

David C. Wyld
Jan Zizka (Eds)

Computer Science & Information Technology

Third International Conference on Advanced Information Technologies &
Applications (ICAITA-2014)
Dubai, UAE, November 07 ~ 08 - 2014



AIRCC

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Jan Zizka,
Mendel University in Brno, Czech Republic
E-mail: zizka.jan@gmail.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-17-5
DOI : 10.5121/csit.2014.41101 - 10.5121/csit.2014.41132

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

Third International Conference on Soft Computing, Artificial Intelligence (SAI-2014) was held in Dubai, UAE, during November 07 ~ 08, 2014. Third International Conference of Data Mining & Knowledge Management Process (CDKP-2014), Third International Conference on Advanced Information Technologies & Applications (ICAITA-2014), Sixth International Conference on Networks & Communications (NeCoM-2014), Third International Conference on Software Engineering and Applications (SEAS-2014), Third International Conference on Control, Modeling, Computing and Applications (CMCA-2014), Fifth International Conference on Ad Hoc, Sensor & Ubiquitous Computing (ASUC-2014) and International Conference on Signal and Image Processing (Signal 2014) were collocated with the SAI-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 SAI-2014, CDKP-2014, ICAITA-2014, NeCoM-2014, SEAS-2014, CMCA-2014, ASUC-2014, Signal 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, SAI-2014, CDKP-2014, ICAITA-2014, NeCoM-2014, SEAS-2014, CMCA-2014, ASUC-2014, Signal 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 SAI-2014, CDKP-2014, ICAITA-2014, NeCoM-2014, SEAS-2014, CMCA-2014, ASUC-2014, Signal 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.

David C. Wyld
Jan Zizka

Organization

General Chair

Sundarapandian Vaidyanathan

Vel Tech University, India

Program Committee Members

Ashraf Shahin	Cairo University, Egypt
Ab El-Aziz Ahmed Abd El-Aziz	Cairo University, Egypt
Abdelkrim AID	University of Mascara , Algeria
Abdolreza Hatamlou	Islamic Azad University, Iran
Abdul Mateen Ansari	King Khalid University,Saudi Arabia
Abdurrahman Celebi	Bedër University, Albania
Adnan M. Al Bar	King Abdul Aziz University, Saudi Arabia
Aftab Alam,King	Khalid University ,Saudi Arabia
Ahmed Arara	University of Tripoli, Libya
Ali Abid D. Al-Zuky	Mustansiriyah University, Iraq
Ali AL-zuky	Mustansiriyah-University, Iraq
Ali Azimi	Ferdowsi University of Mashhad, Iran
Ali Jaoua	Qatar University, Qatar
Allaoua Refoufi	University of Setif, Setif Algeria
Alruily	M,Aljouf University, Saudi Arabia
AmirReza	Islamic Azad University, Iran
Arifa Ferdousi	Varendra University, Bangladesh
Asadollah Shahbahrami	Guilan University, Iran
Ashraf Shahin	Cairo University, Egypt
Assem Mousa	E commerce Tech Support Sys Manager, Egypt
Ayad Ismael	Erbil Polytechnic University, Iraq
Baghdad Atmani	University of Oran, Algeria
BenZidane Moh	University of Constantine2, Algeria
Dac-Nhuong Le	Vietnam National University, Vietnam
Denivaldo Lopes	Federal University of Maranhao, Brazil
Dias N.G.J	University of Kelaniya, Sri Lanka
Ehsan Abbasi	Azad Islamic, Iran
Elmahdi Abousetta	University of Tripoli, Libya
Emilio UR	University of La Rioja, Spain
Farah Harrathi	University of Manouba , Tunis.
Farhad Soleimanian	Hacettepe University, Turkey
Farhan Khan	University of Kent, United Kingdom
Farshchi S. M. R	Tehran University, Iran
Farshchi S.M.R	Seyyed Mohammd Reza Farshchi, Iran
Fatih Korkmaz	Cankiri Karatekin University, Turkey
Gammoudi Mohamed Mohen	Universite de la Manouba,Tunis
Georgeguo	Ningbo University of Technology, China
Girija. Chetty	University of Canberra, Australia

H. AZZOUNE	LRIA USTHB,Algeria.
H.Tebbi,LRIA	USTHB, Algeria.
Hacene BelhadeF	University of Constantine 2, Algeria
Hamadouche Maamar	USD Blida, Algeria
Hamdi hassen	Sfax University, Tunisia
Hamid Azzoune	LRIA USTHB Algiers, Algeria
Hassen Hamdi	University of Sfax, Tunisia .
Hayati MAMUR	Cankiri Karatekin University, Turkey
Hossein Jadidoleslami	MUT University, Iran
Hyung-Woo Lee	Hanshin University, Korea
Isa Maleki	Islamic Azad University, Iran
Islam Atef	Alexandria University,Egypt.
Jawad Talaq	University of Bahrain, Bahrain
Karam Jalal	Nazarbayev University, Kazakhstan
Kenneth Mapoka	Botswana College of Agriculture, Botswana
Khaled Merit	Mascara University, Algeria
Kwan Hee Han	Gyeongsang National University, Korea
Liyakath Unisa	Prince Sultan University, Saudi Arabia
Luisa Aquino	University of Saint Louis,Philippines
M S Kaiser	Jahangirnagar University, Bangladesh
M. R. Molaei	University of Kerman, Iran
M.Hamadouche	USDB, Algeria
Marjan Naderinejad	Tehran University of Medical Sciences,Iran
Martha Rosa Cordero Lopez	School of Computing, Mexico
Maryam K	University of Tabriz, Iran
Masoud ziabari	Mehr Aeen University, Iran
Mohamed el boukhari	University Mohamed First, Morocco
Mohamed Hashem	Ain Shams University, Egypt
Mohammad Masdari	Islamic Azad University, Iran
Mohammad S Khan	Sullivan University, USA
Mohammed Al-kahtani	Salman Bin Abdulaziz University, KSA
Mohammed H	Alexandria University, Egypt
Moslem Sharif Zadeh Javidi	Shiraz University, Iran
Muhammad Naufal Bin Mansor	University Malaysia Perlis, Malaysia
Muhammad. Ahmed	University of Canberra, Australia
Nabila Labraoui	University of Tlemcen, Algeria
Natarajan Meghanathan	Jackson State University, United States
Neda Darvish	Islamic Azad University, Iran
Nguyen Dinh Thuc	University of Science, VNU-HCMC, Vietnam
Nourddine Bouhmala	Buskerud and Vesfold University, Norway
Orteza Saberi	University of Tafresh, Iran
Peiman Mohammadi	Islamic Azad University, Iran
Prakash Kuppuswamy	Jazan University, K.S.A.
Raed Ibraheem Hamed Al-Falahy	University of Anbar Ramadi, Iraq
Reza Ebrahimi Atani	University of Guilan, Iran
Sarada Prasad Dakua	Qatar Science & Technology Park,Qatar
Sary Awad	Ecole des Mines de Nantes, France
Sayed EL-Rabaie	FacFaculty of Electronic Eng, Menouf
Sayed Ziaeddin Alborzi	Nanyang Technological University, Singapore

Seyyed AmirReza Abedini
Seyyed Reza Khaze
Shamim H Ripon
Shohidul Islam
Spits Warnars Harco Leslie Hendric
Suleyman Kondakci
Sumit Chaudhary
Tad Gonsalves
Tebbi Hanane
UERTI Mhania
Umesh Lilhore
Vuda Sreenivasa Rao
William Simpson
Yanet Pena Vazquez
Yasuko Kawahata
Yuhanis binti Yusof
Zhangxt

Islamic Azad University, Iran
Islamic Azad University, Iran
East West University, Bangladesh
University Of Ulsan, South Korea
Surya university, Indonesia
Izmir University of Economics, Turkey
Uttaranchal University, UK
Sophia University, Japan
LRIA USTHB Algiers, Algeria
National Polytechnic School Algiers, Algeria
NIIST Eng College Bhopal, Bhopal
Bahir Dar University, Ethiopia
Institute for Defense Analyses, USA
Universidad de las Ciencias Informaticas, Cuba
Kyungnam University, Japan
Universiti Utara Malaysia, Malaysia
Peking University, China

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

Third International Conference on Soft Computing, Artificial Intelligence (SAI-2014)

An Immune Agents System for Network Intrusions Detection.....	01 - 11
<i>Noria Benyettou, Abdelkader Benyettou and Vincent Rodin</i>	
An Investigation on Switching Behaviours of Vector Controlled Induction Motors.....	13 - 19
<i>Yilmaz Korkmaz, Ismail Topaloglu, Hayati Mamur and Fatih Korkmaz</i>	
Performance Analysis of the Recent Role of OMSA Approaches in Online Social Networks.....	21 - 32
<i>J. Ashok Kumar, S. Abirami and S. Murugappan</i>	
Optimal Buffer Allocation in Tandem Closed Queuing Network with Multi Servers Using PSO.....	33 - 41
<i>K.L.Narasimhamu, V.Venugopal Reddy and C.S.P.Rao</i>	
An Improved Teaching-Learning Based Optimization Approach for Fuzzy Clustering.....	43 - 50
<i>Parastou Shamsamandi E. and Soheil Sadi-nezhad</i>	
A Modified Invasive Weed Optimization Algorithm for MultiObjective Flexible Job Shop Scheduling Problems.....	51 - 60
<i>Souad Mekni and Besma Chaar Fayech</i>	
Iranian Cashes Recognition Using Mobile.....	61 - 71
<i>Ismail Nojavani, Amir Hassan Monadjemi and Azade Rezaeezade</i>	
Comparison of Filtering and Clustering Techniques in Diagnosis of Infants Retinopathy Risk.....	355 - 362
<i>Niousha Hormozi, Seyed Amirhassan Monadjemi and Gholamali Naderian</i>	

Third International Conference of Data Mining & Knowledge Management Process (CDKP-2014)

Qubit Data Structures for Analyzing Computing Systems.....	73 - 81
<i>Vladimir Hahanov, Wajeb Gharibi, Svetlana Chumachenko and Eugenia Litvinova</i>	

Arabic Tweets Categorization Based on Rough Set Theory..... 83 - 96
Mohammed Bekkali and Abdelmonaime Lachkar

Variable Length Key Based Visual Cryptography Scheme for Color Image..... 97 - 104
Akhil Anjekar, Prashant Dahiwale and Suchita Tarare

A Boolean Modeling for Improving the Algorithm Apriori..... 105 - 114
Abdelhak Mansoul and Baghdad Atmani

Combining Decision Trees and K-NN for Case-Based Planning..... 115 - 122
Sofia Benbelkacem, Baghdad Atmani and Mohamed Benamina

Third International Conference on Advanced Information Technologies & Applications (ICAITA-2014)

Competency Model for Information Systems' Specialization Track Utilizing RIASEC and Values Search Models..... 123 - 134
Risty Moyo-Acerado, Lorena W. Rabago and Bartolome T. Tanguilig

Effects of GOP on Multiview Video Coding Over Error Prone Channels..... 135 - 150
A.B Ibrahim and A.H Sadka

Molecular Dynamics Simulation Model of AFM-Based NanoMachining..... 151 - 168
Rapeepan Promyoo, Hazim El-Mounayri and Kody Varahramyan

Factors Influencing Actual Use of Mobile Learning Connected with E-Learning..... 169 - 176
Young Ju Joo, Sunyoung Joung, Eui Kyoung Shin, Eugene Lim and Miran Choi

Sixth International Conference on Networks & Communications (NeCoM-2014)

Distance's Quantification Algorithm in AODV Protocol..... 177 - 187
Meryem Saadoune, Abdelmajid Hajami and Hakim Allali

Dual Band Semi Circular Disk Patch Antenna Loaded with L-Shaped Slot..... 189 - 195
Amel Boufrioua

Low Altitude Airships for Seamless Mobile Communication in Air Travel..... 197 - 206
Madhu D, Santhoshkumar M K, Swarnalatha Srinivas and Narendra Kumar G

On the Modeling of Open Flowbased SDNS: The Single Node Case..... 207 - 217
Kashif Mahmood, Ameen Chilwan, Olav N. Østerbø and Michael Jarschel

SOC Nanobased Integrated Wireless Sensor System 333- 342
Penghua Sun, Maher Rizkalla, and Mohamed El-Sharkawy

Third International Conference on Software Engineering and Applications (SEAS-2014)

Enhancing an ATL Transformation with Traceability..... 219 - 230
Laura Felice, Marcela Ridao, Maria Carmen Leonardi and Maria Virginia Mauco

Solution of Unsteady Rolling Motion of Spheres Equation in Inclined Tubes Filled with Incompressible Newtonian Fluids by Differential Transformation Method..... 231 - 244
Y. Rostamiyan, S.D.Farahani and M.R.Davoodabadi

PDD Crawler : A Focused Web Crawler Using Link and Content Analysis for Relevance Prediction..... 245 - 253
Prashant Dahiwale, M M Raghuvanshi and Latesh Malik

Policy Overlap Analysis to Avoid Policy Conflict in Policy-Based Management Systems..... 255 - 266
Abdehamid Abdelhadi Mansor, Wan Mohd Nasir Wan Kadir and Ahmed Mohammed Elsawi

Third International Conference on Control, Modeling, Computing and Applications (CMCA-2014)

Control of Linear Systems Using Dynamic Output Controllers..... 267 - 278
Anna Filasova and Dusan Krokavec

On Observer Design Methods for a Class of Takagi-Sugeno Fuzzy Systems..... 279 - 290
Dusan Krokavec and Anna Filasova

Cooperating Adaptive Devices Applied in General Game Playing..... 291 - 303
Jose Maria Novaes dos Santos and Joao Jose Neto

**Analysis of Spectrum Sensing Techniques for Detection of DVBT Signals
in Gaussian and Fading Channels.....** 343 - 353
Ireyuwa Igbiosa, Olutayo Oyerinde and Stanley Mneney

**Fifth International Conference on Ad Hoc, Sensor & Ubiquitous
Computing (ASUC-2014)**

Performance Evaluation of VANETS Routing Protocols..... 305 - 315
Abduladhim Ashtaiwi, Abdusadik Saoud and Ibrahim Almerhag

International Conference on Signal and Image Processing (Signal 2014)

Analog Signal Processing Solution for Image Alignment..... 317 - 331
Nihar Athreyas, Zhiguo Lai , Jai Gupta and Dev Gupta

AN IMMUNE AGENTS SYSTEM FOR NETWORK INTRUSIONS DETECTION

Noria Benyettou¹, Abdelkader Benyettou² and Vincent Rodin³

¹University of Science and Technology of Oran Mohamed Boudiaf USTOMB,
SIMPA Laboratory, BP 1505 el Menaouar Oran-Algeria,

²University of Science and Technology of Oran Mohamed Boudiaf USTOMB,
BP 1505 el Menaouar Oran-Algeria,

³European University of Brittany, UMR CNRS 6285, Lab-STICC, CS93837,
29238 Brest CEDX3, France,

¹n.benyettou@gmail.com, ²a_benyettou@yahoo.fr

³Vincent.rodin@univ-brest.fr

ABSTRACT

With the development growing of network technology, computer networks became increasingly wide and opened. This evolution gave birth to new techniques allowing accessibility of networks and information systems with an aim of facilitating the transactions. Consequently, these techniques gave also birth to new forms of threats. In this article, we present the utility to use a system of intrusion detection through a presentation of these characteristics. Using as inspiration the immune biological system, we propose a model of artificial immune system which is integrated in the behavior of distributed agents on the network in order to ensure a good detection of intrusions. We also present the internal structure of the immune agents and their capacity to distinguish between self and not self. The agents are able to achieve simultaneous treatments, are able to auto-adaptable to environment evolution and have also the property of distributed coordination.

KEYWORDS

Intrusion Detection System, Artificial Immune System, Multi-Agents System.

1. INTRODUCTION

Networks safety and the intrusion detection systems are the subject of several works; first models goes back to 1984, they are focused on statistical analysis, expert system, and classification rules (IDES [5, 13], Nides[3,13], MIDAS[8], DIDS [15], NADIR [9], ADAM [4]). These models are already based on the attacks indexed in knowledge base. However, with the networks widening they generate much false alarm, and became less and less reliable to new attack's forms. To overcome difficulties met by these models, new research works are interested in multi-agents systems and immunology principles such as (MAAIS [19], NIDIMA [14], DAMIDAIIS [11], IMASNID [7], etc). These systems succeed in decreasing the false alarm rate thanks to the processes employed; namely communication process between the agents and the distinction process between self and not-self.

That is why, we present in this document a new model a Multi-Agents System (MAS) inspired by an Immune algorithm for the Intrusion Detection. Our choice is justified by the distributed and opened character of networks. Given the failure of the exist methods to detects new attacks; we integrate into our agents the artificial immune system mechanism. Artificial immune systems are inspired by the coordination principles and the parallel functioning of the biological immune system (life cycle, immunizing, immature tolerance, mature and memory lymphocyte).

2. INTRUSION DETECTION SYSTEM CHARACTERISTICS

To neutralize in real time illegal intrusion attempts, intrusions detection system must be executed constantly in the host or in the network.

The major inconveniences of the existing IDS [6] are:

1. Their difficulties to adapt oneself to the changes of the network architecture and especially how to integrate these modifications in the detection methods.
2. Their high rate of false-positives (false alert).

The intrusion detection system is effective if it has the following characteristics [12]:

- **Distribution:** to ensure the monitoring in various nodes of the network the analysis task must be distributed.
- **Autonomy:** for a fast analysis, distributed entities must be autonomous at the host level.
- **Delegation:** each autonomous entity must be able to carry out its new tasks in a dynamical way.
- **Communication and cooperation:** complexity of the coordinated attacks requires a correlation of several analyses carried out in network nodes.
- **Reactivity:** intrusion detection major goal is to react quickly to an intrusion.
- **Adaptability:** an intrusions detection system must be open to all network architecture changes.

Concepts of robustness, emergence, auto-organization, adaptability, and communication and cooperation are part of the basis fundament of the multi-agents systems .For this purpose we judged that the multi-agents systems (MAS) are very suitable to answered to these characteristics.

3. ARTIFICIAL MULTI-AGENTS IMMUNE SYSTEM

3.1 Multi-Agent System

The agents are able to achieve simultaneous treatments, are able to auto-adaptable to the evolution of environment and have also the property of distributed coordination.

The Multi-agents systems can be viewed as a collection of autonomous artificial entities able to perform various tasks through interaction, coordination, communication, collective intelligence and emergence of patterns of behavior.

Artificial immune system (AIS) is a set of algorithms inspired by biological immune system principles and functions. This last exploits the characteristics of natural immune system, as regards the learning and the memorizing in order to solve complex problems in artificial intelligence field.

The biological immune system is a robust and powerful process, known for its distributed simultaneous treatment orders of the operations and adaptive within the limit of its function [17]. Biological and multi-agents systems have common characteristics.

Biological cells are modeled by the agents; each agent is equipped with a set of receiver in its surface and has an internal behavior. Agents are submitted to environment rules and also to other agent's influence [18]. This is why it seems natural to model an intrusion detection system by the MAS based on biological immune systems principles.

Table 1. IB and AI common points

Biological immune system (IB)	Immune Agent (AI)
Antibody	Detector Agent
Antigen	The binary string From ip frame
Immune memory	memory Agent
The binding between antibody and antigen	Any intervals matching rule
Immune cells Lifecycle	detector agent Time- life
antibody/antigen Affinity	frame/ agent-detector Affinity

Table1 summarizes the main common points between biological immunity and the immunity agents in our model.

4. IMMUNE COMPONENTS DESCRIPTION

In this section, the principal immune components which are used in our architecture will be defined.

Antigens: they are considered in different approaches [7][11] as bit strings extracted from ip-packets, including ip address, port number, protocol type. Set $U = \{0,1\}^L$ ($L > 0$), and $Ag \subset U$, and the set U can be divided into self and notself. The self indicates normal network behavior; on the other hand, notself indicates the abnormal network [16].

Antibodies: correspond to bit strings, they have the similar length as antigens; antibodies are constantly in search of antigens in order to match them and also to increase their lifespan.

Set $AB = \{ab / ab = \langle b, t, ag \rangle, b, ag \in U \wedge t \in N\}$.

Where 'b' is the antibody bit string whose length is L, 'ag' is the antigen detected by the antibody and 't' is the antigen number matched by antibody[2]. There exist three states for antibodies: immature, mature and memory. Antibodies are able to detect an intrusion, in our architecture they are represented by D-agents.

Immature stage: Correspond to the first stage of our cell. In this stage, the immature Antibodies (Imb) are randomly generated by the generator detector. Immature immunocytes set is :

$Imb = \{ \langle b, t, ag \rangle \in Match / b \in U, t < \theta, ag = \emptyset \}$ and $Match = \{ \langle x, y \rangle / x, y \in U, f_{match}(x, y) = 1 \}$, which will evolve into Imb through self-tolerance. If an Antibody is not matched with notself for step

evolution; then it will die after a certain period of time.

Mature stage: Correspond to the second stage of our cell. In this stage the mature Antibodies (Mab) have failed to match with notself during activation and evolution;

Mature immunocytes set is

$Mab = \{ \langle b, t, ag \rangle \in Match / b \in U, \theta < t < \theta', ag \neq \emptyset \}$ and

$Match = \{ \langle x, y \rangle / x, y \in U, f_{match}(x, y) = 1 \}$.

In our work, if a Mab is not matched with notself after certain period of time then they will die.

Let us note that, dead is formulate by: $Ab_{dead} = \{ \langle b, t, ag \rangle \in Match / b, ag \in U, t \wedge \leq \theta \}$

Memory Stage: Correspond to the final stage of our cell. In this stage the memory antibodies (Meb) are the results of activation and evolution of the mature antibodies. Memory immunocytes set is

$Meb = \{ \langle b, t, ag \rangle \in Match / b \in U, t > \theta', ag = (ag_1, \dots, ag_n) \}$

and $Match = \{ \langle x, y \rangle / x, y \in U, f_{match}(x, y) = 1 \}$.

They have significant lifespan as long as they succeed matching with not-self.

Affinity characterizes the correlation between Antigens and Antibodies is to determinate the. According to Hamming Distance (HD) this major element is evaluated.

The calculation formula is evaluated according to Hamming Distance (HD).

Let us consider x_i ($i=1 \dots L$) the bit string of length L and y_i ($i=1 \dots L$) another bit string of the same length L . x_i represents Antigen and y_i represents an Antibody. α is the affinity matching threshold value and $HD(x, y)$ is the different sum of the bits in the two strings.

The affinity function is calculated as follows:

$$\begin{cases} 1, & HD(x, y) \geq \alpha \\ 0, & otherwise \end{cases} \quad \text{and} \quad DH(x, y) = \sum_i^L = 1^\alpha \quad \text{with } \alpha = 1 \text{ if } x_i \neq y_i, \alpha = 0 \text{ else}$$

5. RELATED WORKS

The idea of using, the artificial immune systems for intrusion detection, in distributed networks, appears recently; to our knowledge, one of the first work, was developed by Hofmeyer and Forest in 1999 [1]. Their model is implemented on distributed network architecture and in each host, a frame which is received is represented by a binary string non-self.

This string is analyzed by another binary string self of the same size L . The self represent the immune detectors which are randomly generated by the system, and which must match with not-self, in order to evolve and to change state (immature / mature & naive, or death); after the mature detector is activate and had a co stimulation it become a memory detector with an infinite lifespan. This technique is responding quickly to a possible intrusion of the same type. But the communication between the host different network is not existing.

Another architecture is proposed by Sunjun, the Immune Multi-agent Active Defense Model for Network Intrusion (IMMAD) in 2006[16]. This model is built for monitoring multilayer of

network, by a set of agents that communicate and cooperate at different levels. The immune agent (IMA) is the security state of computers monitoring, is installed in each network node, and consists of self, immature or mature antibodies, memory antibodies, etc. This immunological mechanism permits him to detect an intrusion and send the message to Local Monitor Agent (LMA). LMA analyzes the state of the local area network, and vaccinate all the nodes in the same segment, after having evaluated the risk of intrusion, and it informs Central Monitor Agent (CMA) of the new type of intrusion. CMA supervises the whole of the network and increase security across the network.

Another architecture which seem to us interesting, is proposed by NianLiu in 2009. This architecture is called Network Intrusion Detection Model Based on Immune Multi-Agent (NIDIMA)[4]. This model ensures the security in the distributed networks against intrusions. In this architecture, each agent Security situation is composed of several agents immune (IA). The IA are distributed on each node of the network, they are the firsts to identify the events of intrusion and are blocking them. If the attacks are not known through learning and memory, it sends information to the agent SMC. The agent SMC analyzes the intrusion, which it received by each AI in network segments and it surfs at network segments to vaccinate them against another intrusion. Agent security situation evaluation gathers information of subnets and host from each agent security situation, and evaluates the risks to integrated the whole network. This information includes the type, quantity, strength and harmfulness of the attacks.

There exist many other models, but we decided to present those which seem to us closest to our architecture. Let us recall that our aim is to increase immunity and to decrease the rate of false alarm.

6. ARCHITECTURE OF MODEL

In this article, we employed a new model, based on Multi-Agent paradigm and Immune algorithm for Intrusion Detection. We describe a model through the dynamic behavior of immune agents, the distinction between self and notself. We expose the architecture of the distributed model, the agents' behavior for insuring the network security, in order to avoid false alert triggering. The system is installed in each Host/Server, and the system agents cooperate and communicate for best and more reliable intrusion detection.

6.1. The behavior of immune agent in this architecture

Detector agent (D-agent) is the principal component for the distinction between self and non self through the sensor/analyzer, which identifies the frame (these agents are in immature state).

The sensor /analyzer is composed of two bit strings: a random bit string which analyses frame by calculating the Hamming Distance (HD); and a stationary bit string which includes host and network information; the stationary part is identical in all D-agents of the same host. To avoid any false alarm, D-agent sends its 1st report (if $HD_{int} > Val$) to Alert agent (A-agent) when it detects an anomaly.

A-agent will evaluate the intrusion importance according to the results obtained. However, it can not trigger the alarm, while it has not received any confirmation from several D-agents, within the same host or from other A-agent within the network [13]. (See Figure1).

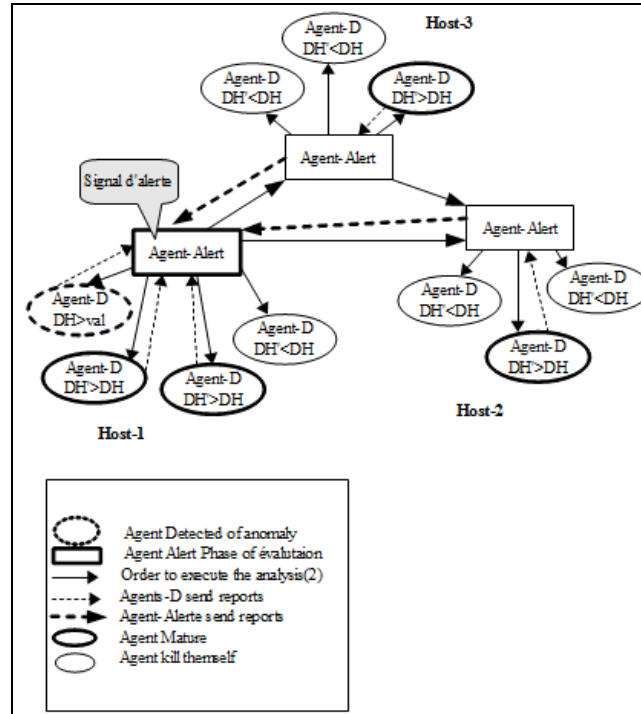


Figure1. Immune Agents Cooperation

Intrusion assessment allows to the A-agent to ignore warning message when the evaluation is tiny; or to be under-monitoring where the evaluation is important.

In this case, the A-agent sends to all D-agents the order to execute the analysis stage (2) for all treated frames.

When D-agent receives this order, a semi-sensor is generated at random, on the basis of the code of D-agent which has detected the anomaly. Thus all frames will be first analyzed by the sensor/analyzer, then by the semi-sensor in each analysis, a new Hamming distance (HD') is evaluated (mature state).

The D-agents which detect ($HD' > HD_{int}$), send their reports to their A-agents and increase lifespan (Memory Phase), the other D-agents decrease their lifespan and when they reached a threshold they kill themselves. D-agents exchange between them the second analysis results, they trigger also the alert if the risk assessment is the same; (See Figure2).

This parallel analysis technique's allows a best management of false alarm and a better network supervision against the intrusion.

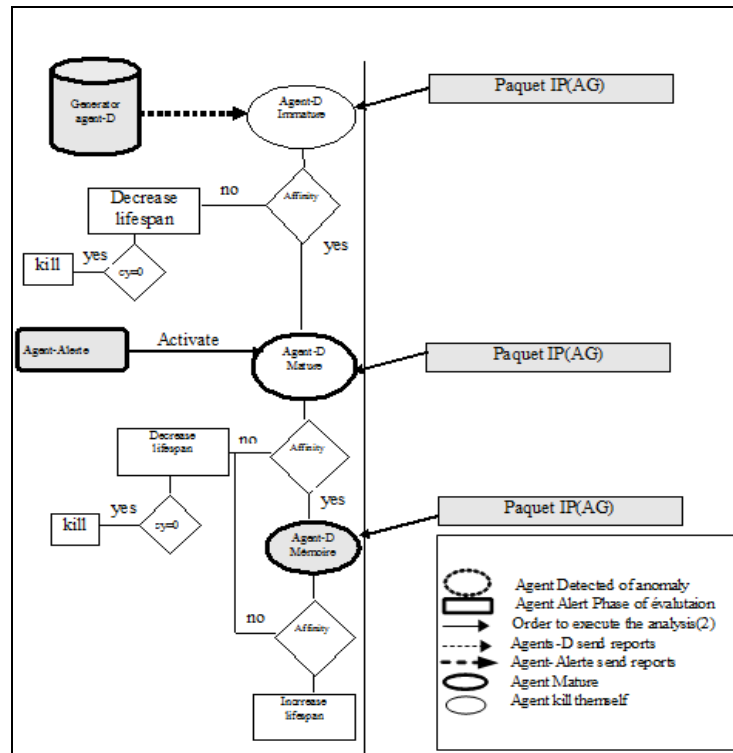


Figure 2. Dynamic evolution of Agent-D

6.2 Analysis process

Immune agents present in our model analyze the incoming IP-packet in by *D-agent* (*memory*) in order to detect intrusions known by the system (acquired immunity).

State Analyses by *D-agent* (*memory*):

- When an anomaly is detected the *D-agent* (*memory*) blocks the frame, and
- If not, IP-packet is transferred to a second analysis (*D-agent* (*mature*)).

***D-agent* (*mature*) State Analyses**

In this stage, *D-agent* (*mature*) analyses the IP-packet by the (sensor/ analyzer). Two cases could occur:

Case 1: When an anomaly is detected,

- *D-agent* sends its 1st report to *A-agent* if ($HD_{int} > Val$).

According to the results obtained *A-agent* will evaluate the intrusion importance. This evaluation is considered as important if the intrusion is detected by several *D-agents*, from the same host.

Intrusion assessment allows *A-agent*

- To ignore the warning message if this evaluation is insignificant or;
- To Activate all *D-agents* (*mature*) to execute (semi-sensor)

When *D-agents* receive this order, they generate at random a semi-sensor on the basis of *D-agent* code (which has detected the anomaly). Thus IP-packets will be first analyzed by the sensor/analyzer, then by the semi-sensor in each analysis, and a new Hamming distance (HD') is evaluated.

The *D-agents* which will detect ($HD' > HD_{int}$),

- Send their report to their *A-agents*
- Block this ip-packet and increase their lifespan. When their life time reached ($T_m = \theta'$), they become memory *D-agents* (with $T_m = T_e$).

The other *D-agents* decrease their lifespan, they kill themselves when the threshold becomes null ($T_m = 0$).

A-agents exchange between them analysis results and they trigger the alert if the risk assessment is similar (see Figure3).

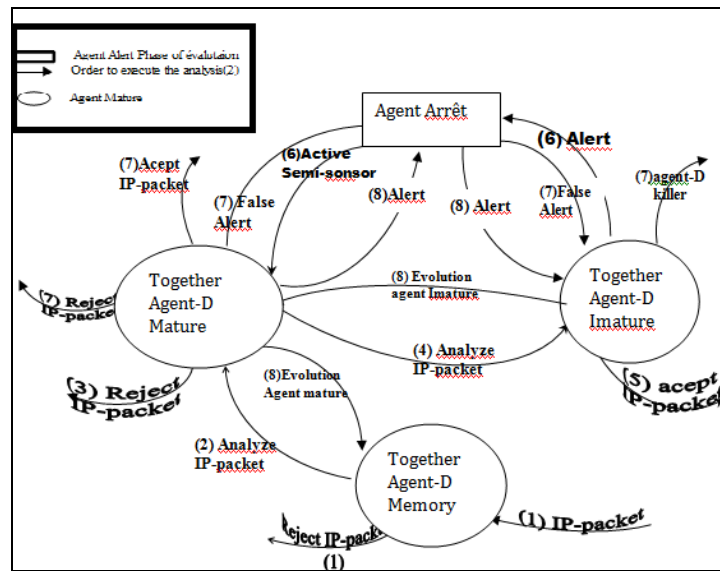


Figure 3. IP-Packet analysis Process

Case-2: if no anomaly is detected

- IP-packet is transferred to the *D-agents* (Immature) (fourth analysis stage).

Immature *D-agent* State Analyses

In this state the *D-agent* (Immature) analysis the IP-packet,

- if no anomaly is detected , IP packet is authorized to pass,
- if an anomaly is detected by this agent, it sends alert to *A-agents*. Thus, two cases could occur:

Case-2.1: *A-agent* rejects this alert

- Then IP-packet is authorized to pass
- The *D-agents* (Immature) concerned by this alert decrease their lifespan, when they arrived at a threshold they kill themselves (see figure 2 & 3).

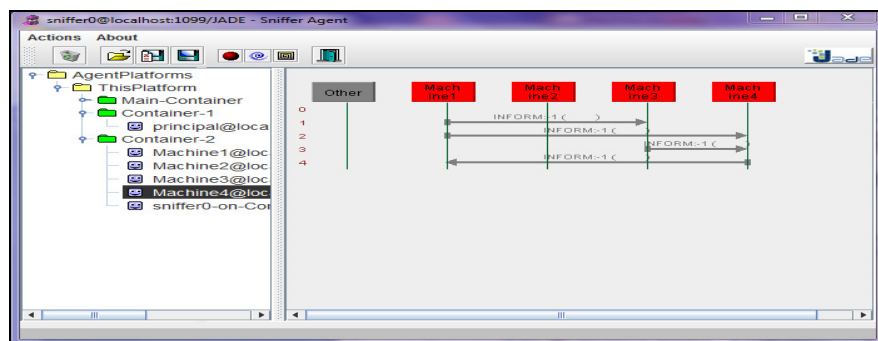


Figure 5 communication between immun-agents in jade

8. CONCLUSION

In this paper we raised problems involved in existing intrusion detection system to cope with the techniques employed by the Hackers. These techniques consist in circumventing the measurements of security by fraudulent behaviors in spread networks; consequently networks became more vulnerable to new types of attacks. A good intrusion detection system must take into account complexity and increasing dynamicity of networks. We proposed a new model of artificial immune system for intrusion detection based on multi-agents systems.

This model is inspired from biological immune principles, by the cooperation of immature D-agent, mature D-agent and memory D-agent.

The D-agent structure allows him to accomplish a double analysis for all frames. This analysis technique permits accelerating the immune response and detecting the intrusion to the shared resources. Furthermore, this distributed analysis mobilized several kind of agent in order to analyze the different sort of intrusion.

Our system adapts to the growing change of the environment of the network thus, it answers favorably at problematic.

REFERENCES

- [1] A.Hofmeyer& S.Forrest, Immunity by Design An Artificial Immune System, In Proceedings of 1999 GECCO Conference, 1999
- [2] C.ChungMing& all:Multi-Agent Artfial Immune Systems(MAAIS)for Intrusion Detection: Abstraction from Danger Theory, KES-AMSTA2009,LNAI5559,pp.11-19,2009
- [3] D.Anderson& all: Next-generation Intrusion Detection Expert System (NIDES): Software Users Manual, 1994
- [4] D.Barbara & all: ADAM: Detecting Intrusions by Data Mining,Proceedings of the IEEE Workshop on Information Assurance and Security, West Point, NY, June 5-6, 2001
- [5] D.E. Denning and P.G. Neumann. Requirements and model for IDESla real-time intrusion detection expert system. Technical report, Computer Science Laboratory, SRI International, Menlo Park, CA (USA), 1985.
- [6] F.Majorczyk& all: Experiments on COTS Diversity as an Intrusion Detection and Tolerance Mechanism. Workshop on Recent Advances on Intrusion-Tolerant Systems (WRAITS). March 2007.
- [7] D. Wang, T. Li, S. J. Liu, G. Liang and K. Zhao. An Immune Multi-agent System for Network Intrusion Detection. In Proceedings of the 3rd International Symposium, ISICA 2008, Wuhan

- (China), 19-21 December, 2008. Springer, Lecture Notes in Computer Science, Vol. 5370, Advances in Computation and Intelligence, pages 436-445, 2008.
- [8] H.Arlowe.D& all.: The Mobile Intrusion Detection and Assessment System (MIDAS), in Proceedings of the Security Technology Conference, Location TBD, October 10-12, 1990, 54-61
- [9] J.Hochberg& all: NADIR: An automated system for detecting network intrusions and misuse, Computers and Security 12(1993)3, May, 253 - 248
- [10] J. Kim &all: Towards an artificial immune system for network intrusion detection: an investigation of clonal selection with a negative selection operator. Proc.Congress on Evolutionary Computation, South Korea, 2001, vol. 2, pp.1244-1252.
- [11] J. Yang, X. Liu, T. Li, G. Liang and S. Liu. Distributed agents model for intrusion detection based on AIS. Knowledge-Based Systems, Elsevier, Volume 22, Issue 2, pages 115-119, March 2009.
- [12] K. Boudaoud, Z. Guessoum. A Multi-agents System for Network Security Management. In Proceedings of the 6th IFIP Conference on Intelligence in Networks (SmartNet), pages 407-418, Vienna, (Austria), 18-22 September 2000.
- [13] N.Benyettou, A.Benyettou,V.Rodin An Immune Multi-Agents System used in the Intrusion Detection System in distributed Network.ICARIS 2012, 11th International Conference on Artificial Immune Systems, Poster session, page 36 (conference programme), Taormina (Italy), 28-31 August 2012.
- [14] N. Liu, S. Liu, R. Li, Y. Liu. A Network Intrusion Detection Model Based on Immune Multi-Agent. International Journal of Communications, Network and System Sciences (IJCNS), Volume 2, Number 6, pages 569-574, September 2009.
- [15] R.Snapp& all.: DIDS (Distributed Intrusion Detection System) - Motivation, architecture and an early prototype, Proc. of the 14th National Computer Security Conference, Washington, D. C., Oct. 1991, 167 – 176.
- [16] S. Liu, T. Li, D. Wang, K. Zhao, X. Gong, X. Hu, C. Xu and G. Liang. Immune Multi-agent Active Defense Model for Network Intrusion. In Proceedings of the 6th International Conference, SEAL 2006, Hefei (China), 15-18 October 2006. Springer, Lecture Notes in Computer Science, Volume 4247, Simulated Evolution and Learning, pages 104-111, 2006.
- [17] U. Aickelin and D. Dasgupta. Artificial Immune Systems. A book chapter in Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques, Ed. E.K. Burke and G. Kendall, Springer, Chapter 13, pages 375-399, 2005.
- [18] V. Rodin, A. Benzinou, A. Guillaud, P. Ballet, F. Harrouet, J. Tisseau and J. Le Bihan. An immune oriented multi-agent system for biological image processing. Pattern Recognition, Elsevier, Volume 37, Issue 4, pages 631-645, April 2004.

INTENTIONAL BLANK

AN INVESTIGATION ON SWITCHING BEHAVIOURS OF VECTOR CONTROLLED INDUCTION MOTORS

Yılmaz Korkmaz¹, İsmail Topaloğlu², Hayati Mamur² and Fatih Korkmaz²

¹Faculty of Technology, Department of Electrical-Electronic Engineering,
Gazi University, Ankara, Turkey

²Faculty of Engineering, Department of Electrical-Electronic Engineering,
Çankırı Karatekin University, Çankırı, Turkey
fkorkmaz@karatekin.edu.tr

ABSTRACT

Field oriented control and direct torque control are the most popular methods in high performance industrial control applications for induction motors. Naturally, the strengths and weaknesses of each control method are available. Therefore, the selection of optimum control method is vitally important for many industrial applications. So, the advantages and the disadvantages of both control methods have to be well defined. In this paper, a new and different perspective has been presented regarding the comparison of the inverter switching behaviours on the FOC and the DTC drivers. For this purpose, the experimental studies have been carried out to compare the inverter switching frequencies and torque responses of induction motors in the FOC and the DTC systems. The dSPACE 1103 controller board has been programmed with Matlab/Simulink software. As expected, the experimental studies have showed that the FOC controlled motors have had a lessened torque ripple. On the other hand, the FOC controlled motor switching frequency has about 75% more than the DTC controlled.

KEYWORDS

Induction motor; machine vector control; motor drives; switching frequency

1. INTRODUCTION

Nowadays, induction motors have wide range use in many industrial applications with its simple and robust structure, low costs and reliable operation. Furthermore, the induction motors can be easily controlled due to development of new control methods in the last few decades. There are two well-known control methods in high performance control of induction motors: Field oriented control (FOC) and direct torque control (DTC).

FOC was first introduced by Blaschke in the 1970s. It was unrivalled in industrial induction motor drivers until DTC was introduced by Takahashi in the 1980s. Since that time, there have been continual discussions and studies in scientific and industrial arenas regarding which one is the best for the high performance control of induction motors [1-3].

The basic idea of vector control is the control of motor flux and torque separately as DC motors. For this aim, motor currents have converted two phase vector components using park or Clarke

transformations. One of these is related with the component controlled flux vector, and the other one is related with the controlled torque vector. The main difference between these two methods is that the FOC controls by a rotor or stator field orientation, while the DTC controls by a stator field observation [4]. Depending on this difference, the structural differences in the two control strategies are that the FOC uses park transformation, more machine parameters, and current regulators, while the DTC uses Clarke transformation, fewer machine parameters, and any current regulators. Thus, comparative studies between the two methods show that the DTC is simplicity, a fast dynamic response, and is robust to parameter changes. Despite all these advantages, the DTC also has some handicaps: the most important of them being high torque and current ripples. Evidently all users have to take into account all these advantages and disadvantages when deciding on which method they will use on the motor drivers [5]. There are some comparative studies regarding comparison of the both methods, which address the motors speed and torque behaviours. Thanks to this research, it is now known that despite a high torque ripple, the DTC has a fast dynamic response [6-7].

In this study, we aimed to give a new and different perspective regarding the comparison of the FOC and the DTC drivers. We compared voltage source inverter switching frequencies on both the FOC and DTC systems. There is no doubt about the importance of switching frequency in motor driver systems because this directly affects switching losses and it means also affects indirectly the efficiency of the drivers. Experimental studies have been carried out to compare switching frequencies and the dSpace 1103 controller board has been programmed with Matlab/Simulink.

2. BASICS OF FOC AND DTC

FOC is the first vector control method developed for induction motors where it is mostly used to control the speed of the motor, not control of moment, due to its low level sensitivity [8].

In FOC, the parts of the motor have to be turned in a d-q reference frame, which then turns in a synchronous speed (park transformation). Thus, the position of the rotor flux needs to be well determined for the success of this transformation. Two basic approaches are used in determining the process of the rotor flux position. The first approach is to use flux sensors to determine the rotor flux position. The second approach is to measure the rotor position with an incremental encoder and calculate the angle between the axis of the rotor and the flux. The stator current d-q components are calculated by using Eq. 1 and Eq. 2.

$$i_{qs}^* = \frac{2}{3} \frac{2}{p} \left(\frac{T_e^*}{\lambda_r^*} \right) \left(\frac{L_r}{L_m} \right) \quad (1)$$

$$i_{ds}^* = \frac{\lambda_r^*}{L_m} \quad (2)$$

Where i_{qs}^* and i_{ds}^* are the stator current d-q components reference values, λ_r^* is the rotor flux reference value, and L_m is the mutual inductance, L_r is rotor inductance and p is motor pole pairs value in these equations.

In Eq. 3., θ indicates the rotor angular position, and can be obtained from the integration of the sum of the rotor angular speed (ω_r) and the sleep angular speed (ω_{sl}) as given below.

$$\theta = \int (\omega_r + \omega_{sl}) dt \quad (3)$$

and the slip angular speed is obtained from Eq. 4.

$$\omega_{sl} = \frac{L_m R_r}{L_r \lambda_r^*} i_{qs}^* \quad (4)$$

The FOC control block diagram is created by using from Eq.1- Eq.4 and is shown in Fig. 1[9].

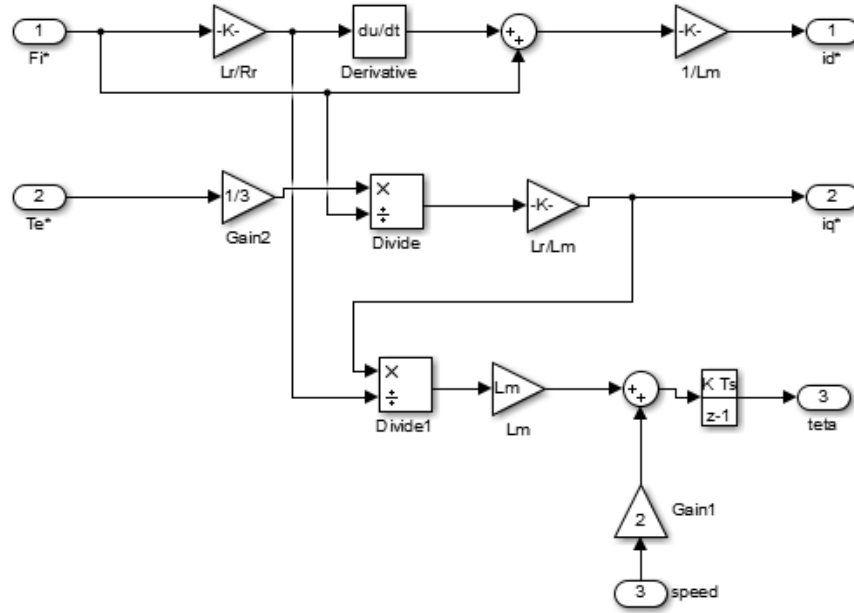


Fig. 1. Simulink block diagram of the FOC controller

In DTC, the stator flux and the torque are directly controlled by the system with using measured currents and voltages.

Instantaneous values of the flux and torque are calculated by using the transformation of the measured currents and the voltages of the motor. The stator flux is calculated as given in Eq.5-7 in a stationary reference frame [10].

$$\lambda_{\alpha} = \int (V_{\alpha} - R_s i_{\alpha}) dt \quad (5)$$

$$\lambda_{\beta} = \int (V_{\beta} - R_s i_{\beta}) dt \quad (6)$$

$$\lambda = \sqrt{\lambda_{\alpha}^2 + \lambda_{\beta}^2} \quad (7)$$

Where, λ_{α} - λ_{β} are stator fluxes, i_{α} - i_{β} are stator currents, V_{α} - V_{β} are stator voltages, α - β components and R_s is the stator resistance. Motor torque can be calculated as given in Eq.8.

$$T_e = \frac{3}{2} p (\lambda_\alpha i_\beta - \lambda_\beta i_\alpha) \quad (8)$$

Where, p is the motor pole pairs. The stator flux vector region is an important parameter for the DTC, and it can be calculated as given in Eq.9:

$$\theta_\lambda = \tan^{-1} \left(\frac{\lambda_\beta}{\lambda_\alpha} \right) \quad (9)$$

The torque and flux errors, which are obtained by comparing the reference and observed values, are converted to control signals by hysteresis comparators. The switching table is used to determine the optimum switching inverter states, and it determines the states by using the hysteresis comparators outputs and the flux region data. The schematic view of the DTC system is given in Fig. 2 [10].

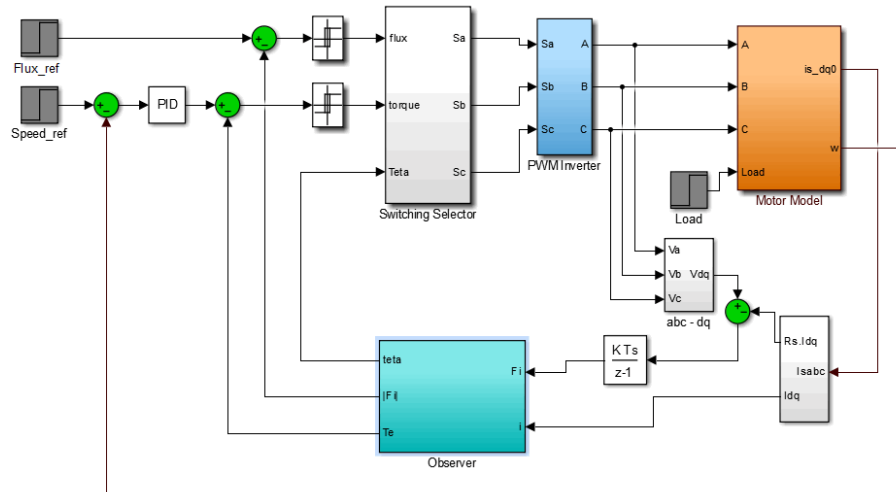


Fig. 2. Simulink block diagram of the conventional DTC

3. EXPERIMENTAL STUDIES

In order to compare the switching frequency between the DTC and the FOC systems, the experimental studies have been carried out by using the dSpace 1103 DSP (Digital Signal Processor) board. The board has been programmed in Matlab-Simulink Real-Time-Workshop environment for the tests.

A new frequency measurement block has created to compare the switching frequency on both systems. The inverter has had three arms and each arm frequency has been measured separately. The inverter switching frequency has been obtained by the addition of the arms switching frequencies, and then dividing the total by three. However, especially in the DTC, the switching frequency has a wide range, so the average switching frequency has been calculated with a different sampling time (100 ms) in the frequency measurement block to obtain healthy comparable results.

The results of the experimental tests obtained in this study are for the induction motor of 4 kW and the parameters of the motor, and the experiments are as given below. The machine model is implemented for the DTC and the FOC schemes using the Matlab/Simulink. The DTC and the FOC systems were tested at under no load conditions. The parameters of the three-phase induction motor, in the SI units are given in Table 1.

Table 1. The parameters of the three-phase induction motor

Motor parameters (SI)	
Power (kW)	4
Voltage (V)	380
Current (A)	8.2
$\cos\phi$	0.85
Speed (rpm)	1425
Frequency (Hz)	50
Stator resistance (Ω)	7.2

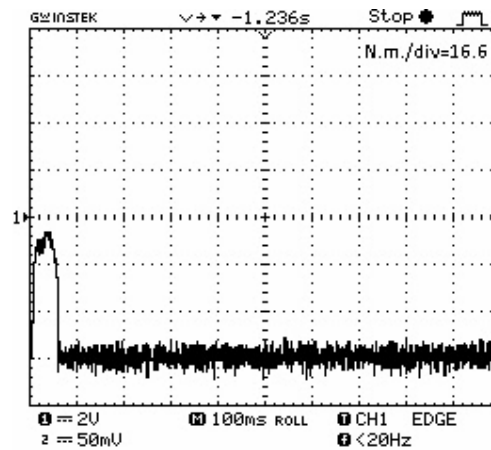


Fig. 3. Torque responses of unloaded DTC controlled motor

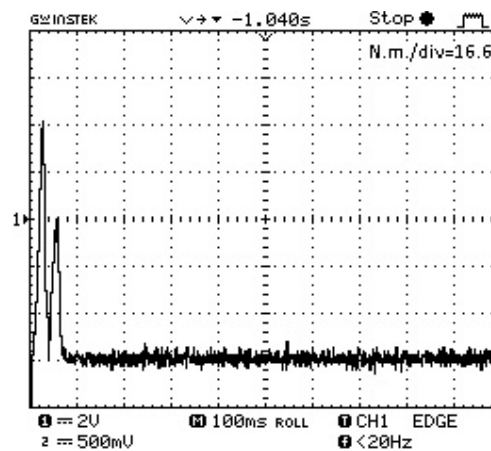


Fig. 4. Torque responses of unloaded FOC controlled motor

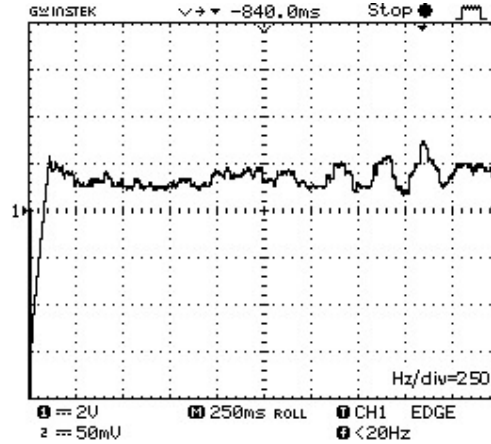


Fig. 5. Switching frequencies of unloaded DTC controlled motor

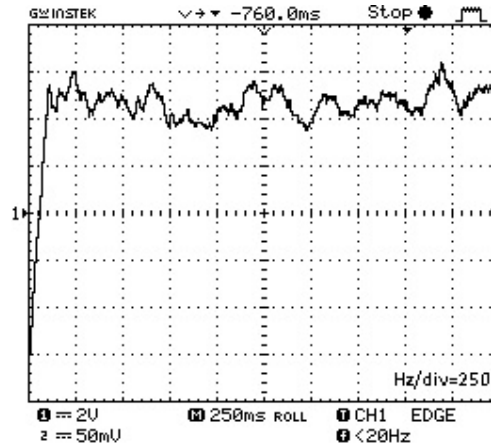


Fig. 6. Switching frequencies of unloaded FOC controlled motor

The motor reference speed is constant at 1500 rpm over all the experimental tests. In the first part of the tests, the motor has been operated as unloaded. The moment and switching frequency data, which has been calculated within the Simulink blocks, have been obtained by using the digital-analogue converters of the dSPACE 1103 controller board. The torque response curves of the motor for the both systems are given in Figure 3 and Figure 4 for unloaded working conditions. As expected, it can be clearly seen that the DTC has much torque ripples when compared the FOC.

Figure 5 and Figure 6 show change on the switching frequencies of the motor for both systems. The switching frequency of the DTC controlled motor is about 900 Hz, while the FOC controlled motor is about 1600 Hz. Therefore, it can be stated that the FOC controlled motor switching frequency has been 75% more than the DTC controlled one.

4. CONCLUSIONS

When designing an industrial application which includes electric motors, it is vitally important to choose motor drive method in order to operate the designed system efficiently. However, many times, it can be difficult to decide for any optimum control method for obtaining high performance regarding to induction motor drivers. Essentially, there are two options for the high

performance induction motor drivers: FOC and DTC. This paper has aimed to give fair comparison between two vector control schemes for induction motor drives. For this aim, experimental tests have been realized to compare of the motor performances on different conditions.

After all experimental studies, it must be pointed out that the DTC method is preferable if the fast dynamic performance has primary importance whereas the FOC method might be a better option when high torque quality is demanded. In addition, a new frequency measurement block has been created to compare the switching frequency on both systems. The inverter switching frequencies have been investigated in the FOC and DTC systems to give a new criterion for the selection of optimum control strategy for induction motor high performance control. The experimental studies have been carried out to compare the switching frequencies in both methods under different load conditions. As a result, the FOC controlled motor switching frequency has been almost 75% more than the DTC controlled one under loaded and unloaded working conditions. The choosing of the DTC scheme will also provide high energy efficiency driver if the dynamic performance of the motor has primary importance.

REFERENCES

- [1] V. Bleizgys, A. Baskys and T. Lipinskas, "Induction motor voltage amplitude control technique based on the motor efficiency observation", *Electronics and Electrical Engineering – Kaunas: Technologija*, 2011. – No. 3(109) – P. 89–92.
- [2] Blaschke, F., "The Principle of Field Orientation Applied to The New Transvector Closed-Loop Control System for Rotating Field Machines", *Siemens-Rev.*, 1972. – Vol. 39 – P. 217–220.
- [3] Takahashi, I., Naguchi, T. A., "New Quick-Response and High-Efficiency Control Strategy of an Induction Motor", *IEEE Transactions on Industry Application*, 1986. IA-22(5) –P. 820–827.
- [4] Liu-Jun; Wang Wan-li, Wang Yang. "Research on FOC and DTC switching control of asynchronous motor based on parameter estimation", *Automation and Logistics ICAL 2008. IEEE International Conference*, 2008. – P. 1754–1758.
- [5] Farasat, M.; Rostami, N.; Feyzi, M. R., "Speed sensorless hybrid field oriented and direct torque control of induction motor drive for wide speed range applications", *Power Electronic & Drive Systems & Technologies Conference (PEDSTC)*, 2010. – P. 243–248.
- [6] Casadei, D.; Profumo, F.; Serra, G.; Tani, A., "FOC and DTC: two viable schemes for induction motors torque control", *Power Electronics, IEEE Transactions*, 2002. –Vol.17– No.5 – P. 779–787.
- [7] Wolbank, T.A.; Moucka, A.; Machl, J.L., "A comparative study of field-oriented and direct-torque control of induction motors reference to shaft-sensorless control at low and zero-speed", *Intelligent Control*, 2002. *Proceedings of the 2002 IEEE International Symposium*. – P. 391–396.
- [8] T. V. Mumcu, I. Aliskan, K. Gulez, G. Tuna, "Reducing Moment and Current Fluctuations of Induction Motor System of Electrical Vehicles by using Adaptive Field Oriented Control", *Electronics and Electrical Engineering – Kaunas: Technologija*, 2013. – Vol. 19 – No.2 – P. 21–24.
- [9] Krishnan, R. *Electric Motor Drives* // Prentice Hall, P.450 2001.
- [10] Fatih Korkmaz, M.Faruk Çakır, İsmail Topaloğlu, Rıza Gürbüz, "Artificial Neural Network Based DTC Driver for PMSM", *International Journal of Instrumentation and Control Systems (IJICS)* Vol.3, No.1, pp. 1-7 Jan. 2013

INTENTIONAL BLANK

PERFORMANCE ANALYSIS OF THE RECENT ROLE OF OMSA APPROACHES IN ONLINE SOCIAL NETWORKS

J. Ashok Kumar¹, S. Abirami², S. Murugappan³

Research Scholar¹, Assistant Professor², Associate Professor³

^{1,2}Department of Information Science and Technology,
Anna University, Chennai, India

jashokkumar83@gmail.com, abirami@annauniv.edu

³Department of Computer Science,
Tamil Nadu Open University, Chennai, India
drmryes@gmail.com

ABSTRACT

In this emerging trend, it is necessary to understand the recent developments taking place in the field of opinion mining and sentiment analysis (OMSA) as part of text mining in social networks, which plays an important role for decision making process to the organization or company, Government and general public. In this paper, we present the recent role of OMSA in Social Networks with different frameworks such as data collection process, text pre-processing, classification algorithms, and performance evaluation results. The achieved accuracy level is compared and shown for different frameworks. Finally, we conclude the present challenges and future developments of OMSA.

KEYWORDS

Sentiment Analysis, Opinion Mining, Classification Algorithms, Social Media.

1. INTRODUCTION

Opinion Mining and Sentiment Analysis (OMSA) plays a vital role in social media to get positive or negative sentiment and opinions expressed by the user's or public using the mode of online feedback forms, emails and OSN websites such as Facebook, Twitter, LinkedIn, YouTube, MySpace, Blogs and forums etc. Shusen Zhou et al. [14] states that OSN sites are one of the most important tools of the Web 2.0 to share or disseminate views. OMSA helps a lot to predict the product sales, service, quality, policy initiatives, Institutions, forecasting political opinions, and news contents for the company or organization, Government and general public. The main task of OMSA is used to classifying the polarity at the document, sentence, or feature / aspect, and which are expressed as positive, negative or neutral. The sentiment analysis research is also done at this polarity level. The general system architecture of OMSA is constructed as shown Fig. 1, and the main characteristics are analyzed like [1] educational data mining approach and reported performances [12]. This paper is organized as follows. Section 2 presents the recent developments in the field of OMSA with different frameworks and algorithms. Section 3 discusses the obtained results by using datasets and its volume. Section 4 states the challenges and future developments. Finally, Section 5 concludes the paper.

David C. Wyld et al. (Eds) : SAI, CDKP, ICAITA, NeCoM, SEAS, CMCA, ASUC, Signal - 2014

pp. 21–32, 2014. © CS & IT-CSCP 2014

DOI : 10.5121/csit.2014.41103

2. RECENT DEVELOPMENTS IN OMSA

This section presents the recent development of OMSA approach with different frameworks, methods, techniques, and algorithms. The main characteristics of OMSA approach is shown in Table 1, which gives the complete or overall reference to the researchers and the process is explained below. It is necessary to understand the present work to carry out future work without duplication. First, the various framework and algorithms in opinion mining with data collection approach, pre-processing stage and classification of polarity. Second, describes the various framework and algorithms in sentiment analysis with data collection approach, pre-processing stage and classification of polarity.

2.1 Twitter Opinion Mining (TOM) Frame work and Polarity Classification Algorithm

Farhan Hassan Khan et al. [5] proposed a new TOM framework to predict the polarity of words into positive or negative feelings in tweets, and to improve the accuracy level of this classification. TOM framework is constructed into three stages. First, data acquisition process, which is used to obtain the Twitter feeds with sparse features through Twitter streaming API from OSN. Twitter4J library has been used to extract only English language tweets. Second, pre-processing, which process each tweets individually for the refinement operations such as detection and analysis of slangs/abbreviations, Lemmatization and correction, and stop words removal. Then the refined tweets pass into the classifier. Third, Polarity Classification Algorithm (PCA), it classifies the twitter feeds on basis of Enhanced Emoticon Classifier (EEC), Improved Polarity Classifier (IPC), and SentiWordNet Classifier (SWNC) by using set of emoticons, a list of positive and negative words, and SentiWordNet dictionary respectively. In this stage, reducing the number of neutral tweets is the major issue. For this problem, final classification is performed to indicate more accurate results than its predecessors based on the scores of EEC, IPC, and SWNC.

2.2 Standard election prediction model by using User Influence factor

Malhar Anjaria et al. [10] introduced a model to predict the election result by applying the user influence factor (re-tweets and each party garners) and extracting opinions using direct and indirect feature on the basis of the supervised algorithms such as NB (simple probabilistic model), MaxEnt (Uniform classification model), SVM (achieves maximum margin hyper plane), ANN (feed forward network), and SVM with PCA (dimension reduction). This model is built into several steps. Step1, data collection approach is used to collect tweets with Candidate's name. Step2, normalization and feature reduction includes the refinement operations Internet acronyms and emoticons, duplicate tweets, candidate accounts, word expansion, URLs, repeated words and repeated characters for to get original sentence format into usable format of tweets. Step3, feature extraction and extended terms used for the purpose of unigram, bigram and a unigram + bigram. Step4, machine learning methods, in which the polarity based classification applied to independent features, and it fails to capture query specific sentiments so that the aforementioned text classification algorithms used for direct features. Setp5, incorporating sentiment analysis considers re-tweets into account for influence factor. In Step6, analyzing gender based votes based on term frequency. Finally, US Presidential Election result 2012 and Karnataka State Assembly Election 2013 results are shown that Twitter provides a reasonable accuracy.

2.3 Similarity shaped membership function

Tapia-Rosero A et al. [15] employed a method to detect similarity shaped membership functions in group decision making process. This method is constructed by using the symbolic notation, similarity measure, and grouping membership functions by shape similarity. The symbