansson [3] devised PL animation, according to him an observer can recognise a person familiar to him when walking with light affixed in the major joint of their body. Since then various studies have been carried out using PL animation, studies which verify identify [4], sex [5] [13], emotion [6], facial expression using PL face [7]. Some studies also concluded that increasing the number of PL point on the human body increases accuracy of the recognition [8].

PL animation was proven to be a strong motion analysis model as studies show that manipulating PL animation data or purposefully degrading the information of PL data by fluctuating dot contrast polarity, blurring dots and spatial-frequency filtering doesn't degrade visual recognition performance of the observer [9].

Inverted PL animation however has some interesting results, studies show that human action are not easily perceivable using inverted PL animation [10]. However, inverted PL animation has an application in Parkinson's disease patients as a walking aid [11].

Studies have also been carried out using the stick view obtained by connecting the dots of PL animation. Hodgins et al. [12] studied tested stick view model, polygonal models, and others to check observer recognition performance in each of the model and reported polygonal model performance was greater than stick view. Gender identification study carried using stick view by Yoo et al. [13] using body contour & joint angles as features with SVM classifier shows good result.

Apart from motion based model, human gait can be also represented by a shape based model. Shape based model can be divided into contour based model [14] and region based model [15] [16]. Contour based model only uses boundary shape information whereas region based methods divide the shape region into parts which are then used for shape representation. Feature extraction methods using Chain coding, Fourier & Wavelet descriptor, Autoregressive, B-spline falls under contour based category, whereas Geometric, Zernike, Legendre, Krawtchouk Moments, Convex Hull falls under region based method.

Over a decade of research predicted that PL animation is the dominant model for motion analysis of humans and other living creatures. Our approach towards classifying gait pattern falls under motion based model viz. PL animation using Support Vector Machines. The study combines the ancient & intuitive idea of feature extraction using powerful classification model. Our throughout study concludes that motion based models are more powerful than shape based motion in terms of classification accuracy with less no. of features.

3. DATA COLLECTION

This section explains about the video capturing procedure followed, typically in terms of camera specification, camera placement, posture of walking subject and other specification. A camera fitted on a tripod was well adjusted so that the side view of walking subject can be captured. 10 videos of each walking subject were captured in 5 different locations (In each location 2 videos of each subject were captured). Different location have different light intensities, shadow fall, wind speed etc. We collected a video of 50 different walking subject. Videos are captured using Nikon S3000 which has Optical Sensor Resolution of 12 megapixels, frame rate of 25 frames/second and lens focal length 4.9-19.6mm/F3.2-5.9. The resolution of video used was 320×240 (QVGA Resolution).



Figure 1: Captured Frames during Data Collection

4. PREPROCESSING

After the data collection step, we go through a series of pre-processing step sequentially to get a perfect silhouette for feature extraction, the feature extracted is then modelled by SVM.

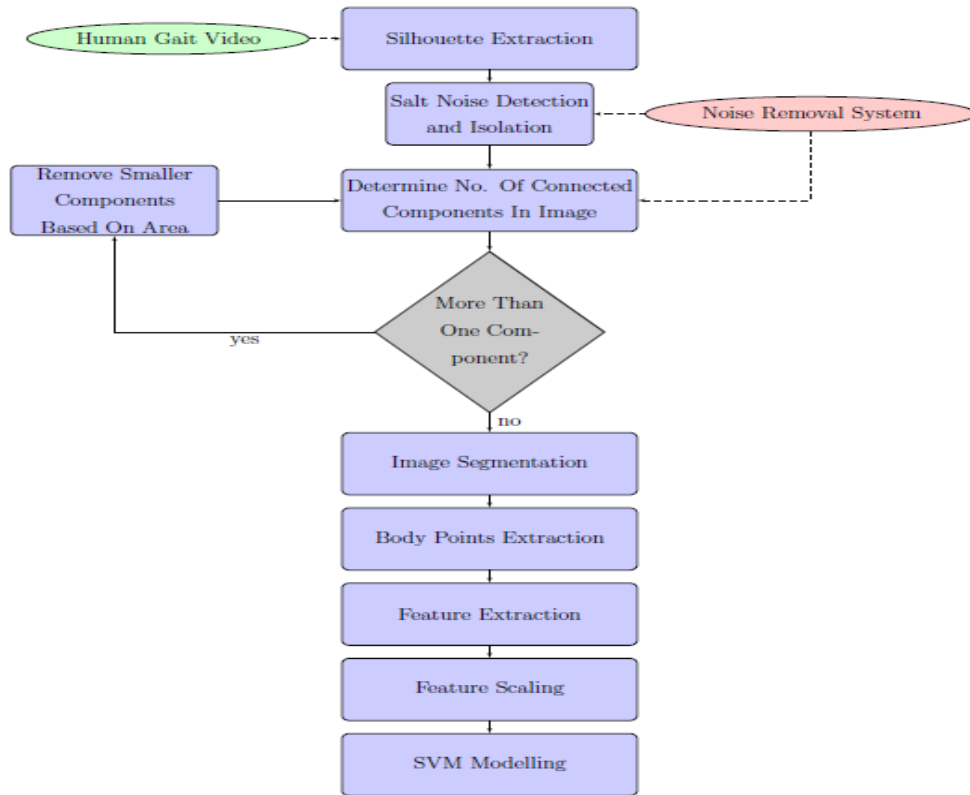


Figure 2: Flowchart of Preprocessing & Feature Extraction

4.1 Silhouette Extraction

Frame difference is the simplest form of background subtraction. The current frame is simply subtracted from the previous frame, and if the difference in pixel values for a given pixel is greater than that of threshold T_h then the pixel is considered part of the foreground [17].

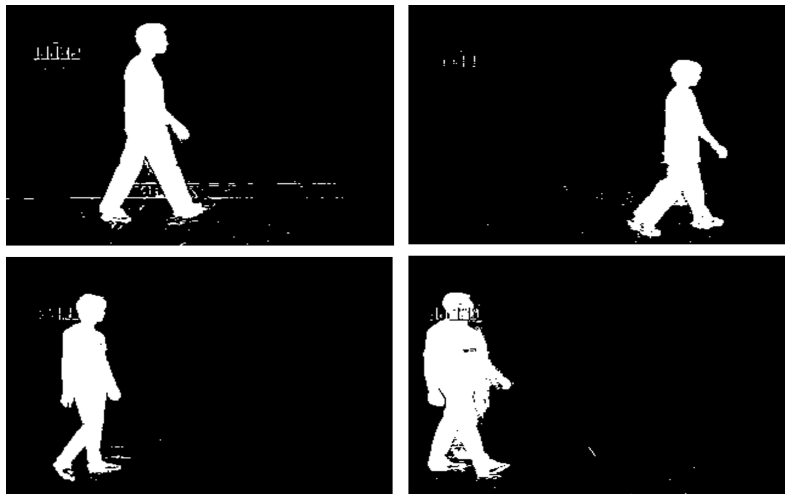


Figure 3: Background Subtraction

$$|frame_i - frame_{i-1}| > T_h$$

The estimated background is just the previous frame and it is very sensitive to the threshold T_h .

The threshold T_h has been determined by hit and trial method.

4.2 Noise Detection and Removal Techniques

The gait representation technique used in this paper may be partially biased by noisy silhouette. To overcome this, in this work the noisy silhouette are processed by the silhouette noise removal algorithms which efficiently removes two types of noise, which are mentioned as follows:

1. Intensity errors spread randomly across the silhouette (also called salt Noise) are removed by using a morphological opening operation on the silhouette with the square structuring element of size 6×6 (The morphological open operation is an erosion followed by a dilation, using the same structuring element for both operations).
2. Errors connected in large blobs in silhouette (greater than the size of structuring elements) are removed by finding the number of connected components in the binary silhouette and then filtering the component based on the area of the components. The low valued areas are filtered out and the largest area of connected component (which is the person in the silhouette) remains in the resulting silhouette.



Figure 4: Original Grayscale Frames Processed after Noise removal Techniques, the algorithm correctly identifies the largest area after filtering out the rest

5. FEATURE EXTRACTION

We identify nine major points of human body, sticks are fitted into the frames using the points. The magnitude of the sticks and angle between the sticks, along with height of the body and velocity of the centroid constitutes a feature vector for a single frames of a single class. A total of 24 features were used.

5.1 Estimation of Body Points

1. For each frame the centroid is calculated (the centroid is always inside the walking person image).
2. Keeping the vertical axis distance from the centroid fixed, the max and min point is searched in vertical direction.
3. Using max and min point the total height (H) of the person is estimated in each frame.
4. For a body height H, an estimate of head is 0% of H, neck 13% of H, Hand is between 35-50% of H, waist is 47% of H, knee is 71.5% and foot is between 93-98% of H. (all the points are distance from the horizontal axis).

The image is segmented further to search other body point. We now search in the horizontal direction keeping the fixed range 35-50% H in vertical direction to find the front and rear hand point (in term of distance from vertical axis) respectively. Similarly, the front and rear coordinates points of the knee and foot are identified.

Body parts were points are fitted are as follows: Head, Neck, Front Hand, Rear Hand, Centroid, Front Knee, Rear Knee, Front foot, Rear foot.

5.2 Computation of major angles between body parts

Apart from this we determine 9 angle values which are 1. Angle between the head and the body [some person walk with their head bend towards front/back] 2. Angle between the front hand and the body [Measures the amount of front arm swing] 3. Angle between the rear arm and the body [Measures the amount of rear arm swing] 4. Angle between front thigh and vertical axis [Measures the amount of front knee swing] 5. Angle between rear thigh and vertical axis [Measures the amount of rear knee swing] 6. Angle between front leg shin and vertical axis 7. Angle between rear leg shin and vertical axis 8. Angle between front leg foot and horizontal axis 9). Angle between rear leg foot and horizontal.

5.3 Total features vector

The relative distances between the points were measured with respect to other points as a human observer always looks at the relative distance with other points. Thus the Euclidean distance between pairs of different point were used as a feature.

We use the following set of features after finding Euclidean distance:

1. Euclidean distance of head point from the neck point.
2. Horizontal distance of front hand from the centroid point.
3. Vertical distance of front hand from the centroid point.
4. Horizontal distance of rear hand from the centroid point.
5. Vertical distance of rear hand from the centroid point.
6. Horizontal distance of front knee from the centroid point.
7. Vertical distance of front knee from the centroid point.
8. Horizontal distance of rear knee from the centroid point.
9. Vertical distance of rear knee from the centroid point.
10. Horizontal distance of front foot from the front knee point.
11. Vertical distance of front foot from the front knee point.
12. Horizontal distance of rear foot from the rear knee point.
13. Vertical distance of rear foot from the rear knee point.



Figure 5: Vectors corresponding to the extracted point are fitted to image. Magnitude and the direction of these vector will be used for classification

Along with these 13 Euclidean distances, 9 angle values (mentioned in sub-section 4.2) along with height and centroid velocity were used to form a 24 dimensional feature vector for each frame of a video. As SVM works on fixed sized vectors, we consider only the first 100 frames of a video, we have 100×24 feature matrix for each video. Since, 50 different persons gait data are present and for each person we have 10 different samples with different parameters changed for example. Therefore, we form 50000×24 feature matrix for SVM modelling using LIBSVM. (i.e., 50 video of different person with 10 different videos of each person with 100 frames of each video vs. 24 features each frame)

5.4 Feature Scaling

The goal is to use the features in LIBSVM (SVM Toolkit), therefore we scale the features in range [0, 1].

$$M'(i, j) = \frac{M(i, j) - \min(i)}{\max(i) - \min(i)} \quad \forall 1 \leq i \leq 50000, 1 \leq j \leq 24$$

Where M and M' represents the unscaled & scaled feature matrix of size 50000×24 respectively. M' has values in range 0 to 1.

6. RESULTS

For SVM modelling LIBSVM [18] toolkit is used. We first divide the feature matrix into training and testing matrix size each of size 25000×24 (5 video of each person is used for training and testing respectively). We then prepare the class label matrix for each of training and testing matrix set. We take the testing matrix and perform cross validation with different c (cost) & g (gamma) values using the default kernel provided by LIBSVM [18] (i.e., radial basis function) to check the performance at unique parameter values. The unique parameter c (cost) & g (gamma) values which yields the highest performance is then fixed and the system is trained to generate a final train model. The system is then tested using training matrix with the train model plugged into the LIBSVM system. The highest accuracy achieved by our method was 93.5%. Confusion matrix for the first 10 classes are shown below in figure 6.

		ACTUAL CLASS LABELS											
		1	2	3	4	5	6	7	8	9	10
	1	23891	0	9	2	0	0	4	0	0	3
	2	1	24816	0	0	0	1	0	2	0	1
	3	0	0	24021	0	0	3	0	0	0	0
PREDICTED	4	0	4	0	23194	0	0	0	0	13	0
CLASS	5	3	0	2	0	22982	0	0	0	8	0
LABELS	6	1	0	0	4	0	23456	0	3	0	0
	7	0	1	0	0	0	0	24781	0	0	0
	8	0	2	0	1	2	0	7	22495	0	0
	9	0	0	0	0	0	0	0	0	24698	0
	10	0	0	0	0	0	0	1	0	0	23003

Figure 6: Testing Confusion Matrix of SVM Testing for first 10 Class

7. CONCLUSION

The result using the 24-dimensional feature vector has better recognition. The 24 features were selected so precisely that these features were the most discriminant than any other features. We can conclude that the motion based model is much stronger representation of human gait than shape based models as motion based feature distinctly recognizes a person with less no. of features than shape based model. For examples, Lee and Grimson [2] used 41 to 57 features by which they get a recognition rate of 98%.

To improve the performance of the algorithm in terms of running time, this algorithm can be tried with anomaly detection algorithm. In which a priority can be assigned to the each features (e.g., the most discriminant feature among all the person gets the top priority). When a videos is fed into the recognition system, we take some top priority features and try to classify it to a classes of existing videos in database. If an anomaly is detected (i.e., if the frames of video cannot be classified in any existing classes) and maximum no. of samples (say x no. of samples) are not classified within a single class, then we add the lower priority feature and classify again. The procedure would yield better running time of the algorithm as its breaking the features into distinct set for classification and computing each feature & testing consumes significant amount of time.

REFERENCES

- [1] Geisheimer, Jonathon L., William S. Marshall, and Eugene Greneker. "A continuous-wave (CW) radar for gait analysis." Signals, Systems and Computers, 2001. Conference Record of the Thirty-Fifth Asilomar Conference on. Vol. 1. IEEE, 2001.
- [2] Lee, Lily, and W. Eric L. Grimson. "Gait analysis for recognition and classification." Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on. IEEE, 2002.
- [3] Johansson, Gunnar. "Visual motion perception." Scientific American (1975).
- [4] Cutting, James E., and Lynn T. Kozlowski. "Recognizing friends by their walk: Gait perception without familiarity cues." Bulletin of the psychonomic society 9.5 (1977): 353-356.
- [5] Kozlowski, Lynn T., and James E. Cutting. "Recognizing the gender of walkers from point-lights mounted on ankles: Some second thoughts." Attention, Perception, & Psychophysics 23.5 (1978): 459-459.
- [6] Clarke, Tanya J., et al. "The perception of emotion from body movement in point-light displays of interpersonal dialogue." Perception-London 34.10 (2005): 1171-1180.
- [7] Bassili, John N. "Facial motion in the perception of faces and of emotional expression." Journal of experimental psychology: human perception and performance 4.3 (1978): 373.

- [8] Blake, Randolph, and Maggie Shiffrar. "Perception of human motion." *Annu. Rev. Psychol.* 58 (2007): 47-73.
- [9] Ahlström, Vicki, Randolph Blake, and Ulf Ahlström. "Perception of biological motion." *Perception* (1997).
- [10] Sumi, Shigemasa. "Upside-down presentation of the Johansson moving light-spot pattern." *Perception* 13.3 (1984): 283-286.
- [11] Dunne, J. W., G. J. Hankey, and R. H. Edis. "Parkinsonism: upturned walking stick as an aid to locomotion." *Archives of physical medicine and rehabilitation* 68.6 (1987): 380-381.
- [12] Hodgins, Jessica K., James F. O'Brien, and Jack Tumblin. "Perception of human motion with different geometric models." *Visualization and Computer Graphics, IEEE Transactions on* 4.4 (1998): 307-316.
- [13] Yoo, Jang-Hee, Doosung Hwang, and Mark S. Nixon. "Gender classification in human gait using support vector machine." *Advanced concepts for intelligent vision systems*. Springer Berlin Heidelberg, 2005.
- [14] Lu, Jiwen, and Erhu Zhang. "Gait recognition for human identification based on ICA and fuzzy SVM through multiple views fusion." *Pattern Recognition Letters* 28.16 (2007): 2401-2411.
- [15] Little, James, and Jeffrey Boyd. "Recognizing people by their gait: the shape of motion." *Videre: Journal of Computer Vision Research* 1.2 (1998): 1-32.
- [16] Pranjit Das, Deepjoy Das and Dr. Sarat Saharia "Gait Analysis and recognition for human identification", National Seminar on Advances in Electronics and Allied Science & Technology, (NaSAEAST- 2013)
- [17] Das, Deepjoy, and Dr Saharia. "Implementation And Performance Evaluation Of Background Subtraction Algorithms." *arXiv preprint arXiv:1405.1815* (2014).
- [18] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011): 27.

AUTHORS

Deepjoy Das

Has done Bachelor of Technology from Tezpur University and was held as Assistant Project Engineer in Indian Institute of Technology for one year. His research interest includes image processing with supervised and unsupervised learning and handwriting recognition.



Dr. Sarat Saharia

He is currently working as Associate Professor in the department of computer science and engineering at Tezpur University, Assam, India. His research interest includes pattern recognition and image processing.



INTENTIONAL BLANK

A MIND MAP QUERY IN INFORMATION RETRIEVAL: THE ‘USER QUERY IDEA’ CONCEPT AND PRELIMINARY RESULTS

Rihab Ayed¹, Farah Harrathi², M. Mohsen Gammoudi² and Mahran Farhat³

¹Tunisia Polytechnic School, University of Carthage, Tunis, Tunisia
rihabb.ayed@gmx.com

²Higher Institute of Multimedia Arts of Manouba, University of Manouba, Tunisia

{harrathi.farah@gmail.com, mohamed.gammoudi@fst.rnu.tn}

³Faculty of Sciences of Tunis, University Tunis El-Manar, Tunisia
farhatmahran@gmail.com

ABSTRACT

Users in Information Retrieval are formulating since many years their queries in a bag of words that should be understandable by the system. The problem of ‘bag of words’ format is that it can cause the deformation of the user’s information need. A question is raised in this paper to discover if there is a more faithful and richer way to users to formulate their idea of search. This paper proposes an approach for users to model their queries in Information Retrieval (IR) based on the use of a brainstorming technique (Mind Mapping). The choice of the query representation model is based on assertions concerning Human Mind and habits of thinking. In this approach, an interpretation is suggested for the use of Mind Map, based on the relative importance weight of terms. Preliminary experimentation on a Medical corpus (CLEF2009) showed the accuracy of our approach.

KEYWORDS

Information Retrieval, Query Formulation, Idea, Associations, Mind Map

1. INTRODUCTION

There are two main components manipulated by an Information Retrieval System (IRS): the user query which is expected to traduce faithfully the user information need and the documentary corpus from which the IRS selects relevant documents [1].

According to a study in 2001 [2], the average query length in the Excite search engine is 2.6 words. A Report in the beginning of 2010 [3] done by the search engine Google, asserts that the length of 54.5% of queries in Google is greater than 3 words. In a 2013-2014’s period, the multi-words queries (producing a click) constitute about 45.92% of all queries [4]. In Google Mobile Search, the average English query length in 2008 is between 2.44 and 2.93 words [5]. In 2012, a study about Mobile searches [6] indicates that average English query length is 3.05 words. These statistical results could reflect the ambiguity problem of users’ multi-words queries in Information Retrieval Systems.

Natarajan Meghanathan et al. (Eds) : ICCSEA, SPPR, VLSI, WiMoA, SCAL, CNSA, WeST - 2014
pp. 197–213, 2014. © CS & IT-CSCP 2014

DOI : 10.5121/csit.2014.4726

The key challenge that remains nowadays is finding ways to capture and integrate contextual information for solving the ambiguity problem of queries expression. This contextual information is important to ameliorate the assimilation of the user's multi-words query and the user information needs in general [7]. Several approaches handle the queries ambiguity by proposing three main solutions [7]: (i) user profiling and personalization (ii) query expansion (i.e. automatic expansion or by recommendation) and (iii) Relevance Feedback. These approaches are generally founding their techniques on an input: a bag of words query, to reach the user's idea of search which would lead to the output: the documents relevant to the user's idea.

Some critics [8,9] tend to say that search engines are becoming more and more intelligent, using these solutions, while the user is becoming more and more dependent and less intelligent. In fact for query formulation, users are dragging since years their unchanged way of expression: the bag of words.

The problem of bag of words expression is that it causes sometimes the deformation of a user information need. This can be due to two facts: (i) the user cannot express clearly the main words of his idea of search and the less important words (which are used to specify the context related to the idea of search) and (ii) the idea of the user is packaged in a linear form which does not express the nodes in the network of the user's mind. As an example for the second point, let Q be the query "Java course Graphical Interfaces swing". This query can be represented in the mind of the user as a related multi-nodes query and can be formulated by the sentence "A course for Graphical Interfaces in Java language using the Swing API". In the query Q, the connections between remembered words are absent, which can misrepresent the flow of the user's idea of search.

Confusions come to evolve concerning the relation between the bag of words query and the findings in Cognitive Sciences, involving questions such as: Is the bag of words form the best natural way to express the user's idea? Is it possible to find a more faithful and not philosophical way of expressing the internal user's idea? We contribute in this paper to answer these questions by studying the cognitive sciences' findings about information recall process in the human mind and comparing existing studies about user expression forms used in IR or in Cognitive Sciences. This is elaborated in order to select the best natural way for user query formulation.

In this paper, we start by presenting in the section two, the state of the art which contains: (i) a dashboard of several Cognitive Maps and (ii) researches in IR that integrate query graphs and/or weighted queries (representing levels of importance in a query). In the third section, we incorporate our contribution of query formulation by expressing user's mind. In the fourth section, the evaluation of the contribution is presented. This is followed by the outcome of this work and some motivations of further works in the section five.

2. RELATED WORK

According to [10], students nowadays are connected and surrounded by networks of information, but a few of them have developed tools of thinking which express their information in a sensed network of knowledge. In fact, in one hand, human ideas are expressed by a dominating way: the linear language. In the other hand, the mind is constantly and naturally trying to find connections between information. So, there is a gap between the *internal* networks in the human mind and the *external* linear way of human expression. The process of thinking about a domain involves reasoning through our minds' Cognitive Maps and finding a way through them [11]. Human beings use these maps unconsciously. The external modelling of these maps helps to understand a domain or a cognitive territory [11].

Following these general assertions about cognition process, in the next section, we explore the kind of maps that could replace the bag of words query.

2.1 Findings about Human Cognitive Maps

A map can be a generic term for several forms: trees, graphs, diagrams, networks, etc. In the past, Cognitive Maps defined by Tolman (1948) [12] are mental representations of physical places. Nowadays, Cognitive Maps are not defined only for physical environments but also for several mapping methods such as: *Causal Mapping*, *Concept Mapping*, etc. [12]. Another popular designation of Cognitive Maps used in Education and Learning is visual tools [13]. Visual tools are divided into three categories (see Figure 1) [14]: (i) *Brainstorming webs* which are dominated by Mind Mapping techniques (1970's), (ii) *Graphic organizers* (1980) which are very structured tools, called also task-specific graphic organizers, and (iii) The Thinking-Process Maps (shown in Figure 1) which contains two main approaches: (a) the *Conceptual Mapping* which is dominated by the Concept Mapping technique (created in 1960 and became famous in 1984) and (b) the *Thinking Maps* (1988) which combine the freedom of thinking of brainstorming tools and the structured aspect of graphic organizers.

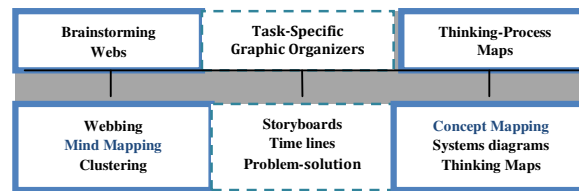


Figure1. Clustering of Visual tools

According to our study, *Thinking Maps* [14] and *Task-specific Graphic Organizers* [10], [15] are not considered convenient for IR query formulation this is due to these reasons: (i) Graphic Organizers structure the ideas for specific goals. In fact, the use of one of the Graphic Organizers is chosen according to some characteristics relative to the user idea (For example, to choose a diagram, the user has to know if the level of his thinking is superior to three or not, or if his idea is hierarchical or not, etc.) [10], [15]. (ii) Thinking Maps are not designed for vague or fuzzy thinking process. Each Thinking Map is designed for a specific task of thinking. As an example of Thinking Maps, the *Bubble Map* [13] is used to describe an object by defining its features, using adjectives.

Further, the *Cognitive Maps* which are considered in our search as potential maps for query formulation are *Mind Maps* and *Concept Maps*. Definitions and examples of these two last maps and their techniques are mentioned respectively in [16,17,18] and in [19]. In the following section, we introduce works in Information Retrieval, representing queries as graphs (Cognitive Maps, Conceptual Graphs of Sowa, etc.).

2.2 Specifying User Context by a Query Graph

Some researchers are trying to represent the query in a graph in the phase of the indexing process (graphs are invisible to the user) such as in [20]. The goal of these works is to integrate the semantic aspect both in the indexing process and in the correspondence process between documents and queries. In this state of the art, we introduce the researches representing the query in a graph in the phase of query formulation. We are interested by these researches since they change the user behavior from formulating his need in unrelated words (bag of words), to

constructing a graph which connects the query's whole idea and specifies more the user's search context.

2.2.1 "Googling from a Concept Map"

Some studies merge to the automatic query formation based on Concept Maps. In [21], the authors propose an automatic generation of Web queries from a user's Concept Map under construction. The goal of "Googling" based on Concept Maps is to provide to users supplementary information to add to their Concept Map while they are constructing it. The steps are: (i) the construction of a Concept Map by a user, (ii) the selection of a concept from the map by the user, (iii) the map is analysed to form automatically a textual query for information retrieval. The textual query is composed of key concepts. The key concepts are selected from the map to describe its topic. The difference between an ordinary search and a Concept Map based search is that the main concept of the query is searched through the web considering its semantic context. The Concept Map-based search showed ameliorated results comparing to the Single concept-based search (in terms of similarity and coverage between the retrieved documents and the query). The inconveniences of the use of Concept Maps for query formulation are mentioned further in the section 3.1.

2.2.2 SyDoM: a Multilingual IRS Based on Knowledge Management

In order to improve the documents content representation, the system SyDoM (Multilingual Documentary System) [23] employs a model of knowledge representation named Semantic graph (an extension of Sowa's Conceptual graph)[23]. This graph is used to represent both the user query and the documents existing in a multilingual corpus. The Semantic graph is wholly constructed by a semantic thesaurus [23] of a specific domain (e.g. Vehicle Mechanics). The main goal of using Semantic graphs in this system is to retrieve concepts from documents that are related in almost the same way they could be related in the query graph. Discussion about the limits of the use of Semantic graphs in query formulation is mentioned further in the section 3.2.

2.2.3 The Google's Wonder Wheel and Knowledge Graph

In a period from 2009 until 2011, the search engine Google proposed a tool named the Wonder Wheel [24]. The tool proposed expansions of the user query, visualized by a *Mind Map*. In this Mind Map, the user query is the central node, and the associated nodes represent the related ideas to the query proposed by the system. The goal of this tool is to guess the search idea of the user (For example: the user query "Mind Maps" has an associated idea: "Mind Maps tutorial"). In this work, the Mind Map is used as a visualization tool and not a way to express the user's own associations in mind. Further, Google has beenproposing since 2012 another associative concept for searching: The *Knowledge Graph* used for retrieving associations (data or facts) related to the user query and presenting the information in an aggregated way. These associations connect information which can be entities or attributes [25]. This last Google project seems to be oriented to queries containing one-concept (For example: "*Albert Einstein*") instead of an idea of search (which can contain many associated words and evolving eventually more than one concept). Also, this project offers *Knowledge associations* in order to enrich user's network of knowledge. We are interested in this paper, by *user's mind associations* in order to improve his attempt of query formulation.

2.2.4 The Pearltrees System

The concept Pearltrees is a collaborative and social bookmarking service. It proposes to the user to formulate and to store his interests in a Mind Map[26], also it allows the user to search for related

interests of other users. In a search phase, the user can either formulate his interest in a bag of words form or select a node(pearl) from his created Mind Map(named pearl tree). When a node is selected, it is considered as well as its sub-nodes in the search. The inconvenience of the Pearl trees system is that it proposes a *semantic clustering* tool for user interests. For example, let an interest I in a form of bag of words be “popular terms Information Retrieval” which can be traduced by the question “what are popular terms in the Information Retrieval domain?”. In the Pearl trees system, the clustered nodes of I would be as follows: a super-pearl: {Information Retrieval} and a sub-pearl {popular terms}. While it appears that this clustering avoids the *user importance degree*, since the main idea of search in the interest I is {popular terms} not {Information Retrieval}.

The problem of bag of words is the non-fidelity of representation of the network in the user's mind. In fact, there is no existence of nodes or relations. For example, the bag of words “definition concept thesaurus” could be in the mind of the user {“definition concept”, “thesaurus”}, where the node “thesaurus” is only used to specify the context of the query (“definition of a concept in a thesaurus”). So, the word “thesaurus” appears to be less important in the query than the central search idea of the user “definition concept”. A lack in a query bag of words format, is that there is no *importance degree* between used terms. Some works in literature integrate importance preferences in the query. We describe these relative works in the next section.

2.3 Adding Preferences in Queries: Weighted Query Languages

A body of research in *Fuzzy Boolean Information Retrieval*, since the 90's [27] until 2010 [28] has followed the postulation that “weighting query terms in a different way would lead to a more effective IRS” [28]. According to authors such as [29], the query formulation subsystem in IRS has to take into consideration the imprecision and uncertainty aspects of human communication. As a potential solution, adding user weights to the query could reduce the imprecision and clarify more the user information need, by mentioning what is important and what is less important in the query. As an example of numeric weighted query languages, the works of [28],[30] propose an enrichment of the Boolean query expression as follows:

$$Q = \langle t1, w1 \rangle OR \langle t2, w2 \rangle AND NOT \langle t3, w3 \rangle \quad (1)$$

Where: $t1, t2$ and $t3$ are search terms and $w1, w2$ and $w3$ are numeric weights in the interval $[0,1]$

A critic concerning user numeric weights holds that “a human being is more capable to qualify the importance of a concept, than to quantify it” [27]. An example of the *fuzzy linguistic approach* is given as follows [27]:

$$Q = \langle t1, I \rangle OR \langle t2, FI \rangle AND \langle t3, NVI \rangle \quad (2)$$

Where: I : Important, FI : Fairly Important, NVI : Not Very Important

More query weighting methods in *Fuzzy IR* exist in [29]. In vector space IR, manual query weighting terms were present in the Microsoft Index Server. For not encumbering the user by weights, some studies tried to *automatically* weight queries by guessing terms importance such as in [31] (this study is based on a cognitive hypothesis non certified by scientists or by experimentation). Other proposals of weighting query terms differently are predicting and estimating weights based on the corpus statistics [32] or documents relevance measures [33] or on the relevance feedback mechanism [34] or on the semantic relations between query terms [35]. We are mainly interested in this paper by the user importance of query terms while these last

works focus on the term importance in the query depending on the documents statistics and relevance or on terms' semantic similarities.

The related works in this paper, focused on two branches: (i) *Meaning making* of words in the query, by finding relations between words and conceptualizing words, using graphs and (ii) *importance of words* in the query, by adding weights. Some of these studies appear to have visible drawbacks while others have to be discussed further. In the following section, an explanation of the inconveniences of these last studies is proposed, where the inconveniences are related to the *query format* used in the IRS query formulation.

3. THE OUTCOME OF RELATED WORK

3.1 Concept Map Query

One of the differences between Concept Maps and Mind Maps is the use of linking phrases in the Concept Maps [19]. This means that the relation between every two nodes (concepts) in the Concept Map should be specified (i.e. labeled). However, in the Mind Map the relations are not specified by the user, they are abstract association in the mind of the user. In Information Retrieval, if we imagine the cognitive situation of a user who is specifying his information need, certain questions would evolve: (i) would this user be able to specify the kind of the connection between the words he is remembering? (ii) Can this user define his words of search as a relation (e.g. let the query: "The goal of development of diesel". "Goal" could be a relation between "development" and "Diesel") or as a term (or a concept)? Considering that the user is confused in the step of query formulation, it would be easier to mention only the words remembered and extracted from his mind naturally by association. Connections between the words in the brain of the user are interesting because they could define better the user thinking. In fact, as Tony Buzan expressed "without connections, thinking would not even exist" [36]. These connections should not be labeled (causality, composition, "is used", "is created by"...) since the user is formulating an idea not knowledge. Knowledge is acquired after learning, it expresses meaningful, solid and organized thoughts [19]. However an idea of search can be a messy or faulty thought. The Concept Maps are knowledge models, they could not be dedicated only for the purpose of query formulation.

3.2 Semantic Graph Query

Like a Concept Map, a semantic graph has concepts and relations between concepts, which would involve the same inconveniences as in the section 3.1. Also, the query formulated by the semantic graph in SyDOM does not mention any *level of importance* between concepts used in the query. It only mentions the kind of relations between concepts.

The outcome of the state of art suggests a query formulation where there are: (i) the *exploitation of the links* happening between remembered terms during user query formulation (ii) the non-labeling of links in the query and (iii) the need of discriminating terms according to their *importance for the user*. A user system is generally reticent to new ways asking to add supplementary information. For example, for IRS query, the user could be reticent to adding manually importance weights for the words in the query, etc. In order to reduce this reticence, a simple and faithful way of idea extraction from the user mind has to be suggested. We propose in the next section a modeling of queries based on Mind Mapping technique. This modeling technique is justified by its convenience with cognitive sciences' assertions about the information recall process from the human mind.

4. OUR APPROACH: MIND MAP QUERY

The information need is relative to a knowledge gap or to an incompleteness or inadequacy situation [37]. As a first step, an individual starts by searching in his internal memory (human memory), how to express this need. Then in a second step, he searches the missing knowledge in an external memory (documents). Considering the first step, cognitive sciences should be explored and used in a way that would accommodate the natural user's mind behavior in the process of query formulation.

4.1 Theoretical Foundations

The information representation in the brain is very complex. Many researches tried to simulate the knowledge representation and the human reasoning. We present in the following, the theories of thinking in the Brain, memory and mind that are agreed by the majority of cognitive scientists.

4.1.1 The Brain: a Detector of Models

Words and concepts are not saved in the human brain in an isolated way. The human brain models and constructs mental maps of information. Human brain organizes the raw data in schemes [17]. It detects models (i.e. structures that organize information) through certain habits of mind. It generates inductively new mental models of knowledge or concepts (i.e. a unit of knowledge) by interconnecting information [17]. According to Monroe and Pendergrass [38], the brain actively joins ideas together through the neurons networks, it deletes some information, joins others in associative schemes or it modifies the existing structures to give sense to new information in the brain. When the brain is confronted to a list of information, it is always trying to reconstruct unconsciously pieces and networks of information in a multitude of overlapped models [17]. This complex organism is partially directed by *sequential processes*. Cognitive sciences studies state that the mind contains linear and procedural knowledge, but the foundation of knowledge going from basic facts to decision making consists of non-linear models [13].

4.1.2 Human Memory: the Association Mechanism

According to Norman and Bobrow[38], the extraction of information from the brain is often triggered by matching a context with a saved concept in the human memory. The remembering of a concept will trigger the recall of other associated concepts. This process is called by Meyer and Schvaneveldt[38,39], the *spreading activation* mechanism (see an example in the Figure 2).

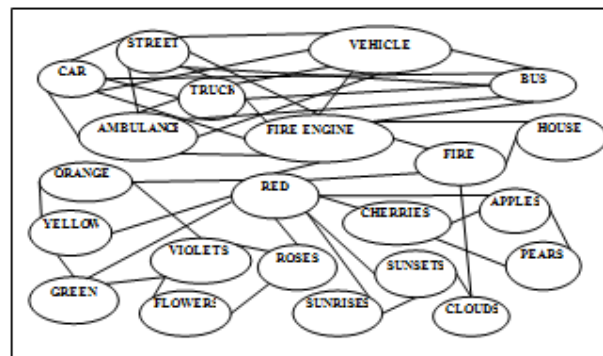


Figure 2. A stereotype of memory concepts representation [39]

The human memory can be divided in two kinds: The *episodic* and the *semantic memory*. According to Eric Jensen, the semantic memory is activated by mechanisms of association, similarity or contrast [17]. According to Capra [17], models existing in the human mind cannot be measured. They have to be extracted out of the mind, by mapping them. Cognitive scientists tried to extract from the mind several kinds of models: *Descriptive models*, *episodic models*, *causal models*, *sequential models*, *generalization/principal models*, *conceptual models*[40]. There are plenty of theories about these mental models in Cognitive Sciences defined in [40, 41] such as the *double-coding theory*, *connectionist models theory*, etc. We did not adopt one of these theories because they are argued by cognitive scientists. We consider in our theoretical background two phenomena: the detection of models by the brain, and the spreading activation of associations in the human memory. In the following section, we introduce our proposal of query formulation using these two cognitive findings.

4.2 Query Expression by User's Idea

An *Idea* is the result of the mind reflection process. It is a Greek word coming from the verb “*idein*” meaning “to see”. An idea is then a tool for the perception of the mind [42]. A user's idea about a subject can change through time, since his mind can omit some details and remember others about the same subject at a specific time. This does not necessarily mean that the knowledge has been modified in the user's mind, but that the *mind recall process* depends on the user's situation and information need. Unlike *knowledge*, an idea can express the information differently according to the user, the situation and the information need.

The query formulation proposed in this paper, aims to make a link between the human brain and Information Retrieval, in order to offer to the user a deeper way for expressing his idea, and potentially reducing the *confusion* and the *omission of words* that the user faces in query formulation.

Our contribution is divided in two steps: (i) a Mind Map query formulation by the user and (ii) an internal representation of the Mind Map query by the IR system.

4.2.1 Query Formulation by Mind Maps

This step consists in representing the user query by a Mind Map. It allows the user to construct his own associations of ideas. The reasons for choosing Mind Maps are the following:

- *The association aspect*: The nature of links between ideas is not identified by users (no obligatory semantic hierarchy, no labeling of links). The links represent simply the associations between ideas that happen in the mind of the user while thinking.
- *The graph aspect*: It represents the radiant aspect of the brain, considering that the brain does not remember the ideas in a list [16].
- *The relative importance of terms*: In the thinking process, the user could accord level of importance to his thought: (a) words representing the central idea and (b) words coming by associations from the mind, which could be important or not very important. The words which are important are mentioned near to the central idea and the others are mentioned far from it.

We illustrate in the following examples, these three different aspects.

Illustration 1: In this example, we illustrate the necessity of the migration from the “bag of words” approach to the Mind Maps modeling approach. Let a user information need: “*The definition of a*

concept in a thesaurus or by standard norms". The bag of words query in IRS would be for example: "definition of concept thesaurus standard norms".

The problems are as follows:

- The system cannot guess the nodes in the brain of the user. Furthermore, for an IRS using techniques of *popular terms* reweighting, or *high weighted terms* in documents, the first documents retrieved would consider more these terms in the query. So, if we assume that "thesaurus" is popular or high weighted term in documents, then the first documents retrieved would deal more with the term "thesaurus" than "definition of concept". The *user importance degree* is absent in the query. The user information need is poorly expressed and controlled by one aspect of information need expression: the terms.
- Considering that in the classical IRS the terms in the query have all the same importance, the user could hesitate which term to add to the query among many others.

We illustrate a representation of the query by a Mind Map (see Figure 3), which tries to solve the problems mentioned above. The benefits of the use of Mind Map query are as follows:

- The associative aspect of terms allows the user to mention the words that comes to his mind with a new dimension of query control: The terms' weights of his idea. In fact, as a contrast with a linear textual list, a Mind Map represents the relative importance of different ideas (by computing their distance with the central idea and the heights between each other).

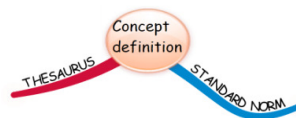


Figure 3. An example of Mind Map query

- The system understands that a user considers a term as not very important in the query. So, the user can mention the remembered terms without worrying about a misunderstood or a loss of his central idea of search through the query words.

Illustration 2: Let an information need of a query be "a good java course which describes inheritance and polymorphism, but also contains other notions of java. This course should contain exercises". If this query is formulated in a bag of words "good java course inheritance polymorphism exercises", we are as a user gambling with the IRS so it understands the terms (or concepts) importance in the query. In fact, the IRS can return java courses that describe only "polymorphism and inheritance" while the Information need is about a "good java course" not only "a java course about inheritance and polymorphism". The information need can be formulated by a Mind Map such as in the Figure 4.

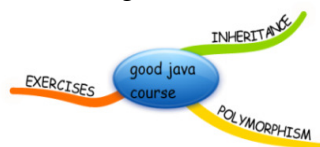


Figure 4. A Mind Map example

Illustration 3: Let the idea of search "the definition of a semantic resource, for example thesaurus, ontology". The user in this query specifies that he is searching for documents that define "what is

a semantic resource?” and mentions some examples of semantic resources to get closer to documents that deepen the subject (see Figure 5). If this query is mentioned in a bag of words, the user would probably avoid mentioning the keywords “ontology” and “thesaurus” that comes to his mind as examples and would formulate instead the query “semantic resource definition”. The user would avoid expressing his whole idea in a bag of words form, in order to diminish IRS misunderstanding.

We present in the following section the internal interpretation of Mind Maps queries, by defining the formula used for nodes levels in the user’s idea of search.

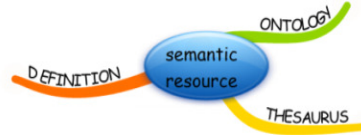


Figure5. A Mind Map example

4.2.2 The Internal Semantic for Mind Map Queries

In classical IRS, the number of occurrences of a term in a query is equal to 1. The weighting of query terms is uniform. These weights reflect the absence of term importance for the user. In our contribution, we consider the user terms’ importance by a graph with levels of importance: The more a remembered idea is far from the central idea of search, the less the term is important. In a quantitative way, terms in the first level of the Mind Map are considered σ times more important for the user than associated words in the second level. The weights of terms (w_i) in the graph are traduced in the IRS by this proposed formula :

$$w_i = \sigma^{(h-p_i-1)} \times a \quad (\text{Error! Bookmark not defined.})$$

Where :

σ : The power of importance between levels ($\sigma > 1$)

a : The weight attributed to the leafs of the graph (it could be Boolean or equal to TF or another weighting formula such as in [32])

p_i : The depth of the node n_i in the query graph

h : The height of the query graph

The proposed measure (3) could be classified in the relative importance semantic of query weighting formulas [28]. The calculus of this measure is illustrated in the following example.

Illustration 4: Let an information need: “Documents about Precision measure in Information Retrieval, for example: GMAP, MAP”. It could be expressed by a Mind Map query (see Figure 6).



Figure6. A Mind Map example

The four nodes (n_i) of the query are:

n_1 : Precision, n_2 : MAP, n_3 : GMAP, n_4 : Information Retrieval. Their respective weights are w_1, w_2, w_3, w_4 . They are calculated as follows, where we suppose that $\sigma = 2$ and $a = TF$ (Term frequency):

$$w_1 = 2^{(2-0-1)} \times TF = 2 \times \left(\frac{1}{5}\right); w_2 = w_3 = w_4 = 2^{(2-1-1)} \times TF = \frac{1}{5}$$

In this example, the weight of “Precision” term is twice important than the weights of “MAP”, “GMAP”, “Information Retrieval” terms.

To summarize, our approach offers these advantages:

- The *indirect* expression of term *importance* by the user
- The expression of the information need according to the *mind functioning* (spreading activation mechanism): For some cases, the user may not be able to define exactly his need but he may have information (related words) which could help in better expressing himself. So, the user can use associative terms (such as related examples or field of study) to get closer to his need. With the bag of words approach, the specification of associated terms to the central need is often misleading, because the IRS could not probably guess the central need and the related words.
- The expression of the user’s idea in a more structured format than the bag of words, which can facilitate the user thinking process while formulating his need.

To prove the accuracy of our approach, we present in the following, experimentation that we elaborated on a Medical collection test.

5. EVALUATION

We have conducted experiments in order to see the impact of Mind Map query on the retrieval accuracy. In order to fulfill this aim, we implemented into a classic retrieval component, the query weighting formula (see formula (3)) traducing the Mind Map in the Retrieval process.

5.1 The Experimental Environment

5.1.1 The Test Collection Features

We elaborated experimentation onto the test collection CLEF2009 (Cross Language Evaluation Forum) within the medical image retrieval track [45]. The corpus collection test contains 74’902 images from 20’000 English journal articles in Radiology. There are 25 queries in the collection test. Both queries and documents are composed of images which are described by an XML text caption. Following our purpose, we used only the text form. The collection CLEF2009 also proposes for the queries, three languages: English, German and French. We opt for the English language.

5.1.2 The IRS Implementation Features

We explain in this part, the specific parameters in the IRS with which the tests were elaborated. The internal model of knowledge representation used in the IRS is the *vector space model*. In the IRS, we used a semantic indexing. The indexing process was elaborated via the *MetaMap* analyzer which uses the UMLS Meta-thesaurus for concepts extraction. After the concepts extraction step,

the formula CF-IDF [46] is used for weighting the documents (queries) concepts. The correspondence between a document and a query is calculated with the cosine correlation measure. In our work, there were 16'514 indexed documents and 25 indexed queries from CLEF2009 collection. In our approach, we intervene in the indexing process only by changing the weights of queries concepts. So, the IRS will have only one changed parameter: The weighting formula of mind map queries.

5.1.3 Methodology of Evaluation

We have considered, in our experiments, Mind Maps with two levels: The first level expresses the central idea of search, and the second one contains associated ideas (it is possible to consider more than two levels in the Mind Map query, if future experiments on users demonstrate the need for it). Moreover, we make the supposition that each node in the Mind Map corresponds to one concept from the indexed query.

Considering that the queries in the test collection are in a linear form (bag of words), the counting of all the possible cases of Mind Maps is necessary. The result of this counting is a set of possible Mind Map queries. The similarity between documents and these possible queries is elaborated. The Mind Map query returning the best results in our approach is then compared to the results of the IRS using a bag of words query (Linear query).

In this experimentation, we tested the importance degree offered within Mind Maps according to hypotheses on the weighting formula (see formula (3)). We describe the hypotheses as follows:

Hypothesis 1: The importance weights of concepts in the query are different from each other, but the whole query weight remains the same as in the bag of words query. We assume that the variable a of the formula (3) is determined by the equation (4).

$$\sum_{i=1}^n w_i = 1 \quad (\text{Error! Bookmark not defined.})$$

Where: w_i is the weight of a node i in the mind map query and n is the number of nodes in the query

Hypothesis 2: A node in the query is σ times more important than another node in the query. Values of σ (min: 2, max: 5) are tested. In order to not lose the significance of a node in the Mind Map query ($w_i \rightarrow 0$), we opt for the maximum $\sigma = 5$. In fact, the more the value of importance σ increases the less the weights of the nodes with a depth $p_i = 2$ are significant especially if the number of nodes are important (n).

The number of experiments to be performed depends on the factors in the Information Retrieval System (IRS). The table 1 illustrates the factors to experiment and their low and high levels.

Table 1. The extreme levelsofthe IRS factors

Factors / Two Levels	Low Level	High Level
Power of importance (σ)	2	5
Height of the query graph (h)	1	2
Default weight of a node (a)	$\frac{1}{n}$	$\frac{1}{n \sum_{i=1}^n \sigma^{(h-p_i-1)}}$

We outperformed 3 experimental runs combining some of the factors and levels (see table 2). The two first experimental runs correspond to our Mind Map IRS and the last one corresponds to the classic IRS.

Table 2. Experimental runs of the IRS

Power of importance (σ)	Height of the query graph (h)	Default weight of a node (α)
2	2	$\frac{1}{n \sum_{i=1}^n \sigma^{(h-p_i-1)}}$
5	2	$\frac{1}{n \sum_{i=1}^n \sigma^{(h-p_i-1)}}$
2	1	$\frac{1}{n}$

5.2 Experimental results

To evaluate our approach, we used three measures: (i) The MAP measure for comparing the global performance of IRSs, (ii) the Precision-Recall couple for evaluating the behavior of IRS and (iii) values of precisions at different levels of documents retrieved (P@n) for evaluating the quality of the first n retrieved results.

We present the results of an IRS based on Mind Map query and an IRS based on bag of words query. In the 25 queries of the collection test, 7 queries are mono-concept (a Mind Map with one node) and 18 queries are multi-concepts. We focus on the results of the 18 multi-nodes queries of the collection test (multi-nodes Mind Maps).

Table 3. The global impact of Mind Map query

Queries set	Importance value σ	MAP		Δ MAP
		Mind Map queries	Linear queries	
18	$\sigma = 2$	0.2452	0.1801	+36.14%
	$\sigma = 5$	0.2542		+41.14%
25	$\sigma = 2$	0.2378	0.1909	+24.56%
	$\sigma = 5$	0.2443		+27.97%

We observe in the table 3 an augmentation of the MAP by 36.14% when the central node of the Mind Map is twice more important than the associated nodes and an improvement of the MAP by 41.14% when the central node is five times more important than the associated nodes. This global improvement of precision encourages the use of the proposed weighting formula of the Mind Map queries in an IRS.

Table 4. The average impact of the Mind Map query on the first documents retrieved (P@5)

Queries set	Importance value σ	P@5		$\Delta P@5$
		Mind Map queries	Linear queries	
18	$\sigma = 2$	0.4888	0.4333	+12.80%
	$\sigma = 5$	0.4555		+5.12%
25	$\sigma = 2$	0.536	0.496	+8.06%
	$\sigma = 5$	0.512		+3.22%

In table 4 and table 5, for the 5 and 10 first documents retrieved, there is an amelioration of precision for Mind Map queries. Due to this observation, modeling queries by Mind Maps could be employed in *precision-oriented systems* such as Medical IRS. We can see that the value of importance $\sigma = 2$ showed better results for the MAP and for the precision at the 5 first retrieved documents of the IRS better than the value $\sigma = 5$. However the value $\sigma = 5$ showed better results for the precision at the 10 first retrieved results than the value $\sigma = 2$.

Table 5. The average impact of the Mind Map query on the first documents retrieved (P@10)

Queries set	Importance value σ	P@10		$\Delta P@10$
		Mind Map queries	Linear queries	
18	$\sigma = 2$	0.4388	0.3722	+17.89%
	$\sigma = 5$	0.4666		+25.36%
25	$\sigma = 2$	0.5	0.4520	+10.61%
	$\sigma = 5$	0.52		+15.04%

Figure 7 shows the positive impact of modeling queries by Mind Maps, on the performance of an IRS. In fact, the curve of Mind Map approach is always higher than the curve of the classical approach either for the value $\sigma = 2$ or the value $\sigma = 5$.

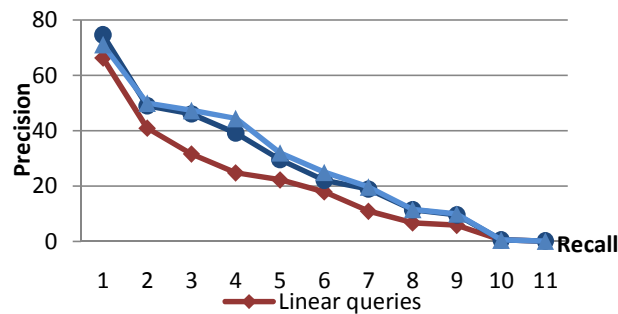


Figure 7. Precision/Recall curves (18 queries)

According to the results, the incorporated importance weights in the Mind Map queries proved to give better accuracy than the uniform weighting of bag of words queries (Linear queries). The

values of importance σ should be further tested and adjusted according to experiments on real users Mind Map queries in an IRS. Moreover, user studies appear to be fundamental in order to perceive the IRS users behavior during query formulation by Mind Maps.

6. CONCLUSION AND PERSPECTIVES

In IR query formulation process, the user tries to find the best way to express his need, by telling the remembered words. These words clarify the user's idea of search. This clarification could not be faithfully fulfilled if all words are specified in an unrelated way (bag of words). We suggested in this work to find a more faithful way for user's query formulation. We depicted from the Human brain's organizing: the *detection of models mechanism* and from the Human memory recall: the *associations' mechanism*. We suggested for IRS query formulation a cognitive map which satisfies these last mechanisms: a Mind Map. Also, we incorporated the importance degree in order to let the user express the important words and the less important ones in the Mind Map query.

Demonstrations on IR users should be elaborated in a further work, in order to study more the relation between associations (in human mind) and levels of importance in a query. Some critics to this work would be about guessing the association of the central idea of search by the system (not by the user). These critics are interesting but in our opinion, we need first to explore and understand better the expression of the users' search idea (by Mind Maps) before the deducing of the user's mind associations by the system.

REFERENCES

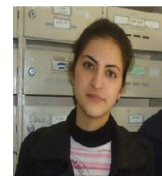
- [1] Göker, A. and Davies, J. (2009) Information Retrieval: Searching in the 21st Century. John Wiley & Sons Ltd.
- [2] Spink, A., Jansen, B., Wolfram, D. and Saracevic, T. (2002) From E-Sex to E-Commerce: Web search changes. IEEE Computer, 35, 3 (March. 2002), 107-109.
- [3] McGee, M. (2010) Google Weighs In on Query Length. Small Business Search Marketing. www.smallbusinesssem.com/google-query-length/3273/
- [4] Trellian. (2013) Keyword and search engines statistics. keyworddiscovery.com/keyword-stats.html?date=2013-01-01
- [5] Kamvar, M., Kellar, M., Patel, R. and Xu, Y. (2009) Computers and iPhones and Mobile Phones, oh my!: a logs based comparison of search users on different devices. In Proceedings of the 18th International Conference on World Wide Web (Madrid, Spain, April 20-24, 2009). WWW'09. ACM, New York, NY, 801-810.
- [6] Song, Y., Ma, H., Wang, H. and Wang, K. (2013) Exploring and Exploiting User Search Behavior on Mobile and Tablet Devices to Improve Search Relevance. In Proceedings of the 22nd International World Wide Web Conference (Rio de Janeiro, Brazil, May 13-17, 2013). WWW'13.ACM, New York, NY, 1201-1212.
- [7] Bhatia, M. and Akshi, K. (2010) Paradigm shifts: from pre web information systems to recent web-based contextual information retrieval. Webology, 7, 1 (Jun. 2010).
- [8] Sparrow, B., Liu, J. and Wegner, D.M. (2011) Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. Science 333, (Aug. 2011), 776-778.
- [9] Carr, N. (2010) What the internet is doing to our brains: The shallows. W.W.Norton & Company Inc., New York, NY
- [10] Hyerle, D. (2008) Thinking Maps : A visual language for learning. In Knowledge Cartography. Thinking Foundation.
- [11] Emberling, D. (2005) About Cognitive Maps. Developmental Consulting Inc. www.developmentalconsulting.com/pdfs/About_Cognitive_Maps_vA.pdf
- [12] Dagan, R. (2002) Cognitive Mapping. Intraspec. <http://intraspec.ca/cogmap.php>

- [13] Hyerle, D. and Piercy, T. D. (2010) Thinking Maps: The Cognitive Bridge to Literacy A Visual Language for Bridging Reading Text Structures to Writing Prompts. In Thinking maps : a new language for learning. Thinking Foundation, USA.
- [14] Hyerle, D. (2008) Thinking Maps: Visual Tools for Activating Habits of Mind. In Learning and Leading with Habits of Mind. ASCD.
- [15] Echanted Learning. Graphic Organizers. www.gmbservices.ca/Jr/GraphicOrganizers.htm.
- [16] Buzan, T. and Buzan, B. (1993) The Mind Map Book: How to Use Radiant Thinking to Maximize Your Brain's Untapped Potential. Plume, New York.
- [17] Hyerle, D. (2009) Visual Tools for Transforming Information into Knowledge. Corwin Press.
- [18] Topic Scape. Mindmaps Directory. www.topicscape.com/mindmaps
- [19] Novak, J. D. and Cañas, A. J. (2008) The theory underlying concept maps and how to construct and use them. Technical Report. Institute for Human and Machine Cognition (IHMC)
- [20] Boughanem, M. and Baziz, M. (2006) An IR model based on a sub-tree representation. In Leading the Web in Concurrent Engineering: Next Generation Concurrent Engineering, P. Ghodous, Ed. IOS Press, 450-457.
- [21] Leake, D., Maguitman, A., Reichherzer, T., Cañas, A. J., Carvalho, M., Arguedas, M. and Eskridge, T. (2004) "Googling" from a concept map : towards automatic concept-map-based query formation. In Proceedings of the 1st International Conference on Concept Mapping (Pamplona, Spain, September 14-17, 2004). 409-416.
- [22] Kwon, S. Y. (2006) The comparative effect of individually-generated vs. collaboratively-generated computer-based concept mapping on science concept learning. Doctoral Thesis. Texas A&M University.
- [23] Roussey, C., Calabretto, S. and Pinon, J.M. (2001) A Multilingual Information Retrieval System for Digital Libraries. In Proceedings 5th East European Conference on Advances in Databases and Information Systems (Vilnius, Lithuania, September 25-28, 2001). ADBIS'01. 98-111.
- [24] Mohidin, F. (2010) The Google Wonder Wheel and Mind Maps. Mind Map Tutor. www.mindmaptutor.com/2010/05/the-google-wonder-wheel-and-mind-maps/
- [25] PremiumSEO Solutions. (2012) Google Knowledge Graph: New search technique. www.premiumseosolutions.com.au/blog/seo-news/google-knowledge-graph-new-search-technique/
- [26] Chuck, F. (2011) Pearltrees extends its mind mapping and curation application to the iPad. The Mind Mapping Software Blog. mindmappingsoftwareblog.com/pearltrees-for-ipad/
- [27] Bordogna, G. and Pasi, G. (1993) A Fuzzy Linguistic Approach Generalizing Boolean Information Retrieval : A Model and Its evaluation. J Am. Soc Inform. Sci. 44,2 (Sept. 1992), 70-82.
- [28] Zadrozny, S. and Kacprzyk, J. (2009) On an interpretation of keywords weights in information retrieval : some fuzzy logic based approaches. Int. J. Uncertain. Fuzz. 17, 1 (Aug. 2009), 41-58.
- [29] Herrera-Viedma, E., Lopez-Herrera, A. G., Alonso, S., Porcel, C. and Cabrerizo, F. J. (2007) A Linguistic Multi-level Weighted Query Language to Represent User Information Needs. In Proceedings of the IEEE International Conference on Fuzzy Systems (London, UK, July 23-26, 2007). FUZZY-IEEE'07. IEEE, 1-6.
- [30] Herrera-Viedma, E., Alonso, S., Cabrerizo, F. J., Lopez-Herrera, A. G. and Porcel, C. (2007) A software tool to teach the performance of Fuzzy IR systems based on weighted queries. In Proceedings of the 1st International Workshop on Teaching and Learning of Information Retrieval (London, UK, January 10, 2007). TLIR'07.
- [31] Arif, A.S.M., Rahman, M.M. and Mukta, S.Y. (2009) Information Retrieval by modified term weighting method using random walk model with query term position ranking. In Proceedings of the International Conference on Signal Processing Systems (Singapore, May 15-17, 2009).
- [32] Singhal, A. (1996) Term Weighting Revisited. Doctoral Thesis. Cornell University.
- [33] Monz, C. (2007) Model Tree Learning for Query Term Weighting in Question Answering. In Proceedings of the 29th European Conference on IR Research (Rome, Italy, April 2-5, 2007). ECIR'07. Springer. 589-596.
- [34] Wang, B., Zhou, Y., Zhang, Q. and Xuanjing, H. (2011) Learning the Weight of the Query Term from the Relevance Feedback. In the 7th International Conference on Natural Language Processing and Knowledge Engineering (Tokushima, Japan, November 27-29, 2011). NLP-KE'11. IEEE. 43-50
- [35] Zheng, W. and Fang, H. (2010) Query Aspect based Term Weighting Regularization in Information Retrieval. In Proceedings of the 32nd European Conference on IR Research (Milton Keynes, UK, March 28-31, 2010). ECIR'10. Springer. 344-356.

- [36] McAdam, T. (2010) Maximize the Power of Your Brain Using Mind Mapping. Self Improvement Information. www.selfimprovementinformation.com/tony-buzan-maximize-the-power-of-your-brain-using-mind-mapping/
- [37] Ingwersen, P. (1996) Cognitive Perspectives of Information Retrieval Interaction : Elements of a Cognitive IR Theory. *J. Doc.* 52, 1 (Mar. 1996), 3-50.
- [38] Hickie, K.M. (2006) An Examination of Student Performance in Reading/Language and Mathematics after Two Years of Thinking Maps Implementation in Three Tennessee Schools. Doctoral Thesis. East Tennessee State University.
- [39] Collins, A.M. and Loftus, E.F. (1975) A spreading-activation theory of semantic processing. *Psychol. Rev.* 82, 407-428.
- [40] Marzano, R.J., Pickering, D.J. and Pollock, J.E. (2001) Classroom instruction that works: Research-based strategies for increasing student achievement. McRel, USA
- [41] Schraw, G. Knowledge Representation. (2011) www.education.com/reference/article/knowledge-representation/
- [42] Rockmore, T. (2010) Kant and Phenomenology. The University of Chicago.
- [43] Ceusters, W., Smith, B. and Goldberg, L. (2005) A Terminological and Ontological Analysis of the NCI Thesaurus. *Method. Inform. Med.* 44, 4 (2005), 498-507.
- [44] IT Informaters. Common Mind Maps. www.informationtamers.com/WikIT/index.php?title=Common_mind_maps
- [45] ImageCLEF. ImageCLEF 2009 medical retrieval task. imageclef.org/2009/medical
- [46] Goossen, F., Ijntema, W., Frasincar, F., Hogenboom, F. and Kaymak, U. (2011) News Personalization using the CF-IDF Semantic Recommender. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics (Sogndal, Norway, May 25-27, 2011). WIMS'11.
- [47] Ussery, B. (2008) Google–Average Number of Words Per Query have Increased!. Bleu Blog. www.beussery.com/blog/index.php/2008/02/google-average-number-of-words-per-query-have-increased/

AUTHORS

Rihab Aayed is a first year PhD student in Tunisia Polytechnic School, University of Carthage, Tunisia, actually specialized in Graph search issues in Information Retrieval Systems. She graduated in 2009 with a 'Computer Science Applied on Management' Bsc degree, and an 'Information Systems' Msc degree in 2011. She studied during her Msc thesis the most natural way to convert the human mind idea to a query understandable by Information Retrieval Systems.



Farah Harrathi is a post doctorate researcher specialized in Information Retrieval and more precisely in Semantic and Multilingual Indexing. He obtained his PhD degree in 2009 from the INSA Lyon, France. He has been an associate professor at Higher Institute of Multimedia Arts of Manouba, University of Manouba and visitor professor at the Faculty of Sciences of Tunis and at the Faculty of Sciences of Gafsa, Tunisia.

Mohamed Mohsen Gammoudi is currently a full Professor at Higher Institute of Multimedia Arts of Manouba, University of Manouba, Tunisia. He is responsible of SCO-ECRI team in Research Laboratory RIADI. He obtained his habilitation to supervise research in 2005 at the Faculty of Sciences of Tunis. He got his PhD in September 1993 in Sophia Antipolis Laboratory I3S/CNRS. Professor Gammoudi's professional work experience began in 1992 when he was assigned as an assistant at the Technical University of Nice. Then he was hired as a visiting professor between 1993 and 1997 at Federal University of Maranhao, Brazil. Since, he has supervised several PhD and master theses.



Mahran Farhat is currently a PhD student at the Faculty of Sciences of Tunis (FST), he is a member of Laboratory RIADI. He obtained his master in 2011 at the Faculty of Sciences of Tunis (FST). He started his PhD in 2012 under the leadership of Professor Mohamed Mohsen Gammoudi. Currently, he served as a contractual assistant professor at ESEN, University of Manouba.



INTENTIONAL BLANK

REAL-TIME DETECTION OF PHISHING TWEETS

Nilesh Sharma¹, Nishant Sharma², Vishakha Tiwari³, Shweta Chahar⁴,
Smriti Maheshwari⁵

¹Software Engineer at Dubizzle, Dubai, UAE
nilesh.sharma7@gmail.com

²Integrated Dual Degree (B.Tech+M.tech), Electronics & Communication
Department, Indian Institute of Technology, Roorkee, India
nishantsharma.iit@gmail.com, nish5uec@iitr.ac.in

^{3,4,5}B.Tech, Computer Science Department,
Hindustan College of Science & Technology, India
vish.twr26@gmail.com
chaharshweta@ymail.com
stellarsmriti19@gmail.com

ABSTRACT

Twitter is an immensely popular social networking site and micro blogging service where people post short messages of 140 characters called tweets. Phishers have started using Twitter as a medium to spread phishing scams because of the fast spread of information.

We deployed our system for end users by providing an easy to use “Web framework” which takes the tweet id and the specific keyword, and in return it will give tweets indicating legitimate or unsafe. We have a green background indicating a legitimate or safe URL and red symbols indicating the malicious or phishing URL with the help of APIs and machine learning algorithm.

KEYWORDS

Phishing Detection, Python, Machine Learning Algorithm, Twitter, Web framework

1. INTRODUCTION

Phishing is the act of attempting to acquire information by masquerading as a trustworthy entity in an electronic communication. Twitter, due to its large audience and information reach, attracts Spammers.

There has been an increase in phishing attacks through social media due to ease spread of information on social networks. Such a rise in phishing attacks on social media presents a dire need for technological solutions to detect these attacks and protect users from phishing scams.

Detecting phishing on social media is a challenge because of (i) large volume of data – social media allow users to easily share their opinions and interests and large volume of data make it difficult to analyse (ii) limited space – social media often impose character limitation (such as Twitter’s 140 character limit) on the content due to which users use shorthand notations. Such shorthand notation is difficult to parse since the text is usually not well-formed; (iii) fast change – content on social media changes very rapidly making phishing detection difficult; and (iv) Shortened URLs – researchers have observed that more than half of the phishing URLs are shortened to obfuscate the target URL and to hide malignant intentions rather than to gain character space.

Recent statistics show that on an average, 8% tweets contain spam and other malicious content .It has been estimated that in the Kaspersky Lab report 37.3 million users experienced phishing in 2013 [1]. Also, 45 % of bank customers are redirected to a phishing site divulge their personal credentials.

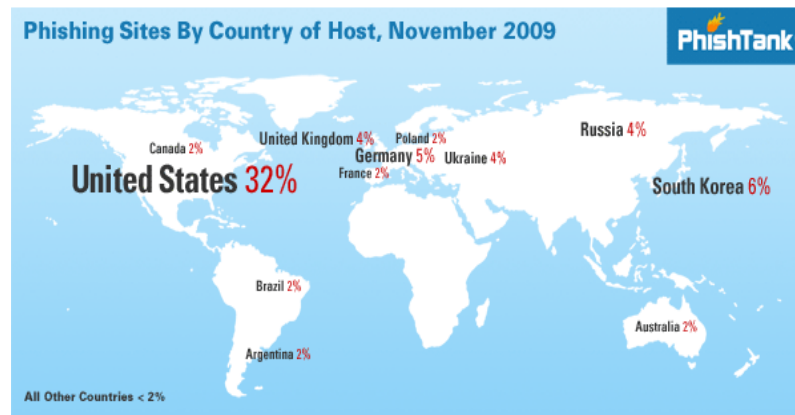


Figure 1:Phishing sites by country of host, November 2009 [2]

In our paper, we have developed a “Web Framework” in Python using Django through which we put a tweet id from Twitter as input, which processes as HTTP request for all the tweet with URLs and HTTP response as a result with legitimate and phishing URLs. We have been using several APIs integrated with the framework.

Along with these APIs we have also created another database for the “new URL” which has not been found in any of the above API. For classifying this “new URL” we use machine learning algorithm. The whole system is user friendly as user can easily input the tweet id or any keyword and we can get all the URLs with classification.

Further, this “**Phishing Detection System**” is time efficient, taking an average of only **0.501** seconds to detect phishing tweets with high accuracy of 94.56%. Such low computation times make it ideal for real world use.

2. NEED OF REAL TIME PHISHING DETECTION

Phishing is a harmful form of spam. These Phishing attacks not only cause the leakage of confidential information, but it also results in a huge amount of monetary losses [3]. Hence, it is important to build a realtime phishing detection mechanism for every OSM to protect its users.

As there is an increase in phishing attacks so we have to deploy a system in which we have built a Web framework for finding the tweet as phishing or safe.

In April 2013, an AP journalist clicked on a spear phishing email disguised as a Twitter email. The phisher, then hacked AP's Twitter account. Stock markets plunged after a phony tweet about an explosion at the White House, erasing \$136.5 billion of value from the S&P 500 index [4]. A most preferred solution used for Twitter by Lee et al. is the Warning-Bird system whose main focus is on suspicious URLs in general but not on detecting phishing. However, Warning Bird may fail if the spammers use a short redirect chain or multiple page-level redirects [5].

There is PhishAri technique which also works in real time, but now it is not in use because the extension could not work for the PhishAri API as now to access the Twitter data we have to firstly have authentication from Twitter using oauth and due to the appliance of various security parameters allowable to Twitter make it difficult to use [6].

After reviewing the above techniques, it was evident that there was very little work done to detect phishing on Twitter in real-time. To fill this gap, we designed and developed a “Web framework” to identify the particular tweet is phishing or safe.

3. TWYTHON

Twython is primary python wrapper for Twitter API, so that we can access the Twitter data easily and supports both normal and streaming Twitter APIs. In other words, Twython is used to query Twitter using the Representational State Transfer (REST) web API to get incoming replies and direct messages.

Twython, has two main interfaces:

- *
- * Twitter's Core API (GET statuses/sample request)
- * Twitter's Streaming API (GET users/lookuprequest)

Search API is used for finding the tweets of a user, finding the tweets with specified keywords. We have a large number of queries, for that purpose Twitter streaming API is used. The RESTful API is useful for getting things like lists of followers and those who follow a particular user, and is what most Twitter clients are built off. In order to allow Twitter to monitor the number of requests we make, we need to follow an authentication protocol, OAuth [7].

Twitter API version 1.1 uses the concept of oauth, in which we need an authentication key whereas in version 1.0 we can access the data of Twitter directly [8]. This provides us the facility of making chrome extension using PhishAri API, which in version 1.1 is not possible.

```
APP_KEY = 'YOUR_APP_KEY'
APP_SECRET = 'YOUR_APP_SECRET'
TWITTER = Twython (APP_KEY, APP_SECRET)
Oauth = TWITTER.get_authentication_tokens(callback_URL='http: //abc.com/callback') [9]
```


4. API INTEGRATION

After getting the permission from Twitter using oauth, web frame work is integrated with several APIs which are working in the background to detect whether the extracted tweets with URL are phishing or not. The APIs used are Phish tank API, Google safe browsing API, Mywot API. These APIs check for the phishing URLs using their own realtime database.

As new phishing techniques contain the short URLs, we can't detect them directly. LongURL API converts the short URLs from long URL, with the extra information added to it [10].

An addition tab “MORE” is used to load more tweets. It takes some time, to overcome with this limitation, we have used the concept of caching. We can store user requested tweets in cache memory, and checks for the update so that the data in cache memory on client site doesn't become outdated.



Figure2: Web framework

Web frameworks also gives the facility of searching the tweets containing a specific keyword along with the specified user. Two text buttons are used for that purpose, and then GET request is sent to the Twitter using Twython, which helps us in accessing the user data, JSON type request is returned. “check_phish.py” file checks for the phishing URLs and gives the result. It shows a red background for phishing URLs and a green background for good URLs, and tweets which do not contain any URL have white background. Database of phishing URLs is stored, which is also referred whenever processing is done.

The machine learning algorithm we used is:

Random Forest- It is the most effective method of machine learning algorithms, it is a collection of CART-like trees specific rules for tree growing, tree combination, self-testing Trees are growing using binary partitioning[11]. This chooses some important set of features which makes it more accurate for classification[12].

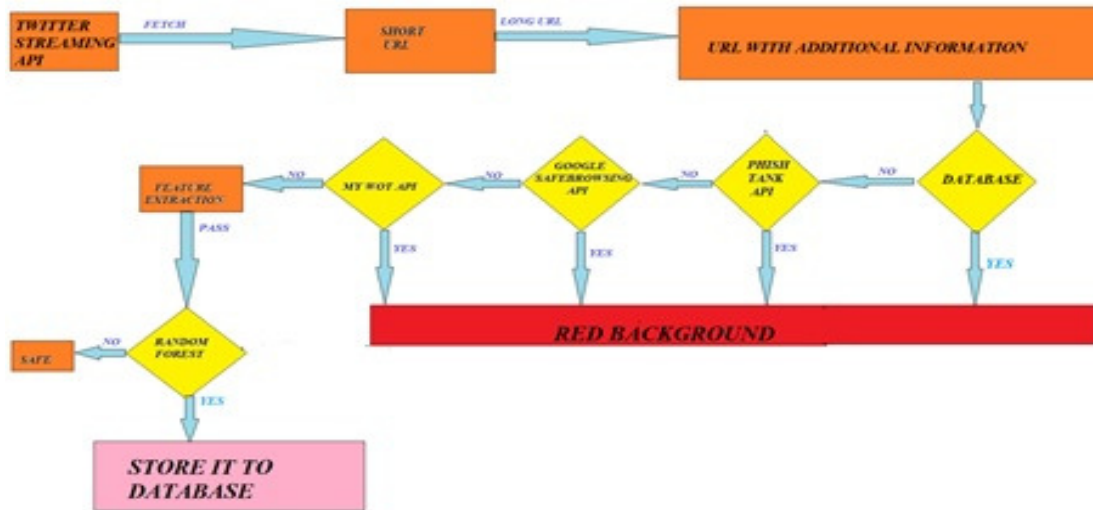


Figure 3: Flow Chart

As figure 3 specifies how a URL is classified as a phishing or safe URL. Tweets are fetched from the Twitter using Twitter Streaming API. LongURL API is used to get the long URL of the shortened URL, which give URL with additional information. It is first checked with the database, if it exist in database, it is labelled as phishing URL, if not then it is further send to different APIs, which check the URL in their database and gives output. Then, final verification is done by machine learning algorithm.

5. FEATURE EXTRACTION

5.1 Features for phishing Detection

Phishing is like a plague in social media .Past studies show that the phishing website can be detected through the analysis of the URL and the content of the website. Phishing websites, often appear identical to the legitimate website, but will generally have one or more characteristics by which we can find out that site are phishing. However, from the past studied it has been observed that the malicious users keeps changing the techniques they used for Phishing, making detection more difficult [13].

Database of phishing URL is used to extract more features and new URLs are classified accordingly. In short, it is learning from previous data.

5.1.1 URL based features

URL based features are defined for the analysis of the suspicious website. The length of the URL, no of dots, length of domain and subdomain, spelling, position of slashes used in the URL. These are some of the common features that help in detecting phishing websites (Table 1). The length of the URL of the phishing websites is normally longer than the legitimate website. A phishing URL contains more number of dotes and sub domain than legitimate [14].

Table1: URL based feature

Feature	Description
Length of URL	Length of expanded URL in number of characters.
Number of dots	Number of dots (.) used
Number of subdomains	Number of subdomains (marked by /) in the expanded URL
Number of Redirections	Number of hops between the posted URL and the Landing page
Presence of conditional redirects	Whether the URL is redirected to different landing page for browser or an automated program
Spelling	use of “l” instead of “I” in URL
Slash	Number of (/) used

5.1.2 Tweet based features

Phishing tweets are designed in such a way so that they can get high visibility by carefully using tags. Twitter specific features are tweet content and its characteristics like length, hashtags, and mentions tags. Other Twitter Features used are the characteristics of the Twitter user posting the tweet such as the age of the account, number of tweets, and the follower-follower ratio (Table 2).

Table 2: Twitter based feature

Feature	Description
Number of @tags	Number of Twitter users mentioned in tweet
Presence of trending #tags	Number of topics mentioned which were trending at that time
Number of RTs	Number of times the tweet was reposted
Length of Tweet	Length of tweet in number of characters
Position of #tags	Number of characters of tweets after which the #tag appears

5.1.3 WHOIS based feature

WHOIS is a TCP based transaction-oriented query/response protocol that is widely used to provide information services to Internet users. It is widely used for querying databases that stores

the registered users or assigns of an internet resource, such as domain name, an IP address block (Table 3). Most malicious users register domains of websites from the same registrar, hence tracking the registrar may aid in detecting phishing. Therefore, we use WHOIS based features to further enhance our phishing detection methodology.

Table 3: WHOIS based feature

Feature	Description
Ownership period	Age of the domain
Time taken to create TwitterAccount	How much time elapsed between creation of domain and the Twitter account

6. RESULT

As we developed a phishing detection system which uses several APIs as well as some features like URL based and Twitter based features to classify tweets accordingly as phishing or safe. In the next step, we create a real time phishing detection system by deploying a Web framework which makes a call to different APIs like Google safe browsing, Web of Trust API, etc. and then marks each tweet as phishing or safe.

In this section, we elaborate the results and observations based on the classification mechanism using the five set of feature sets with the database of phishing URLs. We have implemented random forest algorithm which learns from database.

6.1 Evaluation Metrics

In order to evaluate the effectiveness of our classification method based on the features described, we use the standard information retrieval metrics viz. accuracy, precision and recall. Precision of a class is the proportion of predicted positives in that class that are actually positive. Recall of a class is the proportion of the actual positives in that class which are predicted positive.

Table 4: Results of Classification experiment. We observe that Random forest has the best accuracy of 94.56%.

Evaluation Metric	Random Forest Algorithm
Accuracy	94.56%
Precision (phishing)	96.24%
Precision (Safe)	98.23%
Recall(phishing)	93.21%
Recall(safe)	96.54%

Each entry in the table 5 indicates the number of elements of a class and how they were classified by our classification method. For example, ‘TP’ is the number of phishing tweets which were correctly classified as phishing. Using this confusion matrix, we can compute the precision and recall for both ‘phishing’ and ‘safe’ classes.

We also use the confusion matrix to compute the overall ‘accuracy’ of the classifier [15]. It is the ratio of the correctly classified elements of either class to the total number of elements.

$$\text{Precision phishing} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall phishing} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

6.2 Classification Results

We now describe the results of our classification experiment as described. We use the classification method for our study which is Random Forest. We present the results of classification task using all these methods.

From the 1,689 phishing tweets, we found that 1,573 tweets had unique text. Therefore, is our true positive dataset, we consider these 1,573 phishing tweets and 1,400 safe tweets chosen randomly from the tweets marked as ‘safe’ during our data collection process. We use this dataset for the rest of our classification experiments. We found that Random Forest classifier works best for phishing tweet detection on our dataset with a high accuracy of 94.56%. We also obtain a recall of 94.21% for phishing class and 95.82% for safe class. The results from the classification technique are described in the table 5.

We find that the superior performance of Random Forest for phishing detection on TWITTER also holds true with a high accuracy.

In table 5 we show that we could detect 94.31% phishing tweets correctly. However, we misclassified 8.5% of legitimate tweets as phishing tweets. The false negative percentage is low indicating that we classified only 6.78% phishing tweets as legitimate.

Table5: Precision and recall for phishing detection using Random Forest based on all six feature sets

		Phishing	Prediction Safe
Actual	Phishing	94.31%	8.5%
Actual	Safe	6.78%	95.41%

6.3 Evaluation of various Feature Sets

Most of the previous studies to detect phishing have used features based on the URL of the suspicious page and the HTML source of the landing page. In this study, we propose to use Twitter based features along with URL based features to quickly detect phishing on Twitter at zero-hour.

As described we have used six sets of features in table 6. To evaluate the impact of each feature set, we performed classification task by taking one feature set at a time and then added the other one in the next iteration.

Table 6: Informative features which we found for phishing tweet detection using Random Forest classification

Feature Sets	Precision (Phishing)	Precision (Safe)	Recall (Phishing)	Recall (Safe)	Accuracy
F1	82.27%	88.72%	79.67%	92.35%	84.22%
F1+F2	87.11%	89.34%	82.89%	92.78%	89.35%
F1+F2+F3	92.21%	90.12%	85.29%	93.45%	91.18%
F1+F2+F3+F4	95.85%	92.35%	91.14%	94.35%	92.52%
F1+F2+F3+F4	96.21%	94.62%	92.34%	95.45%	92.87%
F1+F2+F3+F4+F5	97.85%	95.84%	93.56%	95.90%	93.15%
F1+F2+F3+F4+F5+F6	98.43%	96.21%	94.87%	96.56%	94.56%

We observe that when we use only URL based features, we get an overall accuracy of 84.22% and a low precision and recall for ‘phishing’ class. The addition of Twitter based feature sets, user based features and network based features significantly improve the performance of phishing detection and boost the precision of identifying phishing tweets significantly. Hence, Twitter based features are helpful in increasing the performance of classifying phishing tweets.

6.4 Most Informative Features

We now evaluate the most important features which help to decide whether a tweet is phishing or not. We use ‘scikit’ library to find out the most informative features. Random Forests deploy ensemble learning to evaluate the feature importance.

After each random tree is constructed using a set of features, its performance (misclassification rate) is calculated. Then the values of each features is randomly permuted (for each feature) and the new misclassification rate is evaluated.

The best performing features are then chosen as the most informative features.

The domains of malicious and phishing URLs tend to be short lived when compared to the domains of legitimate URLs in order to avoid detection. Similarly the age of Twitter account of the user posting phishing tweets is also generally less. Such users are often detected by Twitter and their accounts are suspended. However, using Phishing Detection System, we could detect a large number of phishing tweets by such users before they were suspended by Twitter[16].

6.5 Test Cases

Here are some test cases, to check the functioning of the webframework.

UserId- shwetachahar1

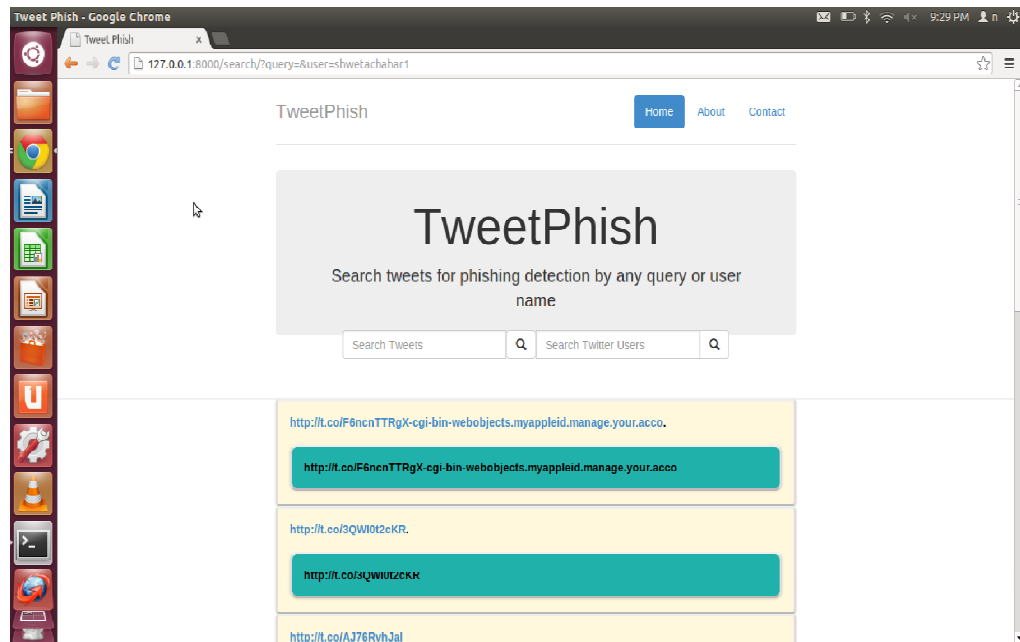


Figure 5: Screen Snapshot 1

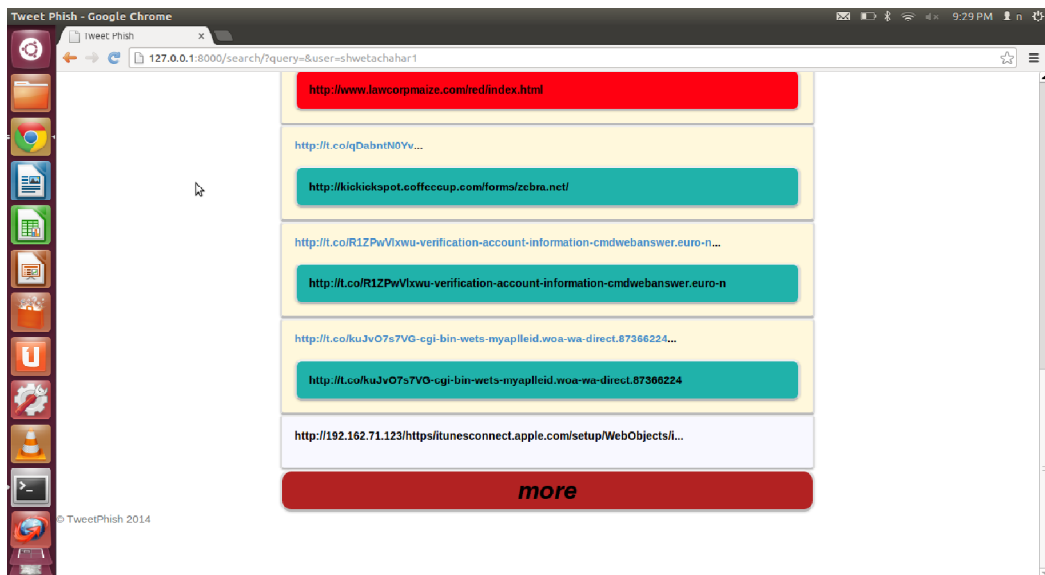


Figure 6: Screen Snapshot 2

Keyword- India

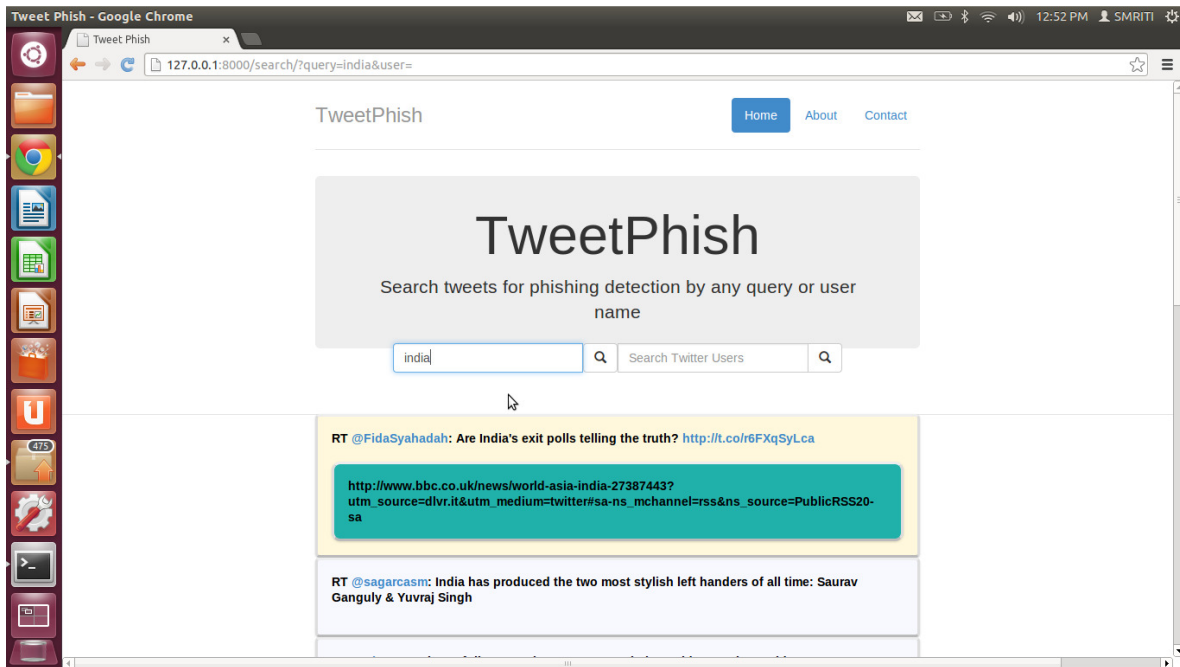


Figure 7: Screen Snapshot 3

7. FUTURE WORK

Now we discuss how we can further improve Phishing detection for more efficient and robust phishing detection.

- * Backend database for faster lookup: In future, we can maintain a cache backend database to capture tweets which have already been marked as either phishing or safe on Twitter. So, if the same tweet appears on Twitter, then we can skip the entire process of feature extraction and classification and lookup in our dataset of phishing URL and safe URL. This will also help us increase our own database of phishing tweets.
- * As future work, it would be interesting to evaluate other feature ranking and selection techniques such as principle component analysis, latent semantic analysis, chi-squared attribute evaluation, etc. and other feature space search methods such as greedy backward elimination, best first, etc.

8. CONCLUSION

This phishing detection realtime web framework allows its user to easily access the tweets using tweet id or using any specific keyword containing in tweets using various APIs. Twython is used for accessing the Twitter data using oauth. This may take some time, for that concept of cache memory is used. For shortened URLs there is the need of longURLs API. Integration of APIs and machine learning algorithm together gives us result with an accuracy of 95.56%. Time taken for detection is a maximum of 0.501 Sec for a tweet. Various features including WHOIS, tweet, network based are under main considerations. White background is used for

tweets without URLs, red background for phishing URL, and green background for safe URL. This method can be improved by using more advanced features and database.

ACKNOWLEDGEMENT

We are very grateful to everyone who has given their valuable feedback and suggestions. This project has been done as a final year B.Tech project.

REFERENCES

- [1] Press Releases, “Kaspersky Lab report”,
http://www.kaspersky.com/about/news/press/2013/Kaspersky_Lab_report_37_3_million_users_experienced_phishing_attacks_in_the_last_year
- [2] Stats from phishtank, “brainfoldb4u” <https://brainfoldb4u.wordpress.com/category/hacking/page/2/>
- [3] Victoria Lund-Funkhouser , “Top 7 Phishing Scams of 2013”, <http://blog.returnpath.com/blog/tori-funkhouser/top-7-phishing-scams-of-2013>
- [4] DanchoDanchev, “How many people fall victim to phishingattacks?”,
<http://www.zdnet.com/blog/security/how-many-people-fall-victim-to-phishing-attacks/5084>
- [5] Manju .C.Nair ,S.Prema (PhD),“A Distributed System for Detecting Phishing in TWITTER Stream”in International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 3, Issue 2, March 2014
- [6] AnupamaAggarwaly , AshwinRajadesingan , PonnurangamKumaraguruy , “Automatic Realtime Phishing Detection on TWITTER”, In seventh IEEE APWG eCrime Researchers Summit (eCRS), 2012
- [7] Hashtags and followers: “An experimental study of the online social network TWITTER Eva GarcíaMartínSchool of Computing”,Blekinge Institute of Technology, Sweden, Thesis no: 1MSC:2013-01 , September 2013
- [8] TWITTER, "Overview: Version 1.1 of the TWITTER API",
<https://dev.TWITTER.com/docs/API/1.1/overview>
- [9] Ryan McGrath, “TwythonDocumentation 3.1.1”,
http://twython.readthedocs.org/en/latest/usage/starting_out.html#oauth-1-user-authentication
- [10] Long URL, “Browse with Confidence and Increased Security!”,<http://longURL.org>
- [11] Mark A. Hall, “Correlation-based Feature Selection forMachine Learning”, the university of waikato, Hamilton, NewZealand, 1999
- [12] Saeed Abu-Nimeh , Dario Nappa , Xinlei Wang , Suku Nair, “A comparison of machine learning techniques for phishing detection”, Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit, p.60-69, October 04-05, 2007, Pittsburgh, Pennsylvania [doi>10.1145/1299015.1299021]
- [13] Stefan Gremalschi, “Random Forest Prediction of Genetic Susceptibility to Complex Diseases”, course : Algorithms CSc4520/6520
- [14] I. Fette, N. Sadeh, and A. Tomasic, “Learning to detect phishing emails,”in Proceedings of the 16th international conference on World Wide Web. ACM, 2007, pp. 649–656.
- [15] Justin Ma , Lawrence K. Saul , Stefan Savage , Geoffrey M. Voelker, “Beyond blacklists: learning to detect malicious web sites from suspicious URLs”, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, June 28-July 01, 2009, Paris, France
- [16] I. Fette, N. Sadeh, and A. Tomasic. “Learning to detect phishing emails”. Technical Report CMU-ISRI-06-112, Institute for Software Research, Carnegie Mellon University, June 2006. <http://reports-archive.adm.cs.cmu.edu/anon/isri2006/abstracts/06-112.html>.

AUTHORS

Nilesh Sharma received the degree in Computer Science from Hindustan College of Science & Technology, India and degree in Information Technology from IIIT Delhi, India. Currently, he is working as a software engineer at Dubizzle, Dubai. His interests are in python (Django web framework), Databases, JavaScript, JQuery and web designing.



Nishant Sharma is pursuing 5th year of Integrated Dual Degree (B.tech+M.tech) in Electronics & Communication from Indian Institute Of Technology, Roorkee, India. His interests include python, networking, communication and algorithms.



Vishakha Tiwari has completed her B.Tech (1st division with honors) in Computer Science branch from Hindustan College of Science and Technology, Mathura, India. She is placed as software developer in Contata Solutions, Noida, India. Her interests are in Java, Python and Databases.



Shweta Chahar has completed her B.Tech in Computer Science branch from Hindustan College of Science and Technology, Mathura, India. She is placed as program analyst in Cognizant Technology. Her interests are in Java, Python and Databases.



Smriti Maheshwari has completed her B.Tech in Computer Science branch from Hindustan College of Science and Technology, Mathura, India. She is placed as a graduate trainee engineer at HCL comnet Limited. Her interests are in Java and Python.



INTENTIONAL BLANK

NEW FUNCTIONS FOR SECRECY ON REAL PROTOCOLS

Jaouhar Fattahi¹ and Mohamed Mejri¹ and Hanane Houmani²

¹LSI Group, Laval University, Quebec, Canada

²University Hassan II, Morocco

ABSTRACT

In this paper, we present new functions for secrecy in cryptographic protocols: the witness-functions. A witness-function is a protocol-dependent function that is able to prove the correctness of a protocol through its growth. It bases its calculation on the static part of a message only in a role-based specification by using derivation techniques. We show here how to build them. Then, we run an analysis on two real protocols. First, we run an analysis on NSL protocol and we prove that it is correct with respect to the property of secrecy. Then, we run an analysis on a variation of Needham-Schroeder protocol in which we show that a witness-function could even help to discover flaws.

KEYWORDS

Cryptographic Protocols, Role-based specification, Secrecy

1. INTRODUCTION

In this paper, we present a new class of functions to analyze cryptographic protocols statically for the property of secrecy: the witness-functions. Intuitively, an increasing protocol keeps the secret. That means that if the security of all atomic messages exchanged in the protocol does not decay between receiving and sending steps in the protocol, the secret is preserved. For that, we need reliable metrics to estimate the security of atomic messages. This approach has been adopted in some prior works. In [1], Steve Schneider presented the notion of rank-functions as tools to analyze protocols in CSP [2, 3]. They were efficient in analyzing many protocols such as Needham-Schroeder protocol. Nevertheless, a such analysis dictates the protocol implementation in CSP algebra. In addition, building rank-functions is not an easy task and their existence is not certain [4]. In [5] Abadi, by utilizing Spi-Calculus [6, 7], asserted that: "If a protocol typechecks, then it keeps the secret". For that, he restricted the exchanged messages to have strictly the following types: {secret, public, any, confounder} in order to easily know the security level of every component in. This approach cannot analyze prior protocols that had been designed with no respect to this condition.

Similarly, Houmani et al. [8–11] presented universal functions that they named the interpretation functions to statically analyze a protocol. An interpretation function needs to meet some conditions to be "enough good" for the analysis. Naturally, less we have restrictions on functions,

Natarajan Meghanathan et al. (Eds) : ICCSEA, SPPR, VLSI, WiMoA, SCAI, CNSA, WeST - 2014

pp. 229–250, 2014. © CS & IT-CSCP 2014

DOI : 10.5121/csit.2014.4728

more we have the chance to define functions and therefore to have the chance to prove the correctness of protocols. In fact, one function may not succeed to prove the growth of a protocol but another function may. In this respect, we note that the conditions on functions were very restrictive. That's why only two functions had been given: DEK and DEKAN.

We think that the condition of full-invariance by substitution, which enables an analysis run on messages of the generalized roles (messages with variables) to be propagated to valid traces (closed messages), is the most limitative one. From the moment that the goal of our approach is to build as more functions as we can, we believe that if we liberate a function from this condition, we will be able to build more functions. However, liberating a function from a condition may oblige us to take extra precautions when using it.

In this paper, we present the witness-functions as new metrics to analyze cryptographic protocols. We give the way to build them. We show that a witness-function provides two bounds that allow us to pass beyond the limitative condition of full-invariance by substitution by introducing the notion of derivative messages. We exhibit the theorem of analysis with the witness-functions that gives a criterion for the protocol correctness. Finally, we run an analysis on two real protocols. First, we run an analysis on NSL protocol where we prove that it is correct with respect to the property of secrecy. Then, we run an analysis on a variation of Needham-Schroeder protocol in which we show that a witness-function could even help to locate flaws.

2. PRELIMINARY AND NOTATIONS

Here, we give some conventions and notations that we use in this paper.

+ We denote by $\mathcal{C} = \langle \mathcal{M}, \xi, \models, \mathcal{K}, \mathcal{L}, \sqsupseteq, \ulcorner, \urcorner \rangle$ the context of verification in which our analysis is run. It contains the parameters that affect the analysis of a protocol:

- \mathcal{M} : is a set of messages built from the signature $\langle \mathcal{N}, \Sigma \rangle$ where \mathcal{N} is a set of atomic names (nonces, keys, principals, etc.) and Σ is a set of functions (*enc*: encryption, *dec*: decryption, *pair*: concatenation (that we denote by "." here), etc.). i.e. $\mathcal{M} = T_{\langle \mathcal{N}, \Sigma \rangle}(\mathcal{X})$. We denote by Γ the set of substitutions from $\mathcal{X} \rightarrow \mathcal{M}$. We denote by \mathcal{A} all the atomic messages in \mathcal{M} , by $\mathcal{A}(m)$ the set of atomic messages (or atoms) in m and by \mathcal{I} the set of principals including the intruder I . We denote by k^{-1} the reverse form of a key k and we assume that $(k^{-1})^{-1} = k$.
- ξ : is the equational theory in which the algebraic properties of the functions in Σ are described by equations. e.g. $dec(enc(x, y), y^{-1}) = x$.
- $\models_{\mathcal{C}}$: is the inference system of the intruder under the equational theory. Let M be a set of messages and m a message. $M \models_{\mathcal{C}} m$ means that the intruder is able to infer m from M using her capacity. We extend this notation to traces as follows: $\rho \models_{\mathcal{C}} m$ means that the intruder can infer m from the messages exchanged in the trace ρ . We suppose that the intruder has the full control of the net as described by Dolev-Yao model in [12]. That is to say that she can intercept, delete, redirect and modify messages. She knows the public keys of all agents. She knows her private keys and the keys that she shares with other agents. She can encrypt or decrypt any message with known keys. Generically, the intruder has the following rules of building messages:

$$(int) : \frac{\Box}{M \models_{\mathcal{C}} m} [m \in M \cup K(I)]$$

$$(op) : \frac{M \models_{\mathcal{C}} m_1, \dots, M \models_{\mathcal{C}} m_n}{M \models_{\mathcal{C}} f(m_1, \dots, m_n)} [f \in \Sigma]$$

$$(eq) : \frac{M \models_{\mathcal{C}} m', m' =_{\mathcal{C}} m}{M \models_{\mathcal{C}} m}, \text{ with } (m' =_{\mathcal{C}} m) \equiv (m' =_{\xi(\mathcal{C})} m)$$

Example 2.1

The intruder capacity can be described by the following rules:

$$\begin{aligned}
 (int) : & \frac{\Box}{M \models_{\mathcal{C}} m} [m \in M \cup K(I)] \\
 (concat) : & \frac{M \models_{\mathcal{C}} m_1, M \models_{\mathcal{C}} m_2}{M \models_{\mathcal{C}} m_1.m_2} \\
 (deconcat) : & \frac{M \models_{\mathcal{C}} m_1.m_2}{M \models_{\mathcal{C}} m_i} [i \in \{1, 2\}] \\
 (dec) : & \frac{M \models_{\mathcal{C}} k, M \models_{\mathcal{C}} m_k}{M \models_{\mathcal{C}} m} \\
 (enc) : & \frac{M \models_{\mathcal{C}} k, M \models_{\mathcal{C}} m}{M \models_{\mathcal{C}} \{m\}_k}
 \end{aligned}$$

In this example, from a set of messages, an intruder can infer any message in this set. She can encrypt any message when she holds the encryption key. She can decrypt any message when she holds the decryption key and concatenate any two messages and deconcatenate them.

- \mathcal{K} : is a function from \mathcal{I} to \mathcal{M} , that returns to any agent a set of atomic messages describing her initial knowledge. We denote by $K_{\mathcal{C}}(I)$ the initial knowledge of the intruder, or simply $K(I)$ where the context is obvious.
- $\mathcal{L}^{\sqsupseteq}$: is the lattice of security ($\mathcal{L}, \sqsupseteq, \sqcup, \sqcap, \perp, \top$) used to assign security levels to messages. An example of a lattice is $(2^{\mathcal{I}}, \subseteq, \cap, \cup, \mathcal{I}, \emptyset)$ that will be used to attribute to an atomic message α the set of agents that are authorized to know it.
- $\lceil \cdot \rceil$: is a partial function that attributes a value of security (or type) to a message in \mathcal{M} . Let M be a set of messages and m a message. We write $\lceil M \rceil \sqsupseteq \lceil m \rceil$ if $\exists m' \in M. \lceil m' \rceil \sqsupseteq \lceil m \rceil$

Our analysis is performed in a role-based specification. A role-based specification is a set of generalized roles. A generalized role is an abstraction of the protocol where the emphasis is put on a specific agent and where all the unknown messages, and on which the agent cannot carry out any verification, are substituted by variables. An exponent i (the session identifier) is added to a fresh message to say that these components change values from one run to another. A generalized role interprets how a particular agent perceives the exchanged messages. It is extracted from a protocol as follows:

- Extract the roles from the protocol.
- Substitute the unknown messages by fresh variables for each role.

The roles are extracted as follows:

- For each agent, extract from the protocol all the steps in which this principal is participating. Then, add to this abstraction a session identifier i in the steps identifiers and in the fresh values. For example, from the variation of Woo and Lam protocol given in the Table 1, we extract three roles, denoted by R_A (for the agent A), R_B (for the agent B), and R_S (for the agent S).
- Introduce an intruder I to express the fact that the received messages and the sent messages are probably sent or received by the intruder.

- Finally, extract all prefixes from those roles where a prefix ends by a sending step.

$$\begin{aligned}
 p = & \langle 1, A \rightarrow B : A \rangle. \\
 & \langle 2, B \rightarrow A : N_b \rangle. \\
 & \langle 3, A \rightarrow B : \{N_b, k_{ab}\}_{k_{as}} \rangle. \\
 & \langle 4, B \rightarrow S : \{A, \{N_b, k_{ab}\}_{k_{as}}\}_{k_{bs}} \rangle. \\
 & \langle 5, S \rightarrow B : \{N_b, k_{ab}\}_{k_{bs}} \rangle
 \end{aligned}$$

Table 1: The Woo and Lam Protocol

From the roles, we generate the generalized roles. In a generalized role, unknown messages are substituted by variables to express that the agent cannot be sure about its integrity or its origin. In the Woo and Lam protocol, the generalized role of S is:

$$\begin{aligned}
 \mathcal{S}_G^1 = & \langle i.4, I(B) \rightarrow S : \{A, \{U, V\}_{k_{as}}\}_{k_{bs}} \rangle. \\
 & \langle i.5, S \rightarrow I(B) : \{U, V\}_{k_{bs}} \rangle
 \end{aligned}$$

The generalized roles of A are:

$$\begin{aligned}
 \mathcal{A}_G^1 = & \langle i.1, A \rightarrow I(B) : A \rangle \\
 \mathcal{A}_G^2 = & \langle i.1, A \rightarrow I(B) : A \rangle. \\
 & \langle i.2, I(B) \rightarrow A : X \rangle. \\
 & \langle i.3, A \rightarrow I(B) : \{X, k_{ab}^i\}_{k_{as}} \rangle
 \end{aligned}$$

The generalized roles of B are:

$$\begin{aligned}
 \mathcal{B}_G^1 = & \langle i.1, I(A) \rightarrow B : A \rangle. \\
 & \langle i.2, B \rightarrow I(A) : N_b \rangle \\
 \mathcal{B}_G^2 = & \langle i.1, I(A) \rightarrow B : A \rangle. \\
 & \langle i.2, B \rightarrow I(A) : N_b \rangle. \\
 & \langle i.3, I(A) \rightarrow B : Y \rangle. \\
 & \langle i.4, B \rightarrow I(S) : \{A, Y\}_{k_{bs}} \rangle \\
 \mathcal{B}_G^3 = & \langle i.1, I(A) \rightarrow B : A \rangle. \\
 & \langle i.2, B \rightarrow I(A) : N_b \rangle. \\
 & \langle i.3, I(A) \rightarrow B : Y \rangle. \\
 & \langle i.4, B \rightarrow I(S) : \{A, Y\}_{k_{bs}} \rangle. \\
 & \langle i.5, I(S) \rightarrow B : \{N_b^i, Z\}_{k_{bs}} \rangle
 \end{aligned}$$

Thus, the role-based specification of the protocol in the Table 1 is $\mathcal{R}_G(p) = \{\mathcal{A}_G^1, \mathcal{A}_G^2, \mathcal{B}_G^1, \mathcal{B}_G^2, \mathcal{B}_G^3, \mathcal{S}_G^1\}$. The role-based specification is used to express the notion of valid traces of a protocol. More details about the role-based specification could be found in [13–16].

- + A valid trace is an interleaving of substituted generalized roles where each message sent by the intruder can be generated by her using her capacity and by the received messages. We denote by $\llbracket p \rrbracket$ the set of valid traces generated by p .
- + We denote by \mathcal{M}_p^G the set of messages (with variables) in $R_G(p)$, by \mathcal{M}_p the set of closed messages generated by substitution in \mathcal{M}_p^G . We denote by R^+ (respectively R^-) the set of sent messages (respectively received messages) by a honest agent in the role R . Conventionally, we devote the uppercase symbols for sets or sequences of elements and the lowercase for single elements. For example, M denotes a set of messages, m a single message, R a role composed of a sequence of steps, r a step and $R.r$ the role ending by the step r .
- + In our analysis, no restriction on the size of messages or the number of sessions in the protocols is made.

3. INCREASING PROTOCOLS DO NOT REVEAL SECRETS

To analyze a protocol, we need interpretation functions to estimate the security level of every atomic message. In this section, we give sufficient conditions on a function F to guarantee that it is enough good (or reliable) to run an analysis and we show that an increasing protocol is correct with respect to the secrecy property when analyzed with such functions.

3.1 C-reliable interpretation functions

An interpretation function F is said well-formed when it returns the bottom value in the lattice, denoted by \perp , for an atomic message α that appears in clear. It returns for it in the union of two sets, the minimum " \sqcap " of the two values calculated in each set separately. It returns the top value, denoted by " \top ", if it does not appear in this set. These facts are expressed by the definition 3.1.

Definition 3.1. (Well-formed interpretation function)

Let F be an interpretation function and \mathcal{C} a context of verification.

F is well-formed in \mathcal{C} if:

$\forall M, M_1, M_2 \subseteq \mathcal{M}, \forall \alpha \in \mathcal{A}(\mathcal{M})$:

$$\begin{cases} F(\alpha, \{\alpha\}) & = \perp \\ F(\alpha, M_1 \cup M_2) & = F(\alpha, M_1) \sqcap F(\alpha, M_2) \\ F(\alpha, M) & = \top, \text{ if } \alpha \notin \mathcal{A}(M) \end{cases}$$

An interpretation function F is said full-invariant-by-intruder if when it attributes a security level to a message α in a set of messages M , the intruder can never produce another message m that decrease this level (i.e. $F(\alpha, m) \sqsupseteq F(\alpha, M)$) using her capacity in the context of verification, except when α is intended to be known by the intruder (i.e. $\ulcorner K(I) \urcorner \sqsupseteq \ulcorner \alpha \urcorner$). This fact is expressed by the definition 3.2.

Definition 3.2. (Full-invariant-by-intruder interpretation function)

Let F be an interpretation function and \mathcal{C} a context of verification.

F is full-invariant-by-intruder in \mathcal{C} if:

$\forall M \subseteq \mathcal{M}, m \in \mathcal{M}. M \models_{\mathcal{C}} m \Rightarrow \forall \alpha \in \mathcal{A}(m). (F(\alpha, m) \sqsupseteq F(\alpha, M)) \vee (\ulcorner K(I) \urcorner \sqsupseteq \ulcorner \alpha \urcorner)$

An interpretation function F is said reliable if it is well-formed and full-invariant-by-intruder. This fact is expressed by the definition 3.3.

Definition 3.3. (Reliable interpretation function)

Let F be an interpretation function and C a context of verification.

F is C -reliable if F is well-formed and F is full-invariant-by-intruder in C .

A protocol p is said F -increasing when every principal generates continuously valid traces (substituted generalized roles) that never decrease the security levels of received components. The estimation of the value of security of every atom is performed by F . This fact is expressed by the definition 3.4.

Definition 3.4. (F -increasing protocol)

Let F be an interpretation function, C a context of verification and p a protocol.

p is F -increasing in C if:

$\forall R.r \in R_G(p), \forall \sigma \in \Gamma : \mathcal{X} \rightarrow \mathcal{M}_p$ we have:

$$\forall \alpha \in \mathcal{A}(\mathcal{M}_p). F(\alpha, r^+ \sigma) \supseteq \ulcorner \alpha \urcorner \sqcap F(\alpha, R^- \sigma)$$

A secret disclosure consists in manipulating a valid trace of the protocol (denoted by $\llbracket p \rrbracket$) by the intruder using her knowledge $K(I)$ in a context of verification C , to deduce a secret α that she is not intended to know (expressed by: $\ulcorner K(I) \urcorner \not\supseteq \ulcorner \alpha \urcorner$). This fact is expressed by the definition 3.5.

Definition 3.5. (Secret disclosure)

Let p be a protocol and C a context of verification.

We say that p discloses a secret $\alpha \in \mathcal{A}(\mathcal{M})$ in C if:

$$\exists \rho \in \llbracket p \rrbracket. (\rho \models_C \alpha) \wedge (\ulcorner K(I) \urcorner \not\supseteq \ulcorner \alpha \urcorner)$$

Lemma 3.6.

Let F be a C -reliable interpretation function and p an F -increasing protocol.

We have:

$$\forall m \in \mathcal{M}. \llbracket p \rrbracket \models_C m \Rightarrow \forall \alpha \in \mathcal{A}(m). (F(\alpha, m) \supseteq \ulcorner \alpha \urcorner) \vee (\ulcorner K(I) \urcorner \supseteq \ulcorner \alpha \urcorner)$$

See the proof 4 in [17]

The lemma 3.6 says that for any atom in a message produced by an increasing protocol, its security level returned by a reliable interpretation function is kept greater or equal than its initial value in the context, if the intruder is not initially allowed to know it. Hence, initially the atom has a certain level of security. This value cannot be decreased by the intruder using her knowledge and the received messages since it is full-invariant-by-intruder. In every new step of a valid trace, involved messages are better protected since the protocol is increasing. The proof is then run by induction on the size of the trace using the reliability properties of the interpretation function in every step of the induction.

Theorem 3.7. (Theorem of Correctness of Increasing Protocols)

Let F be a C -reliable interpretation function and p a F -increasing protocol.

p is C -correct with respect to the secrecy property

Proof.

Let's suppose that p discloses an atomic secret α .

From the definition 3.5 we have:

$$\exists \rho \in \llbracket p \rrbracket. (\rho \models_{\mathcal{C}} \alpha) \wedge (\ulcorner K(I) \urcorner \not\sqsubseteq \ulcorner \alpha \urcorner) \quad (1)$$

Since F is a \mathcal{C} -reliable interpretation function and p an F -increasing protocol, we have from the lemma 3.6:

$$(F(\alpha, \alpha) \sqsupseteq \ulcorner \alpha \urcorner) \vee (\ulcorner K(I) \urcorner \sqsupseteq \ulcorner \alpha \urcorner) \quad (2)$$

From 1 and 2, we have:

$$F(\alpha, \alpha) \sqsupseteq \ulcorner \alpha \urcorner \quad (3)$$

Since F is well-formed in \mathcal{C} , then:

$$F(\alpha, \alpha) = \perp \quad (4)$$

From 3 and 4 we have:

$$\perp = \ulcorner \alpha \urcorner \quad (5)$$

5 is impossible because it is contradictory with: $\ulcorner K(I) \urcorner \not\sqsubseteq \ulcorner \alpha \urcorner$ in 1.

Then p is \mathcal{C} -correct with respect to the secrecy property

The theorem 3.1 asserts that an increasing protocol is correct with respect to the secrecy property when analyzed with a reliable interpretation function. It is worth saying that compared to the sufficient conditions stated in [11], we have one less. Thus, in [11], Houmani demanded from the interpretation function an additional condition: the full-invariance by substitution. That's to say, interpretation function has also to resist to the problem of substitution of variables. Here, we liberate our functions from this limitative condition in order to be able to build more of them. We rehouse this condition in our new definition of an increasing protocol which is required now to be increasing on valid traces (closed messages) rather than messages of the generalized roles (message with variables). Therefore, the problem of substitutions is transferred to the protocol and becomes less difficult to handle.

4. CONSTRUCTION OF RELIABLE INTERPRETATION FUNCTIONS

As seen in the previous section, to analyze a protocol we need reliable interpretation functions to estimate the level of security of any atom in a message. In this section, we exhibit a constructive way to build these functions. We first exhibit the way to build a generic class of reliable selections inside the protection of the most external key (or simply the external key). Then we propose specialized selections of this class. Finally we give the way to build reliable selection-based interpretation functions. Similar techniques based on selections were proposed in previous works, especially in [8, 10, 11] to build universal functions based on the selection of the direct key of encryption and in [18] to check correspondences in protocols. But first of all, we present the notion of well-protected messages that have valuable properties that we will use in the definition of reliable selections. Briefly, a well-protected message is a message such that every non public atom α in it is encrypted by at least one key k such that $\ulcorner k^{-1} \urcorner \sqsupseteq \ulcorner \alpha \urcorner$, after elimination of unnecessary keys (e.g. $e(k, d(k^{-1}, m)) \rightarrow m$). The main advantage of an analysis performed over a set of well-protected messages is that the intruder cannot deduce any secret when she uses only her knowledge in the context of verification (without using the protocol rules).

4.1 Protocol analysis in Well-Protected Messages

We denote by $\mathcal{E}_{\mathcal{C}}$ the set of encryption functions and by $\overline{\mathcal{E}}_{\mathcal{C}}$ the complementary set $\Sigma \setminus \mathcal{E}_{\mathcal{C}}$ in a context of verification \mathcal{C} .

The definition 4.1 defines the application *keys* that returns the encryption keys of any atom α in a message m .

Definition 4.1. (Keys)

Let $M \subseteq \mathcal{M}$, $f \in \Sigma$ and $m \in M$.

We define the application $Keys : \mathcal{A} \times \mathcal{M} \longrightarrow \mathcal{P}(\mathcal{P}(\mathcal{A}))$ as follows:

$\forall t_1, t_2 \dots t_n$ subterms of m :

$$\begin{aligned} Keys(\alpha, \alpha) &= \{\emptyset\} \\ Keys(\alpha, \beta) &= \emptyset, \text{ if } \alpha \neq \beta \text{ and } \beta \in \mathcal{A} \\ Keys(\alpha, f_k(t_1, \dots, t_n)) &= \{k\} \otimes \bigcup_{i=1}^n Keys(\alpha, t_i), \text{ if } f_k \in \mathcal{E}_{\mathcal{C}} \\ Keys(\alpha, f(t_1, \dots, t_n)) &= \bigcup_{i=1}^n Keys(\alpha, t_i), \text{ if } f \in \overline{\mathcal{E}}_{\mathcal{C}} \end{aligned}$$

We extend the application *Keys* to sets as follows:

$$\forall M \subseteq \mathcal{M}. Keys(\alpha, M) = \bigcup_{m \in M} Keys(\alpha, m) \text{ and } Keys(\alpha, \emptyset) = \emptyset.$$

The definition 4.2 is related to equational theory. It fixes the form of a message that we are going to choose. The chosen form (normal form) is the one that provides the smallest set of encryption keys. This in order to eliminate the unnecessary keys (e.g. $e(k, d(k^{-1}, m)) \rightarrow m$).

Definition 4.2. (Equational theory, Rewriting system and Normal Form)

We assume that we can transform the equational theory ξ given in the context of verification to a convergent rewriting system \rightarrow_{ξ} such that:

$$\forall m \in \mathcal{M}, \forall \alpha \in \mathcal{A}(m), \forall l \rightarrow r \in \rightarrow_{\xi}, \quad Keys(\alpha, r) \subseteq Keys(\alpha, l) \quad (6)$$

We denote by m_{\Downarrow} the normal form of m in \rightarrow_{ξ} .

The kind of rewriting systems orientation in the definition 4.2 poses no problem with the most of equational theories in the literature [19–21].

Example 4.3.

Let $m = \{\{\{A.\alpha\}_{k_{ab}}\}_{k_{ab}^{-1}}\}_{k_{ac}}; m_{\Downarrow} = \{A.\alpha\}_{k_{ac}}$

In the definition 4.4, we introduce the application *Access*. Every element of $Access(\alpha, m)$ contains a set of required keys to decrypt α in m after elimination of unnecessary keys by the normal form defined in 4.2.

Definition 4.4. (Access)

Let $M \subseteq \mathcal{M}$, $f \in \Sigma$ and $m \in M$.

We define the application $Access : \mathcal{A} \times \mathcal{M} \longrightarrow \mathcal{P}(\mathcal{P}(\mathcal{A}))$ as follows:

$\forall t_1, t_2 \dots t_n$ subterms of m :

$$\begin{aligned} Access(\alpha, \alpha) &= \{\emptyset\} \\ Access(\alpha, \beta) &= \emptyset, \text{ if } \alpha \neq \beta \text{ and } \beta \in \mathcal{A} \\ Access(\alpha, f_k(t_1, \dots, t_n)) &= \{k^{-1}\} \otimes \bigcup_{i=1}^n Access(\alpha, t_i), \text{ if } f_k \in \mathcal{E}_C \text{ and } f_k(t_1, \dots, t_n) = f_k(t_1, \dots, t_n)_{\Downarrow} \\ Access(\alpha, f(t_1, \dots, t_n)) &= \bigcup_{i=1}^n Access(\alpha, t_i), \text{ if } f \in \overline{\mathcal{E}}_C \text{ and } f(t_1, \dots, t_n) = f_k(t_1, \dots, t_n)_{\Downarrow} \\ Access(\alpha, f(t_1, \dots, t_n)) &= Access(\alpha, f(t_1, \dots, t_n)_{\Downarrow}), \text{ if not.} \end{aligned}$$

We extend the application $Access$ to sets as follows:

$$\forall M \subseteq \mathcal{M}. Access(\alpha, M) = \bigcup_{m \in M} Access(\alpha, m) \text{ and } Access(\alpha, \emptyset) = \emptyset.$$

Example 4.5.

Let m be a message such that: $m = \{\{A.D.\alpha\}_{k_{ab}}.\alpha.\{A.E.\{C.\alpha\}_{k_{ef}}\}_{k_{ab}}\}_{k_{ac}}$;

$$Access(\alpha, m) = \{\{k_{ac}^{-1}, k_{ab}^{-1}\}, \{k_{ac}^{-1}\}, \{k_{ac}^{-1}, k_{ab}^{-1}, k_{ef}^{-1}\}\}.$$

In the definition 4.6, we define a well-protected message. Informally, a well-protected message is a message such that every non-public atom α in it (such that $\ulcorner \alpha \urcorner \sqsubset \perp$) is encrypted by at least one key k such that $\ulcorner k^{-1} \urcorner \sqsupseteq \ulcorner \alpha \urcorner$ after elimination of unnecessary keys by the normal form given in the definition 4.2.

Definition 4.6. (Well-protected message)

Let \mathcal{C} be context of verification, $m \in \mathcal{M}$, $M \subseteq \mathcal{M}$ and $\alpha \in \mathcal{A}(m)$ such that $\ulcorner \alpha \urcorner \sqsubset \perp$.

We say that α is well-protected in m if:

$$\forall \beta \in Access(\alpha, m). \ulcorner \beta \urcorner \sqsupseteq \ulcorner \alpha \urcorner$$

We say that α is well-protected in M if:

$$\forall m \in M. \alpha \text{ is well-protected in } m$$

We say that m is well-protected in \mathcal{C} if:

$$\forall \alpha \in \mathcal{A}(m). \alpha \text{ is well-protected in } m$$

We say that M is well-protected in \mathcal{C} if:

$$\forall m \in M. m \text{ is well-protected in } \mathcal{C}.$$

In the definition 4.7, we define the application $Clear(m)$. Informally, $Clear(m)$ returns the set of all atoms that appear in clear in m after elimination of unnecessary keys by the normal form given in the definition 4.2.

Definition 4.7. ($Clear$)

Let $m \in \mathcal{M}$ and $M \subseteq \mathcal{M}$.

$$Clear(m) = \{\alpha \in \mathcal{A}(m). \emptyset \in Access(\alpha, m)\}$$

We extend this definition to sets as follows :

$$\text{Clear}(M) = \bigcup_{m \in M} \text{Clear}(m)$$

Lemma 4.8.

Let M be a set of well-protected messages in \mathcal{M} . We have:

$$M \models_{\mathcal{C}} m \Rightarrow \forall \alpha \in \mathcal{A}(m). (\alpha \text{ is well-protected in } m) \vee (\ulcorner K(I) \urcorner \sqsupseteq \ulcorner \alpha \urcorner)$$

See the proof 7 in [17]

The lemma 4.8 says that from a set of well-protected messages, all atomic messages beyond the knowledge of the intruder (i.e. $\ulcorner K(I) \urcorner \not\sqsupseteq \ulcorner \alpha \urcorner$) remain well-protected in any message that the intruder could infer. Indeed, since each atom that does not appear in clear (non-public) in this set is encrypted by at least one key k such that $\ulcorner k^{-1} \urcorner \sqsupseteq \ulcorner \alpha \urcorner$, then the intruder has to retrieve the key k^{-1} before she sees α not well-protected in any message (clear). But, the key k^{-1} is in its turn encrypted by at least one key k' such that $\ulcorner k'^{-1} \urcorner \sqsupseteq \ulcorner k^{-1} \urcorner$. The proof is then conducted by induction on the encryption keys.

Lemma 4.9. (Lemma of non-disclosure of atomic secrets in well-protected messages)

Let M be a set of well-protected messages in \mathcal{M} and α an atomic message in M .

We have:

$$M \models_{\mathcal{C}} \alpha \Rightarrow \ulcorner K(I) \urcorner \sqsupseteq \ulcorner \alpha \urcorner$$

Proof.

From the lemma 4.8, we have $M \models_{\mathcal{C}} \alpha$ then:

$$(\alpha \text{ is well-protected in } \alpha) \vee (\ulcorner K(I) \urcorner \sqsupseteq \ulcorner \alpha \urcorner) \tag{7}$$

But α is not well-protected in α , then we have:

$$\ulcorner K(I) \urcorner \sqsupseteq \ulcorner \alpha \urcorner \tag{8}$$

4.2 Discussion and Assumption

The lemma 4.9 expresses an important result. It states that from a set of well-protected messages the intruder cannot deduce any secret that she is not supposed to know when she uses only her knowledge in the context of verification (without using the protocol rules). It is worth saying that verifying whether a protocol operates over a space of well-protected messages or not is an easy task and most of real protocols respect this condition.

4.3 Building reliable selections

Now, we will focus on building selections such that when they are composed to suitable homomorphisms, provide reliable interpretation functions. The definition 4.10 introduces the notion of a well-formed selection and the definition 4.11 introduces the notion of a full-invariant-by-intruder selection.

Definition 4.10. (Well-formed selection)

Let $M, M_1, M_2 \subseteq \mathcal{M}$ such that M, M_1 and M_2 are well-protected.

Let $S : \mathcal{A} \times \mathcal{M} \mapsto 2^{\mathcal{A}}$ be a selection.

We say that S is well-formed in \mathcal{C} if:

$$\begin{cases} S(\alpha, \{\alpha\}) &= \mathcal{A}, \\ S(\alpha, M_1 \cup M_2) &= S(\alpha, M_1) \cup S(\alpha, M_2), \\ S(\alpha, M) &= \emptyset, \text{ if } \alpha \notin \mathcal{A}(M) \end{cases}$$

For an atom α in a set of messages M , a well-formed selection returns all the atoms in \mathcal{M} if $M = \{\alpha\}$. It returns for it in the union of two sets of messages, the union of the two selections performed in each set separately. It returns the empty set if the atom does not appear in M .

Definition 4.11. (Full-invariant-by-intruder selection)

Let $M \subseteq \mathcal{M}$ such that M is well-protected.

Let $S : \mathcal{A} \times \mathcal{M} \mapsto 2^{\mathcal{A}}$ be a selection.

We say that S is full-invariant-by-intruder in \mathcal{C} if:

$\forall M \subseteq \mathcal{M}, m \in \mathcal{M}$, we have:

$$M \models_{\mathcal{C}} m \Rightarrow \forall \alpha \in \mathcal{A}(m). (S(\alpha, m) \subseteq S(\alpha, M)) \vee (\ulcorner K(I) \urcorner \sqsupseteq \ulcorner \alpha \urcorner)$$

The aim of a full-invariant-by-intruder selection is to provide a full-invariant-by-intruder function when composed to an adequate homomorphism that transforms its returned atoms into security levels. Since a full-invariant-by-intruder function is requested to resist to any attempt of the intruder to generate a message m from any set of messages M in which the level of security of an atom, that she is not allowed to know, decreases compared to its value in M , a full-invariant-by-intruder selection is requested to resist to any attempt of the intruder to generate a message m from any set of messages M in which the selection associated to an atom, that she is not allowed to know, could be enlarged compared to the selection associated to this atom in M . This fact is described by the definition 4.11.

Definition 4.12. (Reliable selection)

Let $S : \mathcal{A} \times \mathcal{M} \mapsto 2^{\mathcal{A}}$ be a selection and \mathcal{C} be a context of verification.

S is \mathcal{C} -reliable if S is well-formed and S is full-invariant-by-intruder in \mathcal{C} .

4.3.1 Reliable selections inside the protection of an external key

Here, we define a generic class of selections that we denote by S_{Gen}^{EK} and we prove that any instance of it is reliable under some condition.

Definition 4.13. (S_{Gen}^{EK} : selection inside the protection of an external key)

We denote by S_{Gen}^{EK} the class of all selections S that meet the following conditions:

- $S(\alpha, \alpha) = \mathcal{A}$; (9)

- $S(\alpha, m) = \emptyset$, if $\alpha \notin \mathcal{A}(m)$; (10)

- $\forall \alpha \in \mathcal{A}(m)$, where $m = f_k(m_1, \dots, m_n)$: (11)

$$S(\alpha, m) \subseteq \left(\bigcup_{1 \leq i \leq n} \mathcal{A}(m_i) \cup \{k^{-1}\} \setminus \{\alpha\} \right) \text{ if } f_k \in \mathcal{E}_C \text{ and } \lceil k^{-1} \rceil \supseteq \lceil \alpha \rceil \text{ and } m = m_{\downarrow}$$

- $\forall \alpha \in \mathcal{A}(m)$, where $m = f(m_1, \dots, m_n)$: (12)

$$S(\alpha, m) = \begin{cases} \bigcup_{1 \leq i \leq n} S(\alpha, m_i) & \text{if } f_k \in \mathcal{E}_C \text{ and } \lceil k^{-1} \rceil \not\supseteq \lceil \alpha \rceil \text{ and } m = m_{\downarrow} \text{ (a)} \\ \bigcup_{1 \leq i \leq n} S(\alpha, m_i) & \text{if } f \in \bar{\mathcal{E}}_C \text{ and } m = m_{\downarrow} \text{ (b)} \\ S(\alpha, m_{\downarrow}) & \text{if } m \neq m_{\downarrow} \text{ (c)} \end{cases}$$

- $S(\alpha, \{m\} \cup M) = S(\alpha, m) \cup S(\alpha, M)$ (13)

For an atom α in an encrypted message $m = f_k(m_1, \dots, m_n)$, a selection S as defined above returns a subset (see " \subseteq " in equation 11) among atoms that are neighbors of α in m inside the protection of the most external protective key k including its reverse form k^{-1} . The atom α itself is not selected. This set of candidate atoms is denoted by $\bigcup_{1 \leq i \leq n} \mathcal{A}(m_i) \cup \{k^{-1}\} \setminus \{\alpha\}$ in the equation 11. The most external protective key (or simply the external key) is the most external one that satisfies $\lceil k^{-1} \rceil \supseteq \lceil \alpha \rceil$. A such key must exist when the set of messages generated by the protocol is well-protected, which is one of our assumptions above. By neighbor of α in m , we mean any atom that travels with it inside the protection of the external key.

S_{Gen}^{EK} defines a generic class of selections since it does not identify what atoms to select precisely inside the protection of the external key. It identifies only the atoms that are candidates for selection and among them we are allowed to return any subset.

Proposition 4.14.

Let $S \in S_{Gen}^{EK}$ and \mathcal{C} be a context of verification.

Let's have a rewriting system \rightarrow_{ξ} such that $\forall m \in \mathcal{M}, \forall \alpha \in \mathcal{A}(m) \wedge \alpha \notin Clear(m)$, we have:

$$\forall l \rightarrow r \in \rightarrow_{\xi}, S(\alpha, r) \subseteq S(\alpha, l) \quad (14)$$

We have:

S is \mathcal{C} -reliable.

See the proof 11 in [17]

Remark. The condition on the rewriting system \rightarrow_{ξ} given by the equation 14 in the definition 4.14 is introduced to make sure that the selection in the normal form is the smallest among all forms of a given message. This prevents the selection S to select atoms that are inserted maliciously by the intruder by manipulating the equational theory. Hence, we are sure that all selected atoms by S are honest. For example, let $m = \{\alpha.S\}_{k_{ab}}$ be a message in a homomorphic cryptography (i.e $\{\alpha.S\}_{k_{ab}} = \{\alpha\}_{k_{ab}} \cdot \{S\}_{k_{ab}}$). In the form $\{\alpha.S\}_{k_{ab}}$, the selection $S(\alpha, \{\alpha.S\}_{k_{ab}})$ may select S since S is a neighbor of α inside the protection of k_{ab} , but in the

form $\{\alpha\}_{k_{ab}} \cdot \{S\}_{k_{ab}}$ the selection $S(\alpha, \{\alpha\}_{k_{ab}} \cdot \{S\}_{k_{ab}})$ may not because it is not a neighbor of α . Then, we must make sure that the rewriting system \rightarrow_{ξ} we are using is oriented in such way that it chooses the form $\{\alpha\}_{k_{ab}} \cdot \{S\}_{k_{ab}}$ rather than the form $\{\alpha.S\}_{k_{ab}}$ because there is no guarantee that S is a honest neighbor and that it had not been inserted maliciously by the intruder using the homomorphic property in the theory. We assume that the rewriting system we are using meets this condition.

For the proposition 4.14, it is easy to check that by construction a selection S , that is instance of S_{Gen}^{EK} , is well-formed. The proof of full-invariance-by-intruder is carried out by induction on the tree of construction of a message. The principal idea of the proof is that the selection related to an atom α in a message m takes place inside the encryption by the most external protective key (such that: $\ulcorner k^{-1} \urcorner \sqsupseteq \ulcorner \alpha \urcorner$). Thus an intruder cannot modify this selection when she does not have the key k^{-1} (i.e. $\ulcorner K(I) \urcorner \not\sqsupseteq \ulcorner k^{-1} \urcorner$). Besides, according to the lemma 4.9, in a set of well-protected messages the intruder can never infer this key since it is atomic. So, this selection can only be modified by people who are initially authorized to know α (i.e. $\ulcorner K \urcorner \sqsupseteq \ulcorner k^{-1} \urcorner$ and then $\ulcorner K \urcorner \sqsupseteq \ulcorner \alpha \urcorner$). In addition, the intruder cannot neither use the equational theory to alter this selection thanks to the condition made on the rewriting system in the remark 4.3.1. Therefore any set of candidate atoms returned by S cannot be altered (enlarged) by the intruder in any message m that she can infer, as required by a full-invariant-by-intruder selection.

Example 4.15.

Let α be an atomic message and m a message such that: $\ulcorner \alpha \urcorner = \{A, B\}$ and $m = \{A.C.\alpha.D\}_{k_{ab}}$. Let S_1, S_2 and S_3 be three selections such that: $S_1(\alpha, m) = \{k_{ab}^{-1}\}$, $S_2(\alpha, m) = \{A, C, k_{ab}^{-1}\}$ and $S_3(\alpha, m) = \{A, C, D, k_{ab}^{-1}\}$. These three selections are \mathcal{C} -reliable.

4.4 Instantiation of reliable selections from the class S_{Gen}^{EK}

Now that we defined a generic class of reliable selections S_{Gen}^{EK} , we will instantiate some concrete selections from it, that are naturally reliable. Instantiating S_{Gen}^{EK} consists in defining selections that return precise sets of atoms among the candidates allowed by S_{Gen}^{EK} .

4.4.1 The selection S_{MAX}^{EK}

The selection S_{MAX}^{EK} is the instance of the class S_{Gen}^{EK} that returns for an atom in a message m all its neighbors, that are principal identities, inside the protection of the external protective key k in addition to its reverse key k^{-1} . (MAX means: the MAXimum of principal identities)

4.4.2 The selection S_{EK}^{EK}

The selection S_{EK}^{EK} is the instance of the class S_{Gen}^{EK} that returns for an atom in a message m only the reverse key of the external protective key. (EK means: External Key)

4.4.3 The selection S_N^{EK}

The selection S_N^{EK} is the instance of the class S_{Gen}^{EK} that returns only its neighbors, that are principal identities, inside the protection of the external protective key. (N means: Neighbors)

Example 4.16.

Let α be an atom and m a message such that: $\ulcorner \alpha \urcorner = \{A, C\}$ and $m = \{\{\{\alpha.E\}_{k_{ab}}.F\}_{k_{ac}}.D\}_{k_{ad}}$
 $S_{MAX}^{EK}(\alpha, m) = \{E, F, k_{ac}^{-1}\}$; $S_{EK}^{EK}(\alpha, m) = \{k_{ac}^{-1}\}$; $S_N^{EK}(\alpha, m) = \{E, F\}$

4.5 Specialized C-reliable selection-based interpretation functions

Now, we define specific functions that are a composition of an appropriate homomorphism and instances of the class of selections S_{Gen}^{EK} . This homomorphism exports the properties of reliability from a selection to a function and transforms selected atoms to security levels. The following proposition states that any function that is a composition of the homomorphism defined below and the selections S_{Gen}^{EK} is reliable.

Proposition 4.17.

Let ψ be a homomorphism defined as follows:

$$\begin{aligned} \psi &: (2^A)^\subseteq \mapsto \mathcal{L}^\sqsupseteq \\ M &\mapsto \begin{cases} \top & \text{if } M = \emptyset \\ \bigcap_{\alpha \in M} \psi(\alpha) & \text{if not.} \end{cases} \end{aligned}$$

$$\text{such that: } \psi(\alpha) = \begin{cases} \{\alpha\} & \text{if } \alpha \in \mathcal{I} \text{ (Principal Identities)} \\ \ulcorner \alpha \urcorner & \text{if not.} \end{cases}$$

We have: $F_{MAX}^{EK} = \psi \circ S_{MAX}^{EK}$, $F_{EK}^{EK} = \psi \circ S_{EK}^{EK}$ and $F_N^{EK} = \psi \circ S_N^{EK}$ are C-reliable.

See the proof 17 in [17]

The homomorphism ψ in the proposition 4.17 assigns for a principal in a selection, its identity. It assigns for a key its level of security in the context of verification. This homomorphism ensures the mapping from the operator " \subseteq " to the operator " \sqsupseteq " in the lattice which offers to an interpretation function to inherit the full-invariance-by-intruder from its associated selection. In addition, it ensures the mapping from the operator " \cup " to the operator " \cap " in the lattice, which offers to an interpretation function to be well-formed if its associated selection is well-formed. Generally, every function $\psi \circ S$ remains reliable for any selection S in S_{Gen}^{EK} .

Example 4.18.

Let α be an atom, m a message and k_{ab} a key such that:

$$\begin{aligned} \ulcorner \alpha \urcorner &= \{A, B, S\}; m = \{A.C.\alpha.D\}_{k_{ab}}; \ulcorner k_{ab}^{-1} \urcorner = \{A, B, S\}; \\ S_{EK}^{EK}(\alpha, m) &= \{k_{ab}^{-1}\}; S_N^{EK}(\alpha, m) = \{A, C, D\}; S_{MAX}^{EK}(\alpha, m) = \{A, C, D, k_{ab}^{-1}\}; \\ F_{EK}^{EK}(\alpha, m) &= \psi \circ S_{EK}^{EK}(\alpha, m) = \ulcorner k_{ab}^{-1} \urcorner = \{A, B, S\}; F_N^{EK}(\alpha, m) = \psi \circ S_N^{EK}(\alpha, m) = \{A, C, D\}; \\ F_{MAX}^{EK}(\alpha, m) &= \psi \circ S_{MAX}^{EK}(\alpha, m) = \{A, C, D\} \cup \ulcorner k_{ab}^{-1} \urcorner = \{A, C, D\} \cup \{A, B, S\} = \{A, C, D, B, S\}. \end{aligned}$$

5. INSUFFICIENCY OF RELIABLE INTERPRETATION FUNCTIONS TO ANALYZE GENERALIZED ROLES

So far, we presented a class of selection-based interpretation functions that have the required properties to analyze protocols statically. Unfortunately, they operate on valid traces that contain closed messages only. Nevertheless, a static analysis must be led over the finite set of messages

of the generalized roles of the protocol because the set of valid traces is infinite. The problem is that the finite set of the generalized roles contains variables and the functions we defined are not "enough prepared" to analyze such messages because they are not supposed to be full-invariant by substitution [22–24]. The full-invariance by substitution is the property that allows us to perform an analysis over messages with variables and to export the conclusion made-on to closed messages. In the following section, we deal with the substitution question. We introduce the concept of derivative messages to reduce the impact of variables and we build the witness functions that operate on these derivative messages rather than messages themselves. As we will see, the witness-functions provide two interesting bounds that are independent of all substitutions. This fully replaces the property of full-invariance by substitution. Finally, we define a criterion of protocol correctness based on these two bounds.

5.1 Derivative message

Let $m, m_1, m_2 \in \mathcal{M}$; $\mathcal{X}_m = \text{Var}(m)$; $S_1, S_2 \subseteq 2^{\mathcal{X}_m}$; $\alpha \in \mathcal{A}(m)$; $X, Y \in \mathcal{X}_m$ and ϵ be the empty message.

Definition 5.1. (Derivation)

We define the derivative message as follows:

$$\begin{aligned} \partial_X \epsilon &= \epsilon \\ \partial_X \alpha &= \alpha \\ \partial_X X &= \epsilon \\ \partial_X Y &= Y, X \neq Y \\ \partial_X f(m) &= f(\partial_X m), f \in \mathcal{E}_C \cup \overline{\mathcal{E}}_C \\ \partial_{\{X\}} m &= \partial_X m \\ \partial(\overline{X}) m &= \partial_{\{\mathcal{X}_m \setminus X\}} m \\ \partial_{S_1 \cup S_2} m &= \partial_{S_2 \cup S_1} m = \partial_{S_1} \partial_{S_2} m = \partial_{S_2} \partial_{S_1} m \end{aligned}$$

To be simple, we denote by ∂m the expression $\partial_{\mathcal{X}_m} m$. The operation of derivation introduced by the definition 5.1 (denoted by ∂) eliminates variables in a message. $\partial_X m$ consists in eliminating the variable X in m . $\partial(\overline{X}) m$ consists in eliminating all variables, except X , in m . Hence, X when overlined is considered as a constant in m . ∂m consists in eliminating all the variables in m .

Definition 5.2.

Let $m \in \mathcal{M}_p^G$, $X \in \mathcal{X}_m$ and $m\sigma$ be a closed message.

For all $\alpha \in \mathcal{A}(m\sigma)$, $\sigma \in \Gamma$, we denote by:

$$F(\alpha, \partial[\overline{\alpha}]m\sigma) = \begin{cases} \top & \text{if } \alpha \notin \mathcal{A}(m\sigma), \\ F(\alpha, \partial m) & \text{if } \alpha \in \mathcal{A}(\partial m), \\ F(X, \partial[\overline{X}]m) & \text{if } \alpha \in \mathcal{A}(X\sigma) \wedge \alpha \notin \mathcal{A}(\partial m). \end{cases}$$

A message m in a generalized role is composed of two parts: a static part and a dynamic part. The dynamic part is described by variables. For an atom α in the static part (i.e. ∂m), $F(\alpha, \partial[\overline{\alpha}]m\sigma)$ removes the variables in m and gives it the value $F(\alpha, \partial m)$. For anything that is not an atom of the static part -that comes by substitution of some variable X in m - $F(\alpha, \partial[\overline{\alpha}]m\sigma)$ considers it as the variable itself, treated as a constant and as a block, and gives it all the time the same value: $F(X, \partial[\overline{X}]m)$. It gives the top value for any atom that does

not appear in $m\sigma$. For any F such that its associated selection is an instance of the class S_{Gen}^{EK} , $F(\alpha, \partial[\bar{\alpha}]m\sigma)$ depends only on the static part of m since α is not selected. The function in the definition 5.2 presents the three following major facts :

1. An atom of the static part of a message with variables, when analyzed with a such function, is considered as an atom in a message with no variables (a closed message);
2. A variable, when analyzed by such function, is considered as any component that substitutes it (that is not in the static part of the message) with no respect to other variables, if any;
3. For any F such that its associated selection is an instance of the class S_{Gen}^{EK} , $F(\alpha, \partial[\bar{\alpha}]m\sigma)$ depends only on the static part of m since α is not selected.

One could suggest that we attribute to an atom α in a closed message $m\sigma$ the value returned by the function $F(\alpha, \partial[\bar{\alpha}]m\sigma)$ given in the definition 5.2 and hence we neutralize the variable effects. Unfortunately, this does not happen without undesirable "side-effects" because derivation generates a "loss of details". Let's look at the example 5.3.

Example 5.3.

Let m_1 and m_2 be two messages of a generalized role of a protocol p such that $m_1 = \{\alpha.C.X\}_{k_{ab}}$ and $m_2 = \{\alpha.Y.D\}_{k_{ab}}$ and $\ulcorner \alpha \urcorner = \{A, B\}$;

Let $m = \{\alpha.C.D\}_{k_{as}}$ be a closed message in a valid trace generated by p ;

$$F_{MAX}^{EK}(\alpha, \partial[\bar{\alpha}]m) = \begin{cases} \{C, A, B\} & \text{if } m \text{ comes by the substitution of } X \text{ by } D \text{ in } m_1 \\ \{D, A, B\} & \text{if } m \text{ comes by the substitution of } Y \text{ by } C \text{ in } m_2 \end{cases}$$

Hence $F_{MAX}^{EK}(\alpha, \partial[\bar{\alpha}]m)$ is not even a function on the closed message m since it may return more than one image for the same preimage. This leads us straightly to the witness-functions.

6. THE WITNESS-FUNCTIONS

Definition 6.1. (Witness-Function)

Let $m \in \mathcal{M}_p^G$, $X \in \mathcal{X}_m$ and $m\sigma$ be a closed message.

Let p be a protocol and F be a \mathcal{C} -reliable interpretation function.

We define a witness-function $\mathcal{W}_{p,F}$ for all $\alpha \in \mathcal{A}(m\sigma)$, $\sigma \in \Gamma$, as follows:

$$\mathcal{W}_{p,F}(\alpha, m\sigma) = \bigcap_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m'\sigma' = m\sigma}} F(\alpha, \partial[\bar{\alpha}]m'\sigma')$$

$\mathcal{W}_{p,F}$ is said a witness-function inside the protection of an external key when F is an interpretation function such that its associated selection is an instance of the class S_{Gen}^{EK} .

According to the example 5.3, the application defined in 5.2 is not necessary a function in \mathcal{M}_p^G as a valid trace could have more than one source (or provenance) in \mathcal{M}_p^G and each source has a different static part. A witness-function is yet a function in p since it searches all the sources of the closed message in input and returns the minimum (the union). This minimum naturally exists and is unique in the finite set \mathcal{M}_p^G . A witness-function is protocol-dependent as it depends on messages in the generalized roles of the protocol. However, it is built uniformly for any pair (protocol, interpretation function) in input.

Remark.

For a witness-function inside the protection of an external key, since its associated interpretation function ranks an atom always from a message m having an encryption pattern, i.e. when $f_k \in \mathcal{E}_C$, the search of the sources of the closed message $m\sigma$ in \mathcal{M}_p^G (i.e. $\{m' \in \mathcal{M}_p^G \mid \exists \sigma' \in \Gamma. m'\sigma' = m\sigma\}$) is limited to a search in the encryption patterns in \mathcal{M}_p^G .

6.1 Legacy of Reliability

The proposition 6.2 asserts that an interpretation function F inside the protection of an external key transmits its reliability to its associated witness-function $\mathcal{W}_{p,F}$. In fact, the selection associated with a witness function inside the protection of an external key is the union of selections associated with the interpretation function F , limited to derivative messages. It is easy to check that a witness-function is well-formed. Concerning the full-invariance-by-intruder property, as the derivation just eliminates variables (so some atoms when the message is substituted), and since each selection in the union returns a subset among allowed candidates, then the union itself returns a subset among allowed candidates (the union of subsets is a subset).

Therefore, the selection associated with a witness-function stays an instance of the class SEK_{Gen} , so full-invariant-by-intruder. Since the witness-function is the composition of the homomorphism of F and an instance of the class SEK_{Gen} , then it is reliable.

Proposition 6.2.

Let $\mathcal{W}_{p,F}$ be a witness-function inside the protection of an external key.
We have:

$\mathcal{W}_{p,F}$ inherits reliability from F .

See the proof 18 in [17]

6.2 Bounds of a Witness-Function

In the lemma 6.4, we define two interesting bounds of a witness-function that are independent of all substitutions. The upper bound of a witness-function ranks the security level of an atom α in a closed message $m\sigma$ from one confirmed source m (m is a natural source of $m\sigma$), the witness-function itself ranks it from the exact sources of $m\sigma$ that are known only when the protocol is run, and the lower bound ranks it from the exact sources of $m\sigma$ that are known only when the protocol is run, and the lower bound ranks it from all likely sources of $m\sigma$ (i.e. the messages that are unifiable with m in \mathcal{M}_p^G).

Example 6.3.

Let $\mathcal{M}_p^G = \{\{\alpha.B.X\}_{k_{ad}}, \{\alpha.Y.S\}_{k_{ad}}, \{A.Z\}_{k_{bc}}\}$ with $Var(\mathcal{M}_p^G) = \{X, Y, Z\}$;
 Let $m_1 = \{\alpha.B.S\}_{k_{ad}}$; $m_2 = \{A.\gamma\}_{k_{bc}}$; $\ulcorner \alpha \urcorner = \{A, D\}$; $\ulcorner k_{ad}^{-1} \urcorner = \{A, D\}$; $\ulcorner k_{bc}^{-1} \urcorner = \{B, C\}$;
 • $\mathcal{W}_{p, F_{MAX}^{EK}}(\alpha, m_1)$
 $= \{\text{Definition 6.1}\}$

$$\bigcap_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m' \sigma' = m_1}} F_{MAX}^{EK}(\alpha, \partial[\bar{\alpha}]m' \sigma') = \bigcap_{\substack{\{\{\alpha.B.X\}_{k_{ad}}, \{\alpha.Y.S\}_{k_{ad}}\} \\ \sigma' = \{X \mapsto S, Y \mapsto B\}}} F_{MAX}^{EK}(\alpha, \partial[\bar{\alpha}]m' \sigma')$$

 $= \{\mathcal{W}_{p, F_{MAX}^{EK}} \text{ is well-formed from the proposition 6.2}\}$
 $F_{MAX}^{EK}(\alpha, \partial[\bar{\alpha}]\{\alpha.B.X\}_{k_{ad}}[X \mapsto S]) \sqcap F_{MAX}^{EK}(\alpha, \partial[\bar{\alpha}]\{\alpha.Y.S\}_{k_{ad}}[Y \mapsto B])$
 $= \{\text{Definition 5.2 and derivation in 5.1}\}$
 $F_{MAX}^{EK}(\alpha, \{\alpha.B\}_{k_{ad}}) \sqcap F_{MAX}^{EK}(\alpha, \{\alpha.S\}_{k_{ad}})$
 $= \{\text{Definition of } F_{MAX}^{EK}\}$
 $\{B, A, D\} \cup \{S, A, D\} = \{B, A, D, S\}$
 • $\mathcal{W}_{p, F_{MAX}^{EK}}(\gamma, m_2)$
 $= \{\text{Definition 6.1}\}$

$$\bigcap_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m' \sigma' = m_2}} F_{MAX}^{EK}(\gamma, \partial[\bar{\gamma}]m' \sigma') = \bigcap_{\substack{\{\{A.Z\}_{k_{bc}}\} \\ \sigma' = \{Z \mapsto \gamma\}}} F_{MAX}^{EK}(\gamma, \partial[\bar{\gamma}]m' \sigma') =$$

 $F_{MAX}^{EK}(\gamma, \partial[\bar{\gamma}]\{A.Z\}_{k_{bc}}[Z \mapsto \gamma])$
 $= \{\text{Definition 5.2}\}$
 $F_{MAX}^{EK}(Z, \partial[\bar{Z}]\{A.Z\}_{k_{bc}})$
 $= \{\text{Derivation in 5.1}\}$
 $F_{MAX}^{EK}(Z, \{A.Z\}_{k_{bc}})$
 $= \{\text{Definition of } F_{MAX}^{EK}\}$
 $\{A, B, C\}$

Lemma 6.4.

Let $m \in \mathcal{M}_p^G$ and $\mathcal{W}_{p, F}$ be a witness-function inside the protection of an external key.
 $\forall \sigma \in \Gamma, \forall \alpha \in \mathcal{A}(\mathcal{M}_p)$ we have:

$$F(\alpha, \partial[\bar{\alpha}]m) \sqsupseteq \mathcal{W}_{p, F}(\alpha, m\sigma) \sqsupseteq \bigcap_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m' \sigma' = m\sigma}} F(\alpha, \partial[\bar{\alpha}]m' \sigma')$$

Proof.

For any $\sigma \in \Gamma$ we have:

- $F(\alpha, \partial[\bar{\alpha}]m) \subseteq \mathcal{W}_{p, F}(\alpha, m\sigma)$: since m is obviously one element of the set $\{m' \in \mathcal{M}_p^G \mid \exists \sigma' \in \Gamma. m' \sigma' = m\sigma\}$ of calculation of $\mathcal{W}_{p, F}(\alpha, m\sigma)$ (i.e. m is a trivial source of $m\sigma$) and since $F(\alpha, \partial[\bar{\alpha}]m\sigma)$ does not depend on σ because, by construction, it depends only on the static part of m (denoted simply by $F(\alpha, \partial[\bar{\alpha}]m)$).
- $\mathcal{W}_{p, F}(\alpha, m\sigma) \subseteq \bigcup_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m' \sigma' = m\sigma}} F(\alpha, \partial[\bar{\alpha}]m' \sigma')$: since for all $m \in \mathcal{M}_p^G$ the set $\{m' \in \mathcal{M}_p^G \mid \exists \sigma' \in \Gamma. m' \sigma' = m\sigma\}$ (unifications) is obviously larger than the set $\{m' \in \mathcal{M}_p^G \mid \exists \sigma' \in \Gamma. m' \sigma' = m\sigma\}$ of sources of $m\sigma$ in \mathcal{M}_p^G .

From these two facts and since \mathcal{L}^\sqsupseteq is a lattice, we have the result in the lemma 6.4

6.3 Protocol correctness with a Witness-Function Theorem

Now, we give the protocol analysis with a Witness-Function theorem that sets a criterion for protocols correctness with respect to the secrecy property. The result in the theorem 6.5 derives directly from the proposition 6.2, the lemma 6.4 and the theorem 3.7. The independence of the criterion stated by the theorem 6.5 of all substitutions fully replaces the condition of full-invariance by substitution stated in Houmani's work [8, 11], and hence any decision made on the generalized roles could be propagated to valid traces.

Theorem 6.5. (Protocol analysis with a Witness-Function)

Let $\mathcal{W}_{p,F}$ be a witness-function inside the protection of an external key.

A sufficient condition of correctness of p with respect to the secrecy property is:

$\forall R.r \in R_G(p), \forall \alpha \in \mathcal{A}(r^+)$ we have:

$$\bigcap_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m'\sigma' = r^+\sigma'}} F(\alpha, \partial[\bar{\alpha}]m'\sigma') \supseteq \ulcorner \alpha \urcorner \sqcap F(\alpha, \partial[\bar{\alpha}]R^-)$$

Proof.

Suppose we have: $\forall R.r \in R_G(p), \forall \alpha \in \mathcal{A}(r^+)$

$$\bigcap_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m'\sigma' = r^+\sigma'}} F(\alpha, \partial[\bar{\alpha}]m'\sigma') \supseteq \ulcorner \alpha \urcorner \sqcap F(\alpha, \partial[\bar{\alpha}]R^-) \quad (15)$$

From the lemma 6.4 and since \mathcal{L}^\sqsupset is a lattice we have for all $\sigma \in \Gamma$:

$$\forall \alpha \in \mathcal{A}(\mathcal{M}_p). \mathcal{W}_{p,F}(\alpha, r^+\sigma) \supseteq \bigcap_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m'\sigma' = r^+\sigma'}} F(\alpha, \partial[\bar{\alpha}]m'\sigma') \quad (16)$$

and

$$\forall \alpha \in \mathcal{A}(\mathcal{M}_p). \ulcorner \alpha \urcorner \sqcap F(\alpha, \partial[\bar{\alpha}]R^-) \supseteq \ulcorner \alpha \urcorner \sqcap \mathcal{W}_{p,F}(\alpha, R^-\sigma) \quad (17)$$

From 15, 16 and 17 we have:

$$\forall \alpha \in \mathcal{A}(r^+\sigma). \mathcal{W}_{p,F}(\alpha, r^+\sigma) \supseteq \ulcorner \alpha \urcorner \sqcap \mathcal{W}_{p,F}(\alpha, R^-\sigma) \quad (18)$$

From the proposition 6.2 $\mathcal{W}_{p,F}$ is \mathcal{C} -reliable, then we have from the theorem 3.1 and 18:

p is correct with respect to the secrecy property

7. NSL PROTOCOL ANALYSIS WITH A WITNESS-FUNCTION

In this section, we analyze the NSL protocol with a witness-function. First, let's recall it:

$$\begin{aligned} m_1 : A &\longrightarrow B : \{N_a.A\}_{k_b} \\ m_2 : B &\longrightarrow A : \{B.N_a\}_{k_a} \cdot \{B.N_b\}_{k_a} \\ m_3 : A &\longrightarrow B : A.B.\{N_b\}_{k_b} \end{aligned}$$

The generalized roles of NSL protocol in a role-based specification are $\mathcal{R}_G(p_{NSL}) = \{A_G^1, A_G^2, B_G^1, B_G^2\}$ where:

$$\begin{aligned} A_G^1 &= i.1 \quad A \longrightarrow I(B) : \{N_a^i.A\}_{k_b} \\ A_G^2 &= i.1 \quad A \longrightarrow I(B) : \{N_a^i.A\}_{k_b} \\ &\quad i.2 \quad I(B) \longrightarrow A : \{B.N_a^i\}_{k_a} \cdot \{B.X\}_{k_a} \\ &\quad i.3 \quad A \longrightarrow I(B) : A.B.\{X\}_{k_b} \\ B_G^1 &= i.1 \quad I(A) \longrightarrow B : \{Y.A\}_{k_b} \\ &\quad i.2 \quad B \longrightarrow I(A) : \{B.Y\}_{k_a} \cdot \{B.N_b^i\}_{k_a} \\ B_G^2 &= i.1 \quad I(A) \longrightarrow B : \{Y.A\}_{k_b} \\ &\quad i.2 \quad B \longrightarrow I(A) : \{B.Y\}_{k_a} \cdot \{B.N_b^i\}_{k_a} \\ &\quad i.3 \quad I(A) \longrightarrow B : A.B.\{N_b^i\}_{k_b} \end{aligned}$$

Let's have a context of verification such that: $\lceil A \rceil = \perp$; $\lceil B \rceil = \perp$; $\lceil N_a^i \rceil = \{A, B\}$; $\lceil N_b^i \rceil = \{A, B\}$; $\lceil k_a^{-1} \rceil = \{A\}$; $\lceil k_b^{-1} \rceil = \{B\}$; $(\mathcal{L}, \sqsupseteq, \sqcup, \sqcap, \perp, \top) = (2^{\mathcal{I}}, \subseteq, \cap, \cup, \mathcal{I}, \emptyset)$; $\mathcal{I} = \{I, A, B, A_1, A_2, B_1, B_2, \dots\}$; The set of messages generated by the protocol is $\mathcal{M}_p^G = \{\{N_{A_1}.A_1\}_{k_{B_1}}, \{B_2.N_{A_2}\}_{k_{A_2}}, \{B_3.X_1\}_{k_{A_3}}, \{X_2\}_{k_{B_4}}, \{Y_1.A_4\}_{k_{B_5}}, \{B_6.Y_2\}_{k_{A_5}}, \{B_7.N_{B_7}\}_{k_{A_6}}, \{N_{B_8}\}_{k_{B_8}}\}$

The variables are denoted by X_1, X_2, Y_1 and Y_2 ;

The static names are denoted by $N_{A_1}, A_1, k_{B_1}, B_2, N_{A_2}, k_{A_2}, B_3, k_{A_3}, k_{B_4}, A_4, k_{B_5}, B_6, k_{A_5}, B_7, N_{B_7}, k_{A_6}, N_{B_8}$ and k_{B_8} .

After the elimination of duplicates, $\mathcal{M}_p^G = \{\{N_{A_1}.A_1\}_{k_{B_1}}, \{B_2.N_{A_2}\}_{k_{A_2}}, \{B_3.X_1\}_{k_{A_3}}, \{X_2\}_{k_{B_4}}, \{Y_1.A_4\}_{k_{B_5}}, \{B_7.N_{B_7}\}_{k_{A_6}}, \{N_{B_8}\}_{k_{B_8}}\}$

Let's select the Witness-Function as follows:

$$p = NSL; F = F_{MAK}^{EK}; \mathcal{W}_{p,F}(\alpha, m\sigma) = \bigcap_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m'\sigma' = m\sigma}} F(\alpha, \partial[\bar{\alpha}]m'\sigma');$$

Let's denote the lower bound of the Witness-Function by:

$$\mathcal{W}'_{p,F}(\alpha, r^+) = \bigcap_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m'\sigma' = r^+ \sigma'}} F(\alpha, \partial[\bar{\alpha}]m'\sigma')$$

7.1 Results and Interpretation

The results of analysis of the NSL protocol are summarized in the table 2. We notice from the Table 2 that

α	Role	R^-	r^+	$\mathcal{W}'_{p,F}(\alpha, r^+)$	$\lceil \alpha \rceil$	$F(\alpha, \partial[\bar{\alpha}]R^-)$	Theorem 6.5
N_a^i	A	\emptyset	$\{A.N_a^i\}_{k_b}$	$\{A, B\}$	$\{A, B\}$	\top	Respected
X	A	$\{B.N_a^i\}_{k_a} \cdot \{B.X\}_{k_a}$	$A.B \cdot \{X\}_{k_b}$	$\{B\}$	$\lceil X \rceil$	$\{A, B\}$	Respected
Y	B	$\{A.Y\}_{k_b}$	$\{Y.N_b^i.B\}_{k_a}$	$\{A, B\}$	$\lceil Y \rceil$	$\{A, B\}$	Respected
N_b^i	B	$\{A.Y\}_{k_b}$	$\{Y.N_b^i.B\}_{k_a}$	$\{A, B\}$	$\{A, B\}$	$\{A, B\}$	Respected

Table 2: Compliance of NSL Protocol with the Theorem 6.5

the NSL protocol respects the correctness criterion stated in the theorem 6.5, then it is correct with respect to the secrecy property.

8. A VARIATION OF THE NEEDHAM-SCHROEDER PROTOCOL ANALYSIS WITH A WITNESS-FUNCTION

In this section, we analyze a variation of the Needham-Schroeder protocol with the witness-function. First,

let's recall it:

$$\begin{aligned} 1 : A &\longrightarrow B : \{A.N_a\}_{k_b} \\ 2 : B &\longrightarrow A : \{N_a.N_b.B\}_{k_a} \\ 3 : A &\longrightarrow B : \{N_b\}_{k_b} \end{aligned}$$

The generalized roles of the protocol are $\mathcal{R}_G^p = \{A_G^l, A_G^2, B_G^l, B_G^2\}$ where:

$$\begin{aligned} A_G^l &= i.1 \quad A \longrightarrow I(B) : \{A.N_a^i\}_{k_b} \\ A_G^2 &= i.1 \quad A \longrightarrow I(B) : \{A.N_a^i\}_{k_b} \\ &\quad i.2 \quad I(B) \longrightarrow A : \{N_a^i.X.B\}_{k_a} \\ &\quad i.3 \quad A \longrightarrow I(B) : \{X\}_{k_b} \\ B_G^l &= i.1 \quad I(A) \longrightarrow B : \{A.Y\}_{k_b} \\ &\quad i.2 \quad B \longrightarrow I(A) : \{Y.N_b^i.B\}_{k_a} \\ B_G^2 &= i.1 \quad I(A) \longrightarrow B : \{A.Y\}_{k_b} \\ &\quad i.2 \quad B \longrightarrow I(A) : \{Y.N_b^i.B\}_{k_a} \\ &\quad i.3 \quad I(A) \longrightarrow B : \{N_b^i\}_{k_b} \end{aligned}$$

Let's have a context of verification such that: $\ulcorner A \urcorner = \perp$; $\ulcorner B \urcorner = \perp$; $\ulcorner N_a^i \urcorner = \{A, B\}$ (secret between A and B); $\ulcorner N_b^i \urcorner = \{A, B\}$ (secret between A and B); $\ulcorner k_a^{-1} \urcorner = \{A\}$; $\ulcorner k_b^{-1} \urcorner = \{B\}$; $(\mathcal{L}, \supseteq, \sqcup, \sqcap, \perp, \top) = (2^{\mathcal{I}}, \subseteq, \cap, \cup, \mathcal{I}, \emptyset)$; $\mathcal{I} = \{I(\text{intruder}), A, B, C, A_1, A_2, B_1, B_2, \dots\}$;

The set of messages generated by the protocol is $\mathcal{M}_p^G = \{\{A_1.N_{A_1}\}_{k_{B_1}}, \{N_{A_2}.X_1.B_2\}_{k_{A_2}}, \{X_2\}_{k_{B_3}}, \{A_3.Y_1\}_{k_{B_4}}, \{Y_2.N_{B_5}.B_5\}_{k_{A_4}}, \{N_{B_6}\}_{k_{B_6}}\}$;

The variables are denoted by X_1, X_2, Y_1 and Y_2 ;

The static names are denoted by $A_1, N_{A_1}, k_{B_1}, N_{A_2}, B_2, k_{A_2}, k_{B_3}, A_3, k_{B_4}, N_{B_5}, B_5, k_{A_4}, N_{B_6}$ and k_{B_6} ;

Let's select the Witness-Function as follows:

$$p = NS; F = F_{MAX}^{EK}; \mathcal{W}_{p,F}(\alpha, m\sigma) = \bigcap_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m'\sigma' = m\sigma}} F(\alpha, \partial[\overline{\alpha}]m'\sigma');$$

Let's denote the lower bound of the Witness-Function by:

$$\mathcal{W}'_{p,F}(\alpha, r^+) = \bigcap_{\substack{m' \in \mathcal{M}_p^G \\ \exists \sigma' \in \Gamma. m'\sigma' = r^+\sigma'}} F(\alpha, \partial[\overline{\alpha}]m'\sigma')$$

8.1 Results and interpretation

The results of the analysis of the variation of Needham-Schroeder protocol are summarized in Table 3.

α	Role	R^-	r^+	$\mathcal{W}'_{p,F}(\alpha, r^+)$	$\ulcorner \alpha \urcorner$	$F(\alpha, \partial[\overline{\alpha}]R^-)$	Theorem 6.5
N_a^i	A	\emptyset	$\{A.N_a^i\}_{k_b}$	$\{A, B\}$	$\{A, B\}$	\top	Respected
X	A	$\{N_a^i.X.B\}_{k_a}$	$\{X\}_{k_b}$	$\{B\}$	$\ulcorner X \urcorner$	$\{A, B\}$	Respected
Y	B	$\{A.Y\}_{k_b}$	$\{Y.N_b^i.B\}_{k_a}$	$\{A, B\}$	$\ulcorner Y \urcorner$	$\{A, B\}$	Respected
N_b^i	B	$\{A.Y\}_{k_b}$	$\{Y.N_b^i.B\}_{k_a}$	$\{A, B, A_3\}$	$\{A, B\}$	$\{A, B\}$	Not Respected

Table 3: Compliance of the Variation of Needham-Schroeder Protocol with the Theorem 6.5

We notice from the Table 3 that the variation of Needham-Schroeder protocol does not respect the correctness criterion set by the theorem 6.5 when analyzed with the witness-function $\mathcal{W}_{p,NS,F_{MAX}^{EK}}$. Therefore, we cannot deduce anything regarding its correctness with respect to the secrecy property. The non-growth of the protocol is localized in the sending step of the generalized role of B and it is due to a possible malicious neighbor (denoted by A_3 in our analysis) that could be inserted beside the nonce N_B^i . In the literature, we report a flaw that operates on the decay of the level of security of the nonce N_B^i in the generalized role of B . This flaw is described by the attack scenario in the Figure.1.

9. CONCLUSION AND FUTURE WORK

In this paper, we gave relaxed conditions on interpretation functions to be reliable to run an analysis of a cryptographic protocol on valid traces for the property of secrecy. Afterward, we gave a whole class of reliable functions based on selections inside the external key. Then we introduced the witness-functions that offer two bounds which are independent of substitutions and therefore enable an analysis on the generalized roles of a protocol in a role-based specification. We experimented our approach on real protocols and we showed that a witness-function can even help to locate flaws. It was successful to prove the correctness of others too. In a future work, we intend to define more witness-functions based on the selection inside other keys (other than the most external key) like the most internal key or all encryption keys together, so that we could efficiently deal with algebraic properties in a non-empty equational theory [19–21] such the Diffie-Hellman property. We intend also to take advantage of the bounds of witness-functions to consider exchanged messages in a protocol as encryption keys.

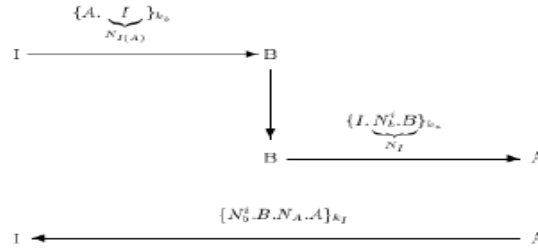


Figure 1: Attack Scenario on the Variation of the Needham-Schroeder Protocol

REFERENCES

- [1] S. Schneider, "Verifying authentication protocols in csp," IEEE Trans. Software Eng., vol. 24, no. 9, pp. 741–758, 1998.
- [2] S. Schneider, "Security properties and csp," in IEEE Symposium on Security and Privacy, pp. 174–187, 1996.
- [3] S. A. Schneider and R. Delicata, "Verifying security protocols: An application of csp," in 25 Years Communicating Sequential Processes, pp. 243–263, 2004.
- [4] J. Heather and S. Schneider, "A decision procedure for the existence of a rank function," J. Comput. Secur., vol. 13, pp. 317–344, Mar. 2005.
- [5] M. Abadi, "Secrecy by typing in security protocols," Journal of the ACM, vol. 46, pp. 611–638, 1998.
- [6] M. Abadi and A. D. Gordon, "Reasoning about cryptographic protocols in the spi calculus," in CONCUR, pp. 59–73, 1997.
- [7] M. Abadi and A. D. Gordon, "A calculus for cryptographic protocols: The spi calculus," in ACM Conference on Computer and Communications Security, pp. 36–47, 1997.
- [8] H. Houmani and M. Mejri, "Practical and universal interpretation functions for secrecy," in SECRIPT, pp. 157–164, 2007.
- [9] H. Houmani and M. Mejri, "Ensuring the correctness of cryptographic protocols with respect to secrecy," in SECRIPT, pp. 184–189, 2008.
- [10] H. Houmani and M. Mejri, "Formal analysis of set and nsl protocols using the interpretation functions-based method," Journal Comp. Netw. and Communic., vol. 2012, 2012.
- [11] H. Houmani, M. Mejri, and H. Fujita, "Secrecy of cryptographic protocols under equational theory," Knowl.-Based Syst., vol. 22, no. 3, pp. 160–173, 2009.
- [12] D. Dolev and A. C.-C. Yao, "On the security of public key protocols," IEEE Transactions on Information Theory, vol. 29, no. 2, pp. 198–207, 1983.
- [13] J. Fattahi, M. Mejri, and H. Houmani, "Context of verification and role-based specification http://web_security.fsg.ulaval.ca/lab/sites/default/files/WF/Ind/Context.pdf," no. 4, pp. 1–4, 2014.
- [14] M. Debbabi, Y. Legaré, and M. Mejri, "An environment for the specification and analysis of cryptoprotocols," in ACSAC, pp. 321–332, 1998.
- [15] M. Debbabi, M. Mejri, N. Tawbi, and I. Yahmadi, "Formal automatic verification of authentication cryptographic protocols," in ICFEM, pp. 50–59, 1997.
- [16] M. Debbabi, M. Mejri, N. Tawbi, and I. Yahmadi, "From protocol specifications to flaws and attack scenarios: An automatic and formal algorithm," in WETICE, pp. 256–262, 1997.
- [17] J. Fattahi, M. Mejri, and H. Houmani, "The witness-functions: Proofs and intermediate results. http://web_security.fsg.ulaval.ca/lab/sites/default/files/WF/Ind/WitFunProofs.pdf," no. 26, pp. 1–26, 2014.
- [18] B. Blanchet, "Automatic verification of correspondences for security protocols," Journal of Computer Security, vol. 17, no. 4, pp. 363–434, 2009.
- [19] H. Comon-Lundh, V. Cortier, and E. Zalinescu, "Deciding security properties for cryptographic protocols. application to key cycles," ACM Trans. Comput. Log., vol. 11, no. 2, 2010.
- [20] V. Cortier and S. Delaune, "Decidability and combination results for two notions of knowledge in security protocols," J. Autom. Reasoning, vol. 48, no. 4, pp. 441–487, 2012.
- [21] V. Cortier, S. Kremer, and B. Warinschi, "A survey of symbolic methods in computational analysis of cryptographic systems," J. Autom. Reasoning, vol. 46, no. 3-4, pp. 225–259, 2011.
- [22] F. Baader and T. Nipkow, Term rewriting and all that. Cambridge University Press, 1998.
- [23] N. Dershowitz and D. A. Plaisted, "Rewriting," in Handbook of Automated Reasoning, pp. 535–610, 2001.
- [24] H. Comon-Lundh, C. Kirchner, and H. Kirchner, eds., Rewriting, Computation and Proof, Essays Dedicated to Jean-Pierre Jouannaud on the Occasion of His 60th Birthday, vol. 4600 of Lecture Notes in Computer Science, Springer, 2007.

WEB SERVICE COMPOSITION BASED ON POPULARITY

Selwa Elfirdoussi¹, Zahi Jarir¹, Mohamed QUAFAFOU²

¹Laboratory LISI, Computer Science Department, Faculty of Sciences,
Cadi Ayyad University, BP 2390, Marrakech, Morocco

s.elfirdoussi@ced.uca.ma ; jarir@uca.ma

²LSIS – UMR CNRS 6168, Domaine universitaire de St Jérôme,
F-13397, Marseille Cedex 20, France

mohamed.quafafou@univ-amu.fr

ABSTRACT

In Web Service research, providing methods and tools to cater for automatic composition of services on the Web is still the object of ongoing research activity. Despite the proposed approaches this issue remains open. In this paper we propose a seamless way to compose automatically web services from expressed abstract process model. The process of composition is based on web service popularity concept. To validate our approach an implementation is presented.

KEYWORDS

Automatic Web service composition, Web service selection, Web service popularity.

1. INTRODUCTION

Web service composition involves combining and coordinating a set of web services with the purpose of achieving functionality that cannot be realized through existing services. The goal of this process is to arrange multiple services into workflows supplying complex and specific user's needs. Hence automating the composition is a really complex and comprehensive problem [1, 7]. In this case the challenge is twofold: (a) How to build a required workflow? and (b) How to discover a more appropriate web service for each node in the built workflow according to user's requirements? In this paper we focus our work to suggest a solution for the second challenge. The first one concerns a future work.

When designing or building a workflow of composite web service statically or dynamically, different approaches are proposed in literature to affect web service to each node of this workflow; for instance each service can be affected in workflow either manually [2, 11], automatically [7, 9] or semi-automatically [3, 10].

For this problem, Said et al. [17] proposes a new SOA architecture called "GenericSOA" that allows dealing with legacy systems problem and enhancing SOA elasticity. The proposed architecture aims to easily integrating the newly developed software components. The main idea behind GenericSOA is to support its users by a set of predefined task templates. These templates can be used in building the new developed services that can be easily integrated in a loosely coupled way to compose the target system.

A more appropriate approach is to affect these web services dynamically since published web services is growing more and more, web services are going offline, new services becoming online, and existing services changing their characteristics, user's requirements are contextual and thereafter not static, dealing with failures that may occur when web services are not available, etc. Consequently and due to the increase of published web services, finding the suitable WS that satisfies the user goals among discovered web services still needs deep investigations. Certainly, QoS requirements represent a more appropriate and decisive factor to distinguish similar WSs. A lot of research efforts in this direction have been made but are still limited due to the complexity and diversity of QoS constraints. In this paper we propose a seamless way to compose web services based on abstract process model representing a needed workflow.

Assuming that for each node a collection of web services have been discovered from web service registry depending on user's requirements, so choosing an appropriate one for each node requires in addition to take into consideration their relationship in order to resolve any conflicts and/or inconsistencies between linked web services. Since inconsistencies may occur at runtime, it may be necessary to predict such events to ensure that the composition will run correctly. An important challenge in providing an automatic web services composition facility is dealing with failures that may occur, for instance as a result of context changes or missing service descriptions.

To affect a more adequate web services in composition taking in consideration their correct relationship, we present in this paper the design and implementation of a framework for web services composition. Experimental evaluation demonstrates that the framework provides an efficient and scalable solution. Furthermore, it shows that our framework transforms goals successfully, increasing the utility of achieved objectives without imposing a prohibitive composition time overhead.

In this paper we suggest a web search composition engine that has the faculty to compose automatically web services based on their popularity. We propose in the first hand a workflow design that can be used to draw the composition diagram, and contain a text that the user can use to define his query. The workflow selects the appropriate web services relative to the query based the most popular and generates the composition according the BPEL process model [2] as the result.

The paper is organized as follows; we describe in section II a related work refereeing to automatic web service composition. In section III we describe and explain our proposed algorithm of the automated composition of web services and the rule of how we define the most popular web service by query. In section IV we describe the implementation of our approach in DIVISE and give an evaluation in the system. We conclude in section V.

2. RELATED WORK

Web service composition lets developers create applications on top of service-oriented computing native description, discovery, and communication capabilities. Such applications are rapidly deployable and offer developers reuse possibilities and users seamless access to a variety of complex services. There are many existing approaches to service composition, ranging from abstract methods to those aiming to be industry standards. As defined in the first section, our approach is based on automatic selection and composition, we define two process. In this section, we describe some research proposed for the automatic Web service selection for composition and automatic process for composition.

Recently, Raj et al. [16] propose an approach on identifying the most appropriate service based on the user's preferences of the requested WS. The given WS description may contain the

parameters, which may have relations with the requested WS of the specific domain in different aspects like name, parameters and types. The domain specific WS classification can be done using Naive Bayes classification algorithm. But this method has some limitations of not considering functionality based classification. The coupling and cohesion properties are not considered in the composed WSs.

2.1. Automate web services selection for composition

The author [4] proposes a novel approach of semantics-based matchmaking, which is named process-context aware matchmaking. This process locates the suitable service during web service composite modeling. During matchmaking, the approach utilizes not only the semantics of technical process but also those of business process of a registered service, thus further improving the precision of matchmaking. The process-context aware matchmaking was integrated with business-process-driven web service composition in a cohesive development environment based on Eclipse. The work describes a way to match web services for composition but doesn't integrate the composition of web services.

To improve the exactitude of a Web service search, Ye and Zhang [9] proposed a method that explicitly specifies the functional semantics of services. They specified a service and a user requirement using object, action and constraints as well as input and output parameters. Utilizing this information, they found a service to satisfy the user requirement. However, they did not consider how the popularity of web services can be applied to service composition.

The authors [11] tried to improve the accuracy of automatic matchmaking of Web services by taking into account the knowledge of past matchmaking experiences for the requested task. In their method, service execution experiences are modeled using case based reasoning. This method can be helpful for improving the exactitude of composite service, but it's still packed of complex problems related to the composition process.

2.2. Automatic Web Services Composition

The University of Georgia implements an extension [5] of GraphPlan [6], an AI planning algorithm, to automatically generate the control flow of a Web process. This extension is does not cover the preconditions and effects of the operations, we also take into consideration in the planning algorithm the structure and semantics of the input and output messages. The approach was presented to solve both the process heterogeneity and data heterogeneity problems. And the system generates outputs, an executable BPEL file which correctly solves non-trivial real-world process specifications. The authors described in parts of their paper the project proposed by BPEL [2] to automate the composition, but neither one of those works propose the automatic selection of web service using the behavior experience named popularity.

In the some context, [8] proposes a composition method that explicitly specifies the uses of functional semantics of web services. Specifically, the proposed method is based on a graph model, which represents the functional semantics of Web services. In this approach, the service functionality of a service is represented by a pair of its action and the object of the action. The information about services is organized and stored in a proposed two-layer graph model. Given a user request, they search for composition paths in the graph model and construct a composite service from the paths discovered. However, the web services selection is not taking in consideration the notoriety to get link in the schema composition.

Liu, Ranganathan and Riabov [10] propose a Web service model in which inputs and outputs of services are expressed using RDF graph pattern, as well as a domain ontology. They improve the

exactitude of composite services without preconditions and effects using semantic propagation based on graph substitution and also they don't take a request user when selecting web services.

3. AUTOMATIC COMPOSITION ALGORITHM

Web services are composed based on QoS metrics [15] by evaluating the utility function and thus maximizing the overall QoS using the hybrid approach in composition patterns. In our approach, we improve the definition presented in [8] that a composite Web service can be defined as a set of transition systems, which has multiple states, arcs and available actions in certain states and represents transitions from an initial state accepting user inputs to a final state providing requested outputs and effects as shown in Figure 1.

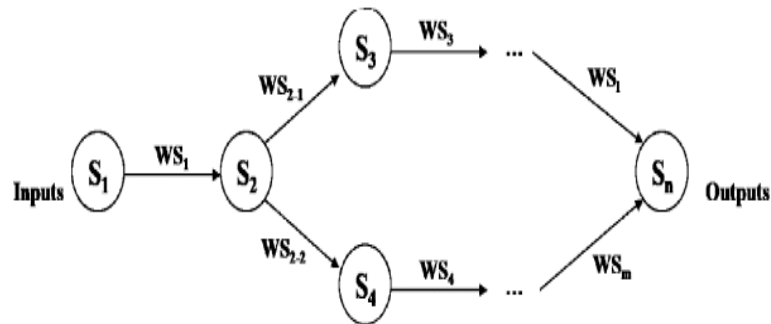


Figure 1: A composite Web Service Representation

In figure 1, “S” represents the web service input variable and “WS” is the web service selected using the query parameter requested by a user while building a graph. As presented in Figure 1, the objective of composing web services is to select required web service for each node in graph based on defined inputs and outputs.

However, the schema mentioned above assumes that every available service is defined by an input and output parameter. For our approach, we present web services as a triple (Input, Query, Output) when Query is the request relative to find the web service. So, we suggest specifying Web Service by its functionality and the sets of I/O as follows:

Web service = (service functionality, input set, output set)

To build the automatic web services composition, we propose, in our algorithm, three fundamental process : (1) Web service composite designer, (2) Web service selection and (3) Web service composite generator. These processes will be presented in details in the following sub-sections. The workflow used is defined in the algorithm defined bellow:

- *Initialize Web Service Composition by defining for each branch of our graph the input, query and output parameter*
- *Search Service Based Popularity: based on query, we calculate the web service criteria value to select the best one*
- *Compose Web Service: define the structure of our composite web service using the web service selected*
- *Generate BPEL File and annexed File: create the web service composite by defining the sequence of activity (Receive, Assign, Invoke and Replay)*

Finally, to complete the composition task, we define the workflow which combines the graph process successively respecting the design produced in the first process. The process selection can be used as a web service discovery to get the service responding to request users sorted by popularity. Note, in our approach the popularity is relative to two criteria, the first one is relative to the frequency of use and the second one will be linked to the appropriate web service.

3.1 Web service composite designer

It's important to begin the web service composition to propose a web service workflow designer. The result proposed in our approach is to generate a BPEL File presenting the web service composite, this module will be used by the client to define the sequence of input, output and a query of Web Service. The architect of our design follows the diagram shown in figure 1; the difference is for each "Invoke" activity in the annotated BPEL in a BPEL Design [13] is that it will be presented by an input text to write the query. In particular, this query information will be used to select the web services in the next process. So, our designer process allows user to draw a sequence of web service call specifying the input and output parameter of each one. it will take a decision on its final result to define the logical orchestration.

3.2 Web Service Selection

This process presents the selection of web services using the user's query. There are two methods for this selection. In the first hand we discover the web service responding the query defined in the input "S", the WS registry will give us a list of result this list will be shorted by the popularity and in the second hand we choose the web service most popular that we present in our workflow to define the "PartnerLink" role defined in the BPEL Process [13].

As defined in the first section, our approach is based on web service selection to automate the composition. In the figure 1, we define the "WS1" designed the first web service, this one will be selected using the frequency presenting the number of web services uses divided by the number of month the formula (1) proposed is [14] :

$$Nb(Invoke(WS))/Nb(Month) = frequency \quad (1)$$

In the second selection, we use the same criteria of frequency and we add the dependency of the previous web service selected presented by the behaviour experience of the link. There are two methods:

- The first one defined in (2) the number of link between the previous web services selected.

$$Nb(Link(WS1, WS2)) \square \square = Notoriety \quad \square \square \square$$
- The second one in (3) but when the previous web service concern two or more the method will be used for the link with the both of them.

$$Nb(Link(WS1, WS2, WS3, \dots WS_n)) \square \square \square = \square Notoriety \square \square \square$$

The process selection provides for each query proposed by user the best web service. In the next step, we extract the web service information that's will be necessary for generating the composite web service as the "accesspoint", the operation and the input and output parameters.

3.3 Web Service composite Generator

The process of web service composition generator is based on creating a BPEL file present the result of composition. This process will be used in two steps. (a) in the first time, we build the BPEL structure based the design provided by the users (b) in the second hand, we execute the

web service selection to add for each invoke tag the web service chooses and we describe the “PartnerLink” tag and “Import” tag to finally generate the web service composite presented by the BPEL File.

4. IMPLEMENTATION AND EVALUATION

The implementation of our approach is based on the BPEL definition [13], because the BPEL process is the most complete and popular language to generate web service composite. In order to test the effectiveness of our proposed algorithm, we have used our implemented framework Discovery and Visual Interactive Web Service Engine (DIVISE) to expose an evaluation of our approach to compose web service by implementing the different algorithms exposed in section III.

Note that DIVISE is an engine that’s has the advantage to discover a required simple, composite or semantic web service and to help user to select the more appropriate Web service from a returned list. This list contains in addition to classical web service information a rate of its previous invocations defined as frequency or detailed description detailed of web service, which is useful for calculating the number of link used between web services [14]. This tool is written in Java and mainly built on the Eclipse- frameworks EMF and GEF and thus is also realized as a set of Eclipse.

4.1 Process Implementation

Referring to our algorithm, we propose to create the orchestration of the web service result in four steps. Each one is presented in a principal class executing as an action as presented bellow. In this section, we describe the role of each class and the orchestration of our automatic composition proposition.

- *InitializeWSCompositeAction*
- *SearchServiceBasedPopularity*
- *ComposeWSAction*
- *GenerateBPELFileFromObject*

The designer starts to receive a sequence of request from the user; each request must contain input, query and output parameter. The input and output parameter have to be used successively to define in the last a tree. After each request has been received, the process initializes a vector containing a triple parameter for each web service asked (Input, query, output). Our process control progressively the logic defined in the sequence to get in the first one input as receive parameter and in the last on output as a replay parameter.

Once the model is finished; the schema will be sending to execute the workflow presented as a set of activities. To orchestral this process, we define each level (WS) as a row which contains a set of attributes. This presentation will be used to add web service and proposes in our form the input text relative to each attributes to define the value if necessary. There are some attributes that’s can be inserted automatically under the generator and it’s not will be presented to user as the url of location or the partner link Etc.

```

public class InitializeWSCompositeAction extends Action {

    public String execute(Map params, HttpServletRequest request, HttpServletResponse response) throws Exception {

        try {

            //Reload the processName and the fileName
            String fileName = (String)request.getParameter("fileName");
            String processName = (String)request.getParameter("processName");

            //Add the file name and process name in request attributes
            if(fileName != null) request.setAttribute("fileName",fileName);
            if(processName != null) request.setAttribute("processName",processName);

            //design vectWSSchema contains hashtable of each WS
            //Vector contains inputs variables {{S1},{S2,S3}}
            String vectWebService = request.getParameter("vectWebService");
            Vector vectWSSchema = new Vector();
            if (vectWebService != null && !vectWebService.equals("null") && !vectWebService.equals("")) {
                vectWSSchema = Utils.explodeList(vectWebService, "|", null, true);
            }
        }
    }
}

```

Figure 2 : Initialize Web Service Schema Composition

The second process is relative to web service composition is the selection of web service based on popularity. For this process, we implement the algorithm defined in the previous section to calculate the Web Service Popularity Score by query. We create a method called “CalculateWSPSFromQuery” taking query as a parameter and we choose the web service that has the best score to propose it in our composition. To integrate this web service, we also need some necessary parameter defined in the file description that’s we extract as the operation, input, output parameters etc.

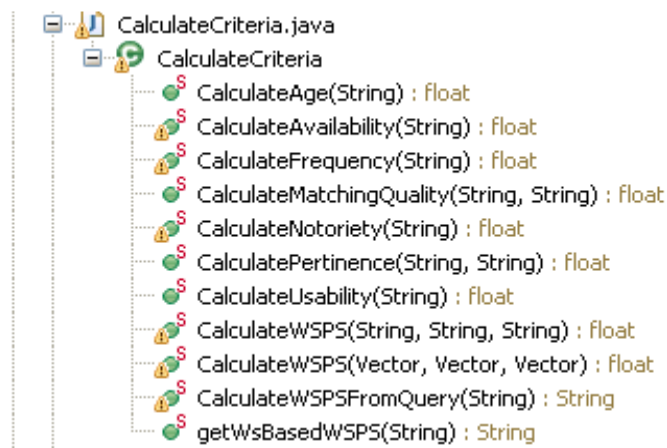


Figure 3 : Methods to calculate Quality Criteria

Based on the schema defined by user and the list of web service selected using the popularity, we create the composition action following the BPEL Model. However, we create a BPEL Structure as defined in the figure 4. In this object, we cover the activities as Receive, Invoke, Assign and Replay to design the sequence of the process and the tag relative to import and variable. To resume, our object contains all BPEL tag defined in the XML definition as presented [12].

```
public class BPELStructure {
    // File Name will be used
    //to generate the BPEL File
    public String fileName;
    //Tag <process>
    public Hashtable processName;
    //Tag <import>, this tag will
    //be generator in the composition generator
    public Vector vectImport;
    //Contains Tags <partnerLinks>,
    //this tag will be generator in the generator
    public Vector vectPartnerLink;
    //Contains Tag <variables>,
    //this tag will be generator in the generator
    public Vector vectVariable;
    //Contains Activity as (<Sequence>,<Invoke>,<Receive>,<Replay>,<forEach>,<While> Ect..)
    public Vector vectActivity;
}
```

Figure 4 : Object of BPEL Structure

This structure defined in figure 4 will be used in the composition process as explained later.

```
public class ComposeWSAction extends Action {

    public String execute(Map params, HttpServletRequest request, HttpServletResponse response) throws Exception {

        String result = "";
        try {

            Vector vectImport = new Vector();
            Vector vectPartnerLink = new Vector();
            Vector vectVariable = new Vector();
            Vector vectActivity = new Vector();

            //construct Object BPEL Structure and generate File
            BPELStructure bpelObject = new BPELStructure();
            //get fileName
            String fileName = (String)request.getParameter("fileName");
            bpelObject.setFileName(fileName);

            //get processName
            String processName = (String)request.getParameter("processName");
            Hashtable rowProcess = new Hashtable();
            rowProcess.put("name", processName);
            rowProcess.put("targetNamespace", "http://"+processName);
            rowProcess.put("suppressJoinFailure", "yes");
            rowProcess.put("xmlns:tns", "http://"+processName);
            rowProcess.put("xmlns:bpel", "http://docs.oasis-open.org/wsbpel/2.0/process/executable");
            bpelObject.setProcessName(rowProcess);
        }
    }
}
```

Figure 5 : Compose Web Service Action

The class defined in the figure 5 allowed composing web service based on the schema proposed by user. This class, in the first hand, defines the process and their attributes as a map. The result is an instance of “BPELStructure” containing the import WSDL, the list of partner list, the variable used by each web service and finally the sequence of activity. Each activity is presented as a map containing key, values of Activity attributes defined in the BPEL language [12].

```
public static String generateBPELFileFromObject(String fileName, BPELStructure bpelObject) throws Exception {
    try
    {
        DocumentBuilderFactory docFactory = DocumentBuilderFactory.newInstance();
        DocumentBuilder docBuilder = docFactory.newDocumentBuilder();

        //root elements
        Document doc = docBuilder.newDocument();

        //process elements
        Element rootElement = doc.createElement("bpel:process");
        Hashtable processRow = bpelObject.getProcessName();
        Enumeration e = processRow.keys();
        while (e.hasMoreElements()) {
            String key = (String) e.nextElement();
            if (processRow.get(key) != null) rootElement.setAttribute(key, (String) processRow.get(key));
        }
        doc.appendChild(rootElement);

        //import elements
        Element importList = doc.createElement("bpel:imports");
        rootElement.appendChild(importList);

        Vector vectImport = bpelObject.getVectImport();
        for (int k=0; k < vectImport.size(); k++){
            Hashtable rowImport = (Hashtable) vectImport.elementAt(k);
            Element importChild = doc.createElement("bpel:import");
            //Tag Attribute defined in hashtable
            e = rowImport.keys();
            while (e.hasMoreElements()) {
                String key = (String) e.nextElement();
                if (rowImport.get(key) != null) importChild.setAttribute(key, (String) rowImport.get(key));
            }
            importList.appendChild(importChild);
        }
    }
}
```

Figure 6 : Method to generate BPEL File

For the implementation of our approach, we have defined a class defined “ComposeUtils” containing each methods useful for the execution of process.

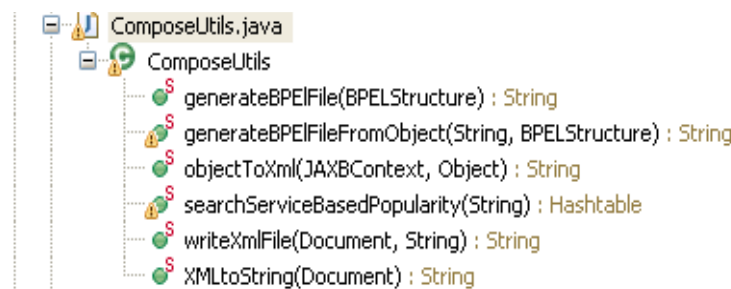
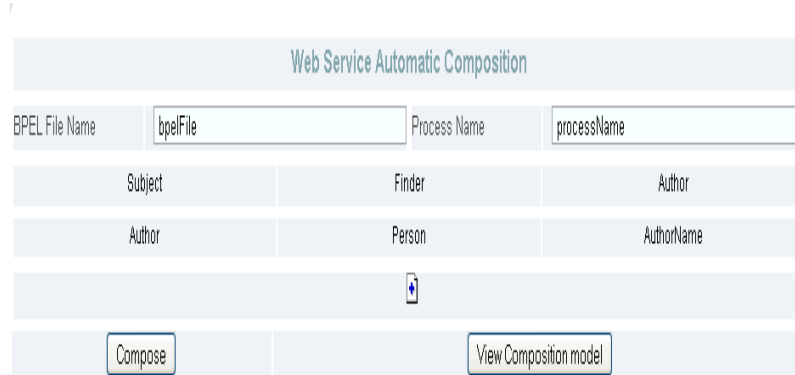


Figure 7 : Utilities for composition process

4.2 Experiment

For improving the efficiencies of our algorithm, we define bellow an example to compose web services using our DIVISE. We have added in the web application menu to compose web services. This link proposes to users to design the workflow process by adding web service request.



The form titled "Web Service Automatic Composition" contains the following fields and buttons:

- BPEL File Name:** A text input field containing "bpelFile".
- Process Name:** A text input field containing "processName".
- Subject:** A text input field containing "Subject".
- Finder:** A text input field containing "Finder".
- Author:** A text input field containing "Author".
- Person:** A text input field containing "Person".
- AuthorName:** A text input field containing "AuthorName".
- Buttons:** "Compose" and "View Composition model".

Figure 8 : Web Service Composite Designer

The frame defined in the figure 8 is proposed to design the workflow that's will be used to compose web service. We have choose to use a link (presented as +) to add web service. This link open a window that's allowing user to add a parameter of web service to be added as presented in figure 9. A refresh of the frame is also sending that saves the web service and proposes to add another. We also, in the last submit operation, a control to the workflow to ensure that the sequence designed is correctly follows and each variable are used as input and output.



The dialog box titled "Add Web Service" contains the following fields and button:

- Input:** A text input field containing "Subject".
- Query:** A text input field containing "Finder".
- Output:** A text input field containing "Author".
- Button:** "Save & Close".

Figure 9 : Frame to add Web Service in model

The final result is presented as a tree containing the sequence web service and their parameters classed from the first to the last. The figure 10 presents the model composition.

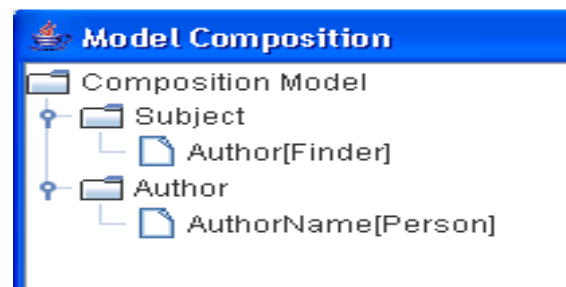


Figure 10 : Tree model composition

After drawing the process, user sends his proposition by clicking compose button. The result of composition is generated in a file containing the orchestration BPEL of our composition process as presented bellow.



Figure 11 : BPEL File generated

From the result frame, we display the BPEL file generated to user that he can, also, download from the link in the top of frame. In our framework, we can invoke our process from the Web Service Invoke module [15] using the file URL.

5. CONCLUSION AND FUTURE WORKS

The contribution of this paper deals with web service composition, particularly building a deployable composition based on abstract process model (abstract graph). The proposed approach is to assign for each node in the defined abstract graph a required web service meeting user requirements and consistencies between composed web services.

In addition we have developed a prototype system and implemented in our DIVISE framework to improve our approach. This prototype has the advantage to compose web services from a design and select the web services deployed in our log Database filtering by frequency and dependency. Our automatic process proposes to user an interface to create and design a BPEL system that defines the sequence of the activities with query input. Each query will be used to select the best web service based on its popularity. After selection the framework DIVISE generates a BPEL code related to build composite web service.

The future work is to integrate in our approach all activities presented in the BPEL Process and define the pre-conditions or effect, etc.

REFERENCES

- [1] P. Bartalos and M. Bielikova, « AUTOMATIC DYNAMIC WEB SERVICE COMPOSITION: A SURVEY AND PROBLEM FORMALIZATION », Computing and Informatics, Vol. 30, 2011, pp. 793–827.
- [2] M. Pistore, P. Traverso, P. Bertoli and A. Marconi, “Automated Synthesis of Composite BPEL4WS Web Services”, in IEEE Intl Conference on Web Services (ICWS’05). 2005
- [3] N. Vuković, “Context aware service composition”, University of Cambridge, Technical Report UCAM-CL-TR-700, October 2007
- [4] W. Han, X. Shi and R. Chen, “Process-context aware matchmaking for web service composition”, Journal of Network and Computer Applications, 2008, pp. 559–576
- [5] Z. Wu, A. Ranabahu, K. Gomadam, A. P. Sheth and J. A. Miller, « Automatic Composition of Semantic Web Services using Process and Data Mediation », Technical Report, LSDIS lab, University of Georgia, February 28, 2007
- [6] S. Russell and P. Norvig, “Artificial Intelligence: A Modern Approach”, Pearson - International edition, 2010
- [7] M. Pistore, P. Traverso, P. Bertoli and A. Marconi, « An Approach for the Automated Composition of BPEL Processes », 6789@ ABCDE FGHC6D• I, 2005, p. 24
- [8] D. H. Shin, K. H. Lee and T. Suda, “Automated generation of composite web services based on functional semantics”, Web Semantics: Science, Services and Agents on theWorldWideWeb, 2009, pp. 332–343
- [9] R. Akkiraju, A. Ivan, R. Goodwin, B. Srivastava and T. Syeda-Mahmood, “Semantic matching to achieve web service discovery and composition”, Proceedings of CEC/EEE’06, IEEE Computer Society, Washington, DC, 2006, p. 70.
- [10] Z. Liu, A. Ranganathan and A. Riabov, “Modeling web services using semantic graph transformations to aid automatic composition”, Proceedings of ICWS’07, IEEE Computer Society, Washington, DC, 2007, pp. 78–85.
- [11] D. Thakker, T. Osman and D. Al-Dabass, “Knowledge-intensive semanticweb services composition”, Proceedings of UKSIM’08, IEEE Computer Society, Washington, DC, 2008, pp. 673–678.
- [12] Eclipse BPEL Project. Eclipse BPEL Designer. <http://www.eclipse.org/bpel/>

- [13] A.Arkin, S. Askary, B. Bloch, F. Curbera, Y. Golland, N. Kartha, C. K. Liu, S. Thatte, P. Yendluri and A. Yiu, “Web Services Business Process Execution Language Version 2.0”, Proceedings of the 13th international conference on World Wide Web, Newyork, USA, 2004, pp. 621 – 630.
- [14] S. Elfirdoussi, Z. Jarir and M. Quafafou, « Discovery and Visual Interactive WS Engine based on popularity: Architecture and Implementation », International Journal of Software Engineering and Its Applications, 2014, Vol 8, No.2, pp 213-228.
- [15] M.Rajeswari, G.Sambasivam, N.Balaji, M.Saleem Basha, T.Vengattaraman, & P. Dhavachelvan, «Appraisal and analysis on various web service composition approaches based on QoS factors»,Journal of King Saud University-Computer and Information Sciences, 2014, vol. 26, no 1, pp 143-152.
- [16] T. RAJ, TF Michael, K. RAVICHANDRAN, K. RAJESH, «Domain Specific Web Service Composition by Parameter Classification Using Naïve Bayes Algorithm», World Applied Sciences Journal , 2014, vol 29, pp 99-105.
- [17] M. SAID,M. HAZMAN, H. HASSAN, et al. «GenericSOA: a Proposed Framework for Dynamic Service Composition», International Journal of Computer Science Issues (IJCSI), 2014, vol. 11, no 2, pp 94-99.

AUTHORS

Selwa ELFIRDOUSSI

Has obtained a diploma of Engineer in Software Engineering from ENSIAS School of engineering, Mohamed V Souissi University, Rabat, Morocco in 2000. Actually, she's a PhD student at Faculty of sciences, Cadi Ayyad University in Marrakech, Morocco since 2008. Her research interest is focalized on service-oriented computing and Web service technologies.



Zahi JARIR

Zahi JARIR Received his postgraduate degree in computer science in 1997 on Natural Language Processing at Faculty of Sciences in Rabat, Morocco. From 1997 to 2006, he was assistant professor at Faculty of sciences, Cadi Ayyad University in Marrakech, Morocco. In 2006, he received academic accreditation from Cadi Ayyad University. Currently, he is a professor of Computer Science at Faculty of Sciences of Cadi Ayyad University. His research interests at LISI laboratory lie mainly with the field of Service-oriented computing and technologies, Cloud computing and security, Computational reflection and Meta level architectures, Adaptive and Mobile Middleware, and Customization techniques of Web Services and Applications.



Mohamed QUAFAFOU

Mohamed QUAFAFOU did his PhD Thesis in 1992 on Intelligent Tutoring Systems at INSA de Lyon, France. From 1992 to 1994, he was ATER at INSA de Lyon and than at Nantes Faculty of Sciences. From 1995 to 2001, he was assistant professor at the Nantes University. During that period, he developed research on Rough Set Theory, concepts approximation, data mining, web information extraction and participated actively with France Telecom to design a new web system dedicated to French web analysis for discovering emergent web communities. He was also chief-scientist at GEOBS where he headed the Geobs Data Analyzer project, which was developing a spatial data mining systems with application to environment, marketing, social analysis, etc. From September 2002, he was professor at the Avignon University and moved in 2005 to the Aix-Marseille University where he joined the LSIS CNRS and leads research on Data Mining Theory and Applications focusing on new contexts for learning (crowdsourcing, interconnected data, and big data) with application to user's behavior analysis, web services, cloud automatic auto-scaling, social network analysis, etc.



INTENTIONAL BLANK

A FORMALIZED MODEL FOR SEMANTIC WEB SERVICE SELECTION BASED ON QoS PARAMETERS

Divya Sachan¹, Saurabh Kumar Dixit¹, and Sandeep Kumar¹
Department of Electronics and Computer Engineering, Indian Institute of
Technology Roorkee, Roorkee, India
{divyasachan22, saurabh4551, sandeepkumargarg}@gmail.com

ABSTRACT

Selecting the most relevant Web Service according to a client requirement is an onerous task, as there are innumerable functionally same Web Services (WS) that are able to satisfy the request of user. However non-functional attributes also matter a lot. A web Service Selection Process involves two major points: Recommending the pertinent Web Service and avoiding unjustifiable web service. The deficiency in keyword based searching is that it doesn't handle the client request accurately. UDDI and search engines are based on keyword search, which is insufficient for providing the relevant service. So the search mechanism must be incorporated with the Semantic behavior of Web Services. In order to strengthen this approach, the proposed model is incorporated with Quality of Services (QoS) based Ranking of web services.

This paper enlightens on various concepts of Quality of Service associated with web services. Various QoS parameters like performance, availability, reliability and stability etc. are formalized in order to enhance the pertinence of web service selection. A QoS mediator agent based Web Service Selection Model is proposed where QoS Consultant acts as a Mediator Agent between clients and service providers. Model suggests user's preferences on QoS parameter selection. The proposed model helps to select pertinent Web Service as per user's requirement and reduce the human intervention.

KEYWORDS

UDDI, QoS, SOAP, Web Service Selection (WSS), Ontology.

1. INTRODUCTION

Web Services [1] assists in providing solutions for distributed business processes and applications which are accessible via the Internet. In case a single WS doesn't meet the complex requirements, several web services combine together to provide a composite solution. In such cases selecting several Web Services for Web Service Composition becomes a major step in the overall process.

WS are nothing more than software ingredients that interact with one another by sending XML messages wrapped in SOAP envelops. WS communication is built on SOAP. SOAP is

XML based information packaging definition. It provides a structured way for information exchange between peers in a distributed environment.

Web Services are defined as “self-contained, self-describing, modular application that can be published, located and invoked across the Web” [1]. These web services are described by using standards like WSDL and then service descriptions are published in some UDDI [2] registries. Whenever a service request is invoked, a search is performed between service request description and available web service description which can satisfy the functional requirement of request.

As we know Service Oriented Architecture (SOA) is not only the service’s architecture as per technology basis but it also renders the policies, practices and frameworks to assure that pertinent services are provided and consumed by users. Goals of SOA are, firstly service provider offers several services and secondly prospective users of the services dynamically choose the best service from the set of services offered. In reference to current Web when we put a query in a search engine, we find a long list of WS as per the similar keywords. Now we have to ad-hoc decision to choose a WS. Now it is just a matter of chance that we select a relevant WS to perform our work on Web. So we can say that above mentioned goals of SOA, are partially executed as WS are described and listed in public registries but there is no means to choose the best from the set of services offering the same functionality. A consumer is thus forced to make an ad-hoc decision of choosing a service from multiple services offered for the same functionality.

In such scenario Quality of Services (QoS) assist in ranking the WS and selecting the best WS from a list of candidate web service having similar functionality based on their QoS descriptions, in response to a service request made by the user. The QoS information is used for categorizing web services in regards to a request of QoS demands [5]. Such QoS information comprises of performance (in terms of response time, latency etc.), accessibility, availability, throughput, security etc. which are expressed as a set of QoS properties. These QoS information have considerable impact on expectation of a user and the experience of using a Web Service. Hence it can be used as a main factor to distinguish and rank Web Services. The service which gets the highest QoS value is selected first. However it should be clear that this ranking step is performed only after the functional matching with the user’s request has been done.

The remaining paper is structured as follows: Section 2 deals with the related work in the field of modeling of QoS parameters. Section 3 gives an overview of the proposed model. Section 4 focused on modeling of different QoS parameters. Section 5 gives the implementation of QoS based WSS model. Section 6 gives an overview of the Simulation Environment and the Implementation aspect of the proposed model. It also provides Simulation results and evaluation. Section 7 deals with conclusions and future prospects of this work.

2. RELATED WORK

Web Service Delegation model [9] provides safety and privacy. This work shows how a Delegation Web Service increases the security for Web Services, but doesn’t consider other QoS based parameter for selection. Web Service discovery based on QoS [10] suggests QoS enhanced UDDI architecture and discusses different QoS parameters, but doesn’t provide methods to calculate them and computing all the QoS parameters for service selection approach may lead to miss the relevant Web Service with low QoS parameters. [11] Introduces a

model to calculate QoS parameters of different Web Services and advocates the use of the Web Service Broker as selector architecture.

Combining QoS-based Service Selection with Performance Prediction [12] selects Web Service based on performance prediction (availability, reliability, bandwidth, request time) using artificial neural network. Performance prediction model lags in other QoS parameters like security, correctness, failure masking etc. Performance criteria might be different with respect to functionality of Web Services and user's interest. Model of Pareto principle based QoS Web Service Selection [13] uses 80-20 rule to compute QoS rank of Web Services. Model reduces computation complexity of service selection as only 20 % of Web Services are ranked according to QoS parameter.

3-Way Satisfaction [14] for Web Service Selection Preliminary Investigation uses selection in community of similar Web Services. Master Web Service calculates SCORE of other slave Web Services based on capacity, execution time and availability. This approach solves the problem of selection Web Service within a community.

[19] Has proposed a novel modeling approach using associative classification. A CBA [19] algorithm is used to classify the candidate WS to different QoS levels. They classified the WS within each class, with respect to their distances from the user's demand for the QoS criteria. In nutshell approach uses the classification data mining algorithm to select the most eligible services respect to the user demand. Further approach uses semantic similarity between WS by semantic links it increases the accuracy of proposed modeling.

Benaboud and Maamri [20] presented a framework for WS discovery and selection based on intelligent mediator agents. In order to add dynamism to WSS model, they have applied OWL-S and domain ontology concepts. Agent based framework is implemented using JADE [21], which was implemented using JENA API. Modeling approach is based on matching and domain ontology of WS, it does not consider parameter based selection.

Guo and Le [22] proposed that discovery of WS should be based on the semantic match between WS providers and consumer query. It contributes by providing procedures to represent WS by the OWL-S profile and OWL based language for service description. A description of the design and implementation of a WS matchmaking mediator which acts on OWL-S ontology is made. It also uses an OWL reasoner to compare ontology based WS descriptions.

A SWSS model based on QoS attribute is presented in [23]. Framework is modeled by adding semantics of QoS attributes with web service profiles. It describes the design and implementation of a WS matchmaking agent. Agent uses an OWL-S based ontology and an OWL reasoner to compare ontology based service descriptions. [11,12] provide a sketch for framework implementation, but how to exactly formalize and retrieve QoS values from WS profiles, still requires a novel work.

Different models are suggested in the field of web service selection, but the proposed model in this paper additionally support security, reputation, availability, correctness and reliability for efficient service selection. In addition to QoS based WS selection, our approach takes user's preference of QoS for service selection. As Request is about journals and research work there is no need to calculate rank of WS using security, availability and performance. Similarly if user requests for online purchasing then cost, security are important to consider rather than

correctness. So in nutshell model selects most relevant service among the functionally similar Web Services, as per user's preference. User can specify any QoS parameter which should get the preference but in the absence of user's input, over all RankQoS based on weighted sum of specified QoS parameters is considered.

3. QUALITY OF SERVICE (QoS)

The Quality of Services Ranking describes the quality of web services. It is an important consideration when the consumer makes decision on service selection. Normally, the QoS attributes can be classified in two categories: dynamic and static - as described in [17]. Li et al. explain in [17] that dynamic attributes could be changed in the execution time, for example response time and throughput; static attributes are defined by service providers before service executions and are usually not updated during the execution. Table 3.1 presents some example attributes by this classification.

Attributes	QoS parameters
Dynamic	Availability, response time, throughput, reputation, stability etc.
Static	Scalability, capacity, accuracy, security, price, etc.

Table 3.1 behavior of QoS parameter

QoS based selection translates user's vision into business processes more efficiently, since a Web Service can be designed according to QoS metrics.

QoS allows for the evaluation of alternative strategies when adaptation becomes necessary. The unpredictable nature of the surrounding environment has an important impact on the strategies, methodologies, and structure of WPs. Thus, in order to complete a WP according to initial QoS requirements, it is necessary to expect to adapt and reschedule a WP in response to unexpected progress, delays or technical conditions [3].

- It allows for the selection and execution of WPs based on their QoS, to better fulfill customer expectations.
- This approach help to fulfill the service oriented architecture's goal. Now users are not forced to make Ad-hoc decisions to select pertinent service among the set of services which are functionally equivalent.
- It makes possible the monitoring of WPs based on QoS. WPs must be rigorously and constantly monitored throughout their life cycles to assure compliance both with initial QoS requirements and targeted objectives.

4. PROPOSED MODEL

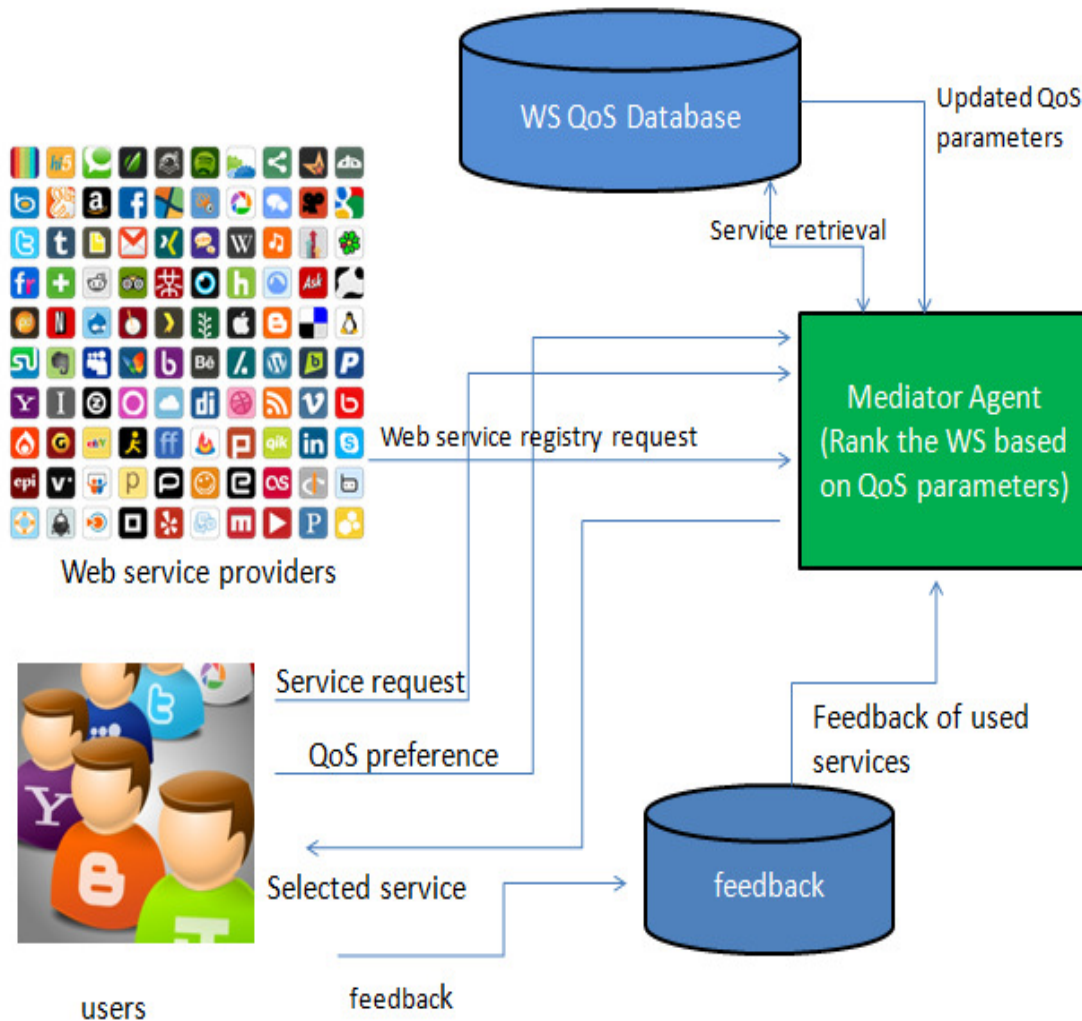


Fig. 1 QoS mediator agent based Web Service Selection Model.

A *QoS mediator agent based Web Service selection* and Web Service registry model (shown in Fig. 1) is proposed in this paper. Whenever a search is performed, the mediator agent selects the list of matched services from the service pool and provides to the client. Client set the preference to QoS parameter and highest ranked service will be provided to the client. Locating the desired Web Service to a client requirement is a difficult task as similar Web Services are readily available to satisfy a request. If there are many equivalent results returned by the QoS database, then the *service_selection* method is called that takes input as matched services on user request and then depending on the linearity of the constraints and the user's preferences, it executes the appropriate WS selection algorithm and returns back the results to user. Whenever a service registry is requested by service providers, a new Web Service is entered in database and respective Web Service is joined in given web-service set.

A. Algorithm for Service Registry:

Step 1: Web Services provided by different providers are stored in QoS database as their functional type and QoS values.

Step2: Normalization is performed using below formulae.
For negative parameters ("lower the better")

$$Q_i = \begin{cases} \frac{Q_j^{\max} - Q_i}{Q_j^{\max} - Q_j^{\min}}, & \text{if } Q_i^{\max} - Q_i^{\min} \neq 0 \\ 1, & \text{else} \end{cases} \quad [4.1]$$

For positive attributes ("higher the better")

$$Q_i = \begin{cases} \frac{Q_i - Q_i^{\min}}{Q_i^{\max} - Q_i^{\min}}, & \text{if } Q_i^{\max} - Q_i^{\min} \neq 0 \\ 1, & \text{else} \end{cases} \quad [4.2]$$

Step 3: Calculate Initial_Rank_{QoS}.

Step 4: For each Web Service if initial QoS parameters are satisfying minimal criteria, assign them AverageRank_{QoS}, to participate in Service Selection procedure.

$$\text{checkif}(\text{Initial Rank}_{QoS} \geq \text{ThresholdRank}_{QoS})$$

$$\text{then Rank}_{QoS} \leftarrow \text{AverageRank}_{QoS}$$

Step 5: According to performance and feedback, Rank_{QoS} will be keep updating, on every selection of Respective Web Service.

If a WS has published good range of QoS parameters during its registry with Mediator agent, we have assigned it *AverageRank_{QoS}*. As every new WS should get an equal chance to be selected in QoS based WSS system, further on the basis of user's feedback and other WS availability, its QoS Rating will be updated

B. Algorithm for Service Selection:

/ Web Services provided by different providers are stored in QoS database as functional type and Rank_{QoS} */*

Step 1: From each user's query the semantically matched WS are extracted from database.

Step 2: Depending on the query the QoS parameters provided by the user is retrieved.

Step 3: Rank_{QoS} is calculated. A listing of the QoS parameter based on weighing schemes and the user's preference is also made used of.

Step 4: If user's required QoS parameters are not specified then rank the Web Services on standard weighing schemes and calculate Rank_{QoS}.

Step 5: The best Rank_{QoS} Web Service is selected.

Where

$$Rank_{QoS} = \sum Q_i \times W_i \quad [4.3]$$

$$AverageRank_{QoS} = \frac{\sum Q_i \times W_i}{N} \quad [4.4]$$

Where N is number of Web Services of given type and

Q_i = {performance, availability, reliability, feedback, execution time, throughput, security, scalability} is normalized value of considered QoS parameters and W_i is weighted contribution of selected QoS parameter.

5. FORMALIZATION OF QoS PARAMETERS

Performance: It is a measure of how well a WS performs during its execution [5]. This is a prominent ingredient in overall having “higher the better” tendency. Performance is a composite attribute comprises of weighted sum of latency, throughput, and response time.

$$Q_{performance} = \sum_{i=1}^{i=3} W_i * Q_i \quad [5.1]$$

Where

- *Q_{performance}* : is a performance based rating of WS ranging in between 0 to 1.
- *Q_i = {Q_{latency}, Q_{throughput}, Q_{response time}}*
- *Latency and response time* have “lower the better” tendency for *QoS_{performance}* rating.
- *On the other hand Throughput* has “higher the better” contribution on *QoS_{performance}* ranking.

Availability: Availability of WS play a vital role in QoS rating with the behavior “higher the better”. It is measured as the probability of WS that the service will be up after selection. As the name shows it have complementary behaviour with unavailability. In this WSS modelling we have considered a WS with low is considered as unavailable. Low is assigned to a WS if WS is less than *thresholdRank_{QoS}*.

$$QoS_{availability} = 1 - QoS_{unavailability} [5.2]$$

Where

- $QoS_{unavailability}$: show QoS value of unavailability of a WS, varying between 0 and 1.
- $QoS_{unavailability}$ has “lower the better”

Experience: It is a dynamic parameter that increases as more number of times service gets selected in QoS based WSS system. Number of times service has been selected $QoS_{experience}$ value gets incremented by 1. It is also a positive parameter, shows “higher the better” contribution on QoS rating of a WS. With prior simulation we concluded that including *Experience* directly to QoS rating will biased the system towards earlier registered WS. So instead of directly adding $QoS_{experience}$, *Experience* is used to calculate other QoS parameters like *Reputation*, *throughput* etc.

Reputation: Reputation of WS shows satisfaction of users. It is collective built over the time as per user’s feedback. Feedback may be positive or negative.

$$Q_{Reputation} = \frac{N_{pos} - N_{neg}}{Experience} [5.3]$$

Where

- N_{pos} : Number of times service is ranked with positive feedback.
- $Q_{reputation}$: is a reputation based ranking of WS ranging -1 to +1.
- N_{neg} : Number of times service is ranked with positive feedback
- *Experience*: is a QoS parameter.

Incompleteness: Number of times service was not successfully completed. Incompleteness is considered as a number of counts respective WS was being selected but not completed its execution. It is a negative parameter as it has “lower the better” contribution in QoS ranking.

Reliability: Reliability is the ability of a WS to perform well over a given time span. Reliability[15] of a WS is measured in terms of *MTBF*(mean time between failure), *recoverability*, *performance* and *availability*.

$$Q_{reliability} = \sum_{i=1}^{i=4} Q_i * W_i [5.4]$$

Where

- $Q_i = \{Q_{MTBF}, Q_{recoverability}, Q_{performance}, Q_{availability}\}$
- Q_{MTBF} is the quality attribute of WS measured as mean of the time spans , when WS was in failure condition. It has “lower the better” contribution in QoS rating.
- W_i : is the weight associated with respective parameter.
- $Q_{recoverability}$ Is the ability of WS from failover and disaster [15].

Security: Security is the measure of how much it is secure to use a WS regarding different security threats [16]. Different security mechanisms used by Web Services are ranked Initially by mediator agent [16]. To calculate Risk of WS.

$$Q_{security} = F(\text{security protocol, encryption methods, auditability, Risk}) [5.6]$$

Where

- Risk should be minimum for good $Q_{security}$.
- Risk is measured in terms of inbound and outbounds attacks [16].

Scalability: scalability of WS is measured in the term of maximum number of simultaneous transactions on WS without decreasing its performance. It is taken as a core parameter at the time of service registry and used at the time of load balancing. It is a static parameter as it is computed by mediator agent at the time of WS registry.

$$\text{Scalability} = \{\max(N_s) \mid \text{Performance is constant}\} [5.7]$$

Where

- N_s : number of simultaneous transactions on WS
- Performance is QoS value of WS.

Throughput : is the vital contributor in QoS rating with “higher the better” behavior. [5] Throughput is defined as total number of completed transactions by a Web service over a time period.

$$Q_{throughput} = 1 - \frac{Q_{incomplete}}{Q_{experience}} [5.8]$$

Where

- $Q_{throughput}$: is a throughput based ranking of WS ranging between 0 to 1.
- $Q_{incomplete}$ and $Q_{experience}$ are the QoS parameter we have formalized earlier

Dependability: It refers to the service delivering capability of a service that can be trusted justifiably. It is calculated from the complete transaction of its sucesor and predecessor services at dynamic composition.

ExecutionTime: Initially published by Web Service provider. Execution Time of a service should be updated at the time of web service registry at QoS Agent side. and further updated as per user's feedback of execution time {exact, delayed}.

ResponseTime: Response time is the overall time required to complete a service request. It is a composite quality attribute comprises of latency and network delay with "lower the better" tendency towards QoS Rating of a WS. Equation 4.1.9 is used for computation of response time.

$$Q_{responsetime} = 100 * \frac{(e^{-(r^2)} * e^{(-r+\beta r)})}{(1 + e^{(-r+\beta r)})} [5.9]$$

Where

- r is the response time, which is measured by running the service.
- βr is the Service Level Agreement (SLA) value for response time.

The first component deals with the contribution of the present response time in the quality rating. The second component deals with the contribution of the overall difference from the published value. It is the contribution of the overall deviation of response time from the published SLA. Mei and Meeuwissen in [18] explained that the SLA is a concept to get QoS guarantees between service providers and consumers at the network level. We can regard the SLA as a measure standard, which can be used to evaluate the service quality.

Stability: As we know stability refers constancy of WS [5]. A good WS should have lesser variation in its QoS rating. Stability concerns whether service is dependable or not. As much variation in QoS attributes shows dynamic behavior of WS. A less stable WS cannot be predicted for its performance. It may be the scenario that after selection WS is not executing according to modeling prediction or its performance is not on a par. It leads to diminish the efficiency of SWSS model. In nutshell, Stability is a considerable attribute for QoS based SWSS modeling.

Stability is a positive attribute of a WS, it show dependability on Web Services. However Li-Li and Yan [25] has defined Stability as rate change of web services parameters. But they did not discuss how to measure it or retrieve from web service profile. Stability is inversely proportional to rate change of dynamic QoS attributes like performance, response time. It depends on deviated value of dynamic QoS attributes. Equation 5.10 is used for stability computation in this SWSS model

$$Q_{stability} = 1 - \sum_{k=0}^3 \Delta Q_i * W_i [5.10]$$

Where

- $Q_{stability}$ is QoS rating of WS which lies in between 0 to 1.
- ΔQ_i is deviated value of QoS parameters $\{Q_{performance}, Q_{responsetime}, Q_{reputation}\}$.
- W_i is weighted contribution of respective QoS parameters in order to calculate QoS value of stability.

While computation of Δ is done on the basis of its previous QoS value and current value of respective QoS attribute. Previous QoS value is stored in web service profile's ontology as the mean of previous value.

$$\Delta Q_i = |\text{mean } Q_i - \text{new } Q_i|$$

Mean value of QoS attributes is keep updating after every selection, using equation.

Where

$$\text{Mean QoS}_{\text{new}} = \frac{(\text{Mean QoS}_{\text{previous}} * \text{Experience} + \text{QoS}_{\text{new}})}{\text{Experience} + 1} \quad [5.11]$$

- $\text{Mean QoS}_{\text{previous}}$: denotes previously stored mean value of a particular QoS parameter.
- $\text{Mean QoS}_{\text{new}}$: Refers to the updated mean value of a QoS parameter.
- Experience : is an attribute associated with WS profile which refers to the number of times WS gets selected.

Up to this level we have formalized different QoS attributes regarding a WS. Every QoS attribute has its own range of rating.

So to calculate cumulative QoS Rank, we need to normalize them in a common range. In our system we normalized different QoS attributes in the range of 0 to 1 using equation.

Now all the attributes are of same range so we can directly use them for their weighted contribution on QoS rating of WS.

$$\text{Rank}_{QoS} = \sum Q_i \times W_i$$

$$\text{AverageRank}_{QoS} = \frac{\sum Q_i \times W_i}{N}$$

Where

- N is number of Web Services of given type.
- $Q_i = \{\text{performance, availability, reliability, feedback, execution time, throughput, security, dependability, scalability}\}$ is normalized value of considered QoS parameters and W_i is weighted contribution of selected QoS parameter.

5. IMPLEMENTATION OF QoS BASED SEMANTIC WSS MODEL

5.1. Modeling Architecture for QoS based Semantic WSS

In current web some core Quality of service (QoS) parameters of a web service are registered in the UDDI entry. These QoS information can be retrieved by the consumers or the brokering systems. So there is a requirement of database to store all the Quality of service information regarding web services (as we have used for Net-logo simulation of SWSS). Maintaining such a large database for QoS for innumerable web service is an onerous task. Li and Zhou elaborated in [17] that such mechanisms based on the UDDI registry, are less efficient due to the fact that their selection results always contain irrelevant and unjustified values. There is no mean to associate QoS parameters with service profiles. This problem is caused by lack of semantic support. To address this problem, QoS based semantic WSS approach is proposed where services are built as OWL-S profile and QoS parameters ontology is attached with profile itself. Figure 5.1 shows outline of model implementation. OWL-S, QoS ontology and service profile generation are the important steps towards SWSS implementation, which are discussed later in this chapter.

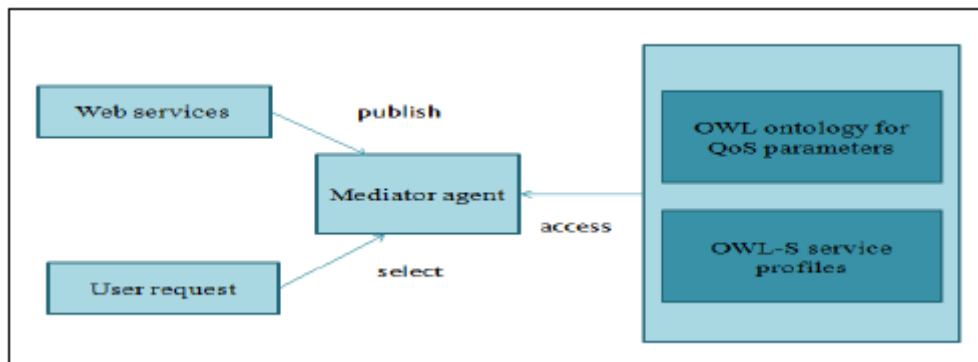


Figure 5.1 Modeling architecture for QoS based semantic WSS.

In order to implement QoS based SWSS model, ontology of QoS based Semantic WSS modeling (figure 5.1) and ontology of QoS parameters (as shown in figure 5.2) was developed using protégé [26]. Figure 5.2 shows a generic ontology of the QoS parameters that has been used in the proposed model design.

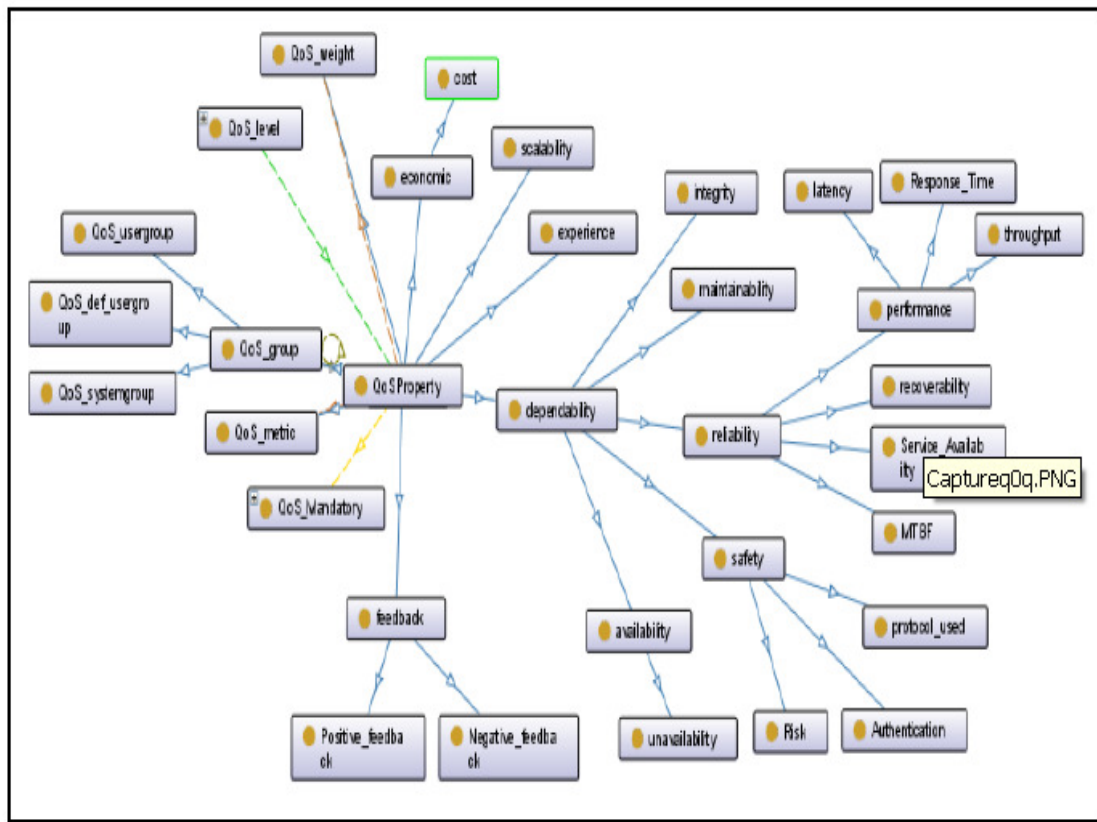


Figure 5.2 Generic Ontology of the QoS Parameters

QoS ontology developed for proposed semantic WSS modeling depicted in Figs- 5.2. Ontology shown in figure 5.2 relates to various roles in defining QoS information like QoS description, QoS mandatory, QoS level, QoS weight, QoS formula, QoS priority, and QoS group. While ontology shown in fig 5.2 shows a set of core QoS properties.

5.2. Integration Of QoS Modeling With Semantic Service Profile

The QoS model architecture shown above is integrated with the Semantic Service Profile. Figure 5.3 shows the results which were obtained after incorporating the service profiles with our model design. The programming was done on Eclipse [27] platform with Java being the programming language. Jena library [28] was used for interaction with the service profiles which were prepared using Protégé 3.2.1. The formalizations which were made earlier in the report were incorporated in the model prepared here. The result clearly validates the proposed model design.

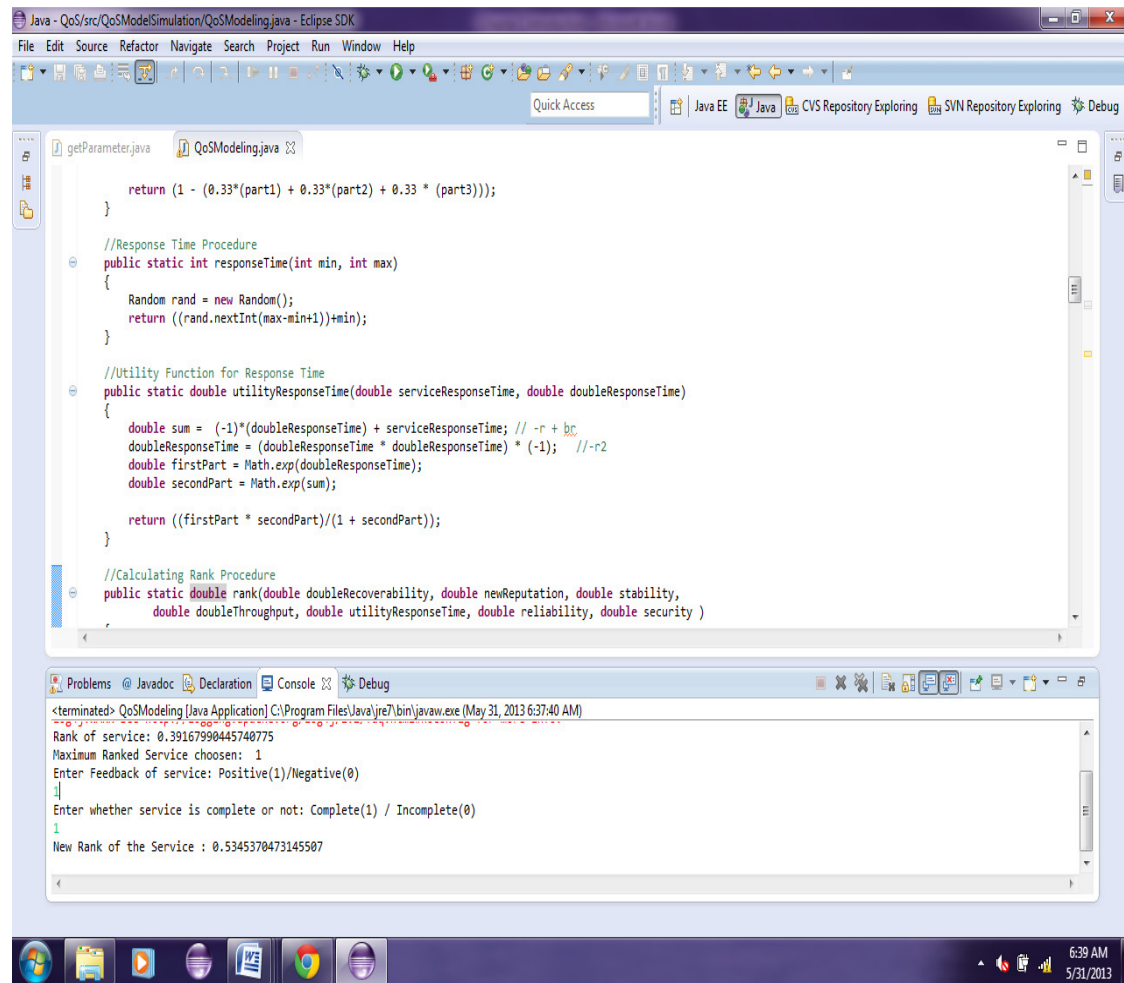


Figure 5.3 Screen shot of programming interface

The code prepared is modular with separate functionalities assigned to separate modules. `getRank()` is one such function which gets the previous rank of the service from the service profiles and is used to find the best available service in the market. Similarly many other functions like `rank()`, `utilityResponseTime()`, etc. has been made which work together to assign new rank to the service depending on the feedback provided by the user.

6. SIMULATION ENVIRONMENT, IMPLEMENTATION, RESULTS AND COMPARATIVE ANALYSIS

The system is developed using NETLOGO [7] and MYSQL [8] database. Initially service given by providers is assumed in text format of particular functional types, then *core* QoS parameters published by service provider are normalized and store in database. This QoS database is being processed by NETLOGO MYSQL extension.

Every service registry request will generate a new Web Service in system linked with *Service Agent 0*. Respective entry for Web Service is done in database and initial rank is calculated

and $AverageRank_{QoS}$ is assigned. For further selection, service's updated Rank is assigned to $Rank_{QoS}$.

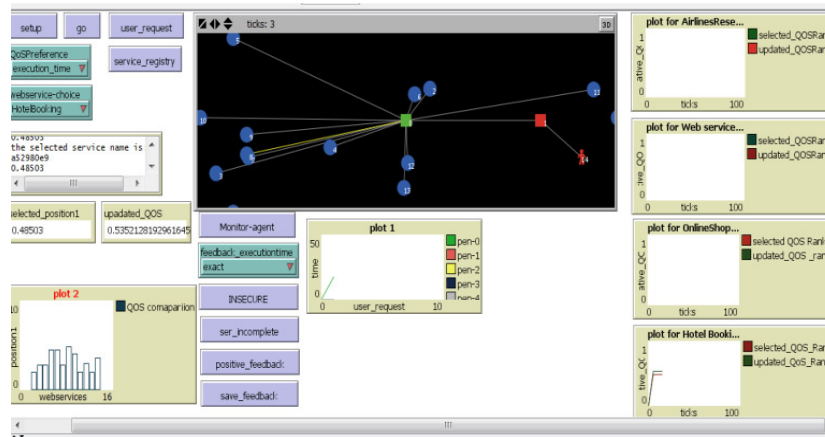


Figure 6.1 Best Service Selection of type “hotel booking” on User’s request.

In order to evaluate the efficiency of proposed model, we have drawn a comparison graph between $selected_QoS_Rank$ of Web Service and $updated_QoS_Rank$ of Web Services. As in the above graphs $selected_QoS_Rank$ and $updated_QoS_Rank$, both are varying in similar fashion. Where $Selected_QoS_Rank$ is QoS value of selected Web Service while $updated_QoS_Rank$ is calculated QoS on user’s feedback. So we can conclude that selection of service was efficient, as we are selecting best ranked QoS service, so its updated value proportionally affects overall QoS_Rank for a particular type of service.

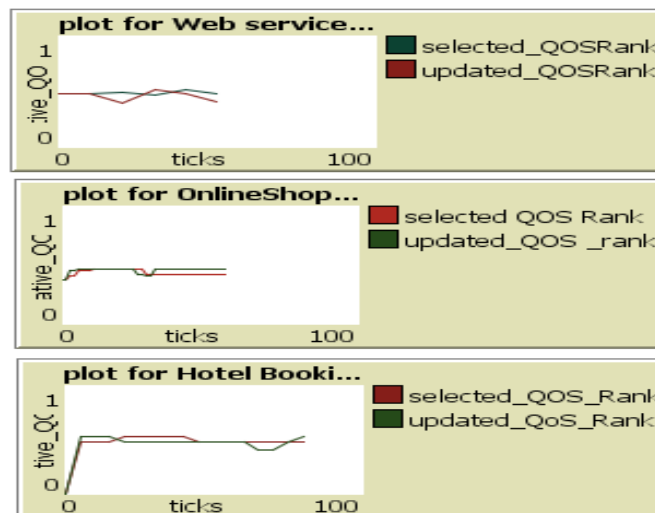


Fig. 6.2 Plot for $selected_QoS_Rank$ and $updated_QoS_Rank$ for different services.

To quantify the mediator agent based Web Service selection model, we have made a graphical comparison between $Rank_{QoS}$ and successful number of services (*experience* –

incomplete number of services) provided by Web Services at a given time period. As both are varying in similar way, means highest QoS rank Web Service always have high performance as per user's request. High valued RankQoS service should be selected maximum time. As experience and RankQoS are following the same graph pattern, so we can say that proposed model is selecting the best service among the same functionality Web Services. As in the above graph experience is not NULL for low valued RankQoS, this infers that every service is getting chance to select.

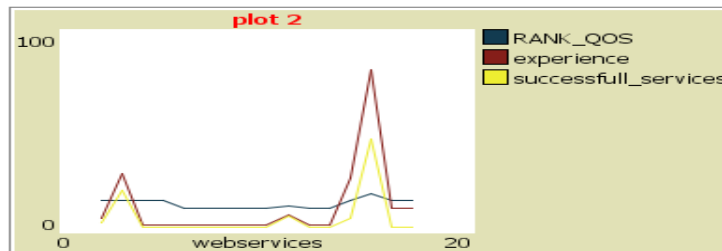


Fig. 6.3 Service Selection comparison for a type of Web Service

After comparison with the Selection model using Pareto Principle [13] we found that this Pareto model do not rank the web services according to user's requirements and feedback. On the other hand the model proposed in this paper takes user's requirements and satisfaction into consideration for ranking the web services. 3-Way Satisfaction for Web Service Selection [14] gives selection method with in a community and uses *Score (availability, correctness, execution time)*.

A comparative analysis based on QoS parameters is done in order to justify the usability of proposed modeling. Selecting most pertinent Web Service and not missing the appropriate Web Services are the two complimentary issues in Web Service selection process. Selecting Web Services according to user's requirement will accomplish both the issues. Feedback is also collected for further selection of same Web Service and to rank web services as per user's requirement.

Table 1 show that Web service ranking with the proposed model uses additional parameters like security, reliability, throughput, reputation etc. for efficient and effective selection as shown in table.

Table 1 Comparison of SWSS modeling based on QoS parameters.

QoS parameter used in	AHP [15]	FLM M [3]	Pareto Prin. SM	3-way Satisfacti on Model	Req. based Broker Arch. [5]	WSS based on Naïve Bayes	Proposed QoS based WSS
Availability	Yes	Yes	No	Yes	Yes	No	Yes
Reliability	Yes	Yes	No	No	Yes	No	Yes
Depend-ability	No	Yes	No	No	No	No	Yes
Reputation	Yes	Yes	Yes	Yes	Yes	No	Yes

Feedback	No	No	No	No	Yes	No	Yes
Failure Semantic	No	No	No	No	No	No	No
Flexibility Scalability	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Security	No	No	No	No	No	No	Yes
Response Time	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Throughput	Yes	Yes	No	No	Yes	Yes	Yes
Integrity	Yes	Yes	No	No	Yes	No	No
Stability	Yes	Yes	No	No	Yes	No	Yes

7. CONCLUSIONS AND FUTURE WORK

In order to enhance efficiency of Web Service selection based on QoS Ranking, this model presents mediator Agent based approach for QoS based Web Service selection. It presents a uniform and lightweight solution for mining QoS parameters. The model selects most relevant service among the functionally similar Web Services, as per user's preference. User can specify any QoS parameter which should get the preference but in the absence of user's input, over all RankQoS based on weighted sum of specified QoS parameters is considered.

A modeling approach is proposed for QoS based semantic WSS model (SWSS) and formalization for various QoS attributes (like reliability, stability, availability, incompleteness etc.) are presented. Mediator agent based semantic web service model is proposed in the modeling approach. Further service registry and service selection algorithms are listed for pertinent WSS based on QoS parameters. Later on different QoS attribute associated with web services are discussed with proper formalization.

The proposed model is simulated on Net-Logo and the results were used to check the efficiency of the proposed model. A generic ontology for the QoS parameters was prepared and the model was integrated with the semantic service profile. In order to read the ontology of the website Jena library is used. The results obtained show the validity of our proposed model. A comparative study of other research work on QoS parameters is also made to emphasize upon the large number of QoS parameters taken into consideration in the proposed model.

In our future work, we would like to explore more QoS parameters and continue our research in the field of service composition in semantic environment. Parameters like integrity, compliance, etc. can be modeled and formulated accordingly to improve the existing service selection models available in this research area. Further QoS based SWSS model can be collaborated with context attribute based parameters to increase the effectiveness and efficiency of the models.

REFERENCES

- [1] W. Junhao, G. Jianan , J. Zhuo. Semantic Web Service Selection based on QoS . International Joint Conference on Service Sciences, 2011: 163-169
- [2] UDDI .<http://www.uddi.org/pubs/uddi v3.html>.
- [3] G. Guo, Fei Yu and Dong Xie. A Four level Model For Web Service Selection Based on QoS Ontology .Third International Symposium on Information Science and Engineering, 2010: 630-634.
- [4] S.Chitra , K. Vidhya and G. Aghila .A Web Service Selection based on Ranking of QoS using Naïve Bayes .ICCCT,2010: 782- 789
- [5] K. Kritikos, D. Plexousakis. Requirements for QoS-Based Web Service Description and Discovery, IEEE Transactions on Service Computing ,2009: 320-337.
- [6] GUO De-keREN, YanCHEN Hong-huiXUE, Qun-weiLUO, Xue-shan. A Web Services Selection and Ranking Model with QoS Constraints. journal of Shanghai Jiaotong University,2007, 41(6):870-875.
- [7] NETLOGO. <http://www.netlogo.org/pubs/netlogo v5.html>.
- [8] J. Blom, R. Quakkelaar M.Rotteveel “NetLogo SQL Wrapper User manual”ccl.northwestern.edu/netlogo
- [9] H. S. Hwang, H. J. Ko, K. I. Kim, U. M. Kim. Agent-Based Delegation Model for the Secure Web Service in Ubiquitous Computing Environments. In Proceedings of the 2006 International Conference on Hybrid Information Technology. Volume 1,pp.51–57. Nov. 2006.
- [10] S. Ran. A Model for Web Services DiscoveryWith QoS. ACM SIGecom Exchanges 4(1):1–10, 2003.
- [11] D. A. DMello, V. S. Ananthanarayana. Quality Driven Web Service Selection and Ranking. In Proceedings of the Fifth International Conference on Information Technology:New Generations. Volume 5, pp. 1175–1176. Chicago, Apr. 2008.
- [12] Zhengdong Gao, Gengfeng Wu . combining QoS based service selection with performance prediction. In proceeding IEEE International Conference on e-Business Engineering 2005.
- [13] Lican Huang. Pareto Principle to Improve Efficiency for Selection of QoSWeb Services, 7th IEEE Consumer Communications and Networking Conference (CCNC), 2010 :1-2.
- [14] Erbin Lim, Philippe Thiran , Zakaria Maamar, Jamal Bentahar. 3-Way Satisfaction For WebService Selection Preliminary Investigation, IEEE International Conference on Services Computing (SCC), 2011: 731-732.
- [15] V. X. Tran, H. Tsuji and R. Masuda, A new QoS ontology and its QoS-based ranking algorithm for web services. ELSEVIER, Simulation modellingpractice and theory (17) 2009: 1378-1398.
- [16] V. Prasath, Modeling the Evaluation Criteria for Security Patterns in Web Service Discovery, International Journal of Computer Applications 2010: 0975 – 8887.
- [17] S. Li and J. Zhou, "The WSMO-QoS Semantic Web Service Discovery framework", International Conference on Computational Intelligence and Software Engineering, 2009: 1-5.
- [18] D.A. Menasce and V. Dubey, “Utility-based QoS Brokering in Service Oriented Architectures”, In Proceedings of IEEE International Conference on Web Service (ICWS 2007), pp. 422 - 430, Salt Lake City, UT, 2007.
- [19] R.D. van der Mei and H.B. Meeuwissen “Modelling End-to-end Quality-of-Service for Transaction-Based Services in Multi- Domain Environments”, In pro ceedings of IEEEInternational Conference on Web Services, pp.3 – 462, Washington, DC, USA, 2006
- [20] M. Makhluhian, S. M. Hashemi, Y. Rastegari and E. Pejman, Web Service Selection Based On Ranking Of Qos Using Associative Classification, International Journal On Web Service Computing (IJWSC), Vol.3, No.1, March 2012: 1-14.
- [21] R. Benaboud, R. Maamri, and Z. Sahnoun, Semantic Web Service Discovery Based on Agents and Ontologies, International Journal of Innovation, Management and Technology, Vol. 3, No. 4, August 2012: 467-472.
- [22] F. Bellifemine, G. Caire, T. Trucco, and G. Rimassa, (2003). Jade Programmer’s Guide. [Online]. Available:<http://sharon.cse.it/projects/jade/>.

- [23] R. Guo^{1,2}, J. Le and X. Ling Capability Matching of Web Services Based on OWL-S Xia³ Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05) IEEE 2005.
- [24] S. CHAARI, Y. BADR and F. BIENNIER, Enhancing Web Service Selection by QoS-Based Ontology and WS-Policy 23rd Annual ACM Symposium on Applied Computing SAC'08, March 16-20, 2008: 2426-2432.
- [25] QU Li-li and C. Yan, "QoS Ontology Based Efficient Web Services Selection", International Conference on Management Science & Engineering (16th), pp. 45-50, 2009.
- [26] Protégé, Ontology Editor, 2007.<http://protege.stanford.edu/>. cited on 3rd june 2013. [27] Eclipse", <http://www.eclipse.org/>, last visited on June 2013.
- [28] "Jena Library",<http://jena.apache.org/>, last visited on June 2013.

AUTHOR INDEX

Ahmed Cuneyd Tantug 21

Alok Mittal 135

Alphonsa Kuriakose 41

Anu Gupta 135

Anuradha B 147

Arjun B Krishnan 75

Banu Diri 21

Chiranjib Patra 31

Deepjoy Das 187

Deepjoy Das 87

Divya Sachan 265

Eray Yildiz 21

Farah Harrathi 197

Hanane Houmani 229

Hemlata Dakhore 175

Jaouhar Fattahi 229

Jayaram Kollipara 75

Kanchan M. Taiwade 49

Krishna Naik 87

Mahran Farhat 197

Manoj Franklin 01

Manusha S 111

Matthew Lentz 01

Mohamed Mejri 229

Mohamed QUAFAFOU 251

Mohsen Gammoudi M 197

Nilesh Sharma 215

Nishant Sharma 215

Pawel Rosciszewski 09

Prakash S. Mohod 49

Prasanna SRM 87

Prateek Asthana 121

Pratyush 31

Pushpalatha N 147

Ramakrishnan M 157

Rani K 67

Rihab Ayed 197

Rituparna Devi 87

Roja Reddy B 59

Sandeep Kumar 265

Sangeeta Mangesh 121

Sarat Saharia 187

Saumya Vij 135

Saurabh Kumar Dixit 265

Selwa Elfirdoussi 251

Shubhangi N. Burde 175

Shweta Chahar 215

Sirisha K 111

Smriti Maheshwari 215

Subhankar Ghosh 87

Sukanyathara J 41

Suman V Patgar 67

Sushanta Pradhan 97

Thottempudi Pardhu 111

Uttarakumari M 59

Vasudev T 67

Vishakha Tiwari 215

Zahi Jarir 251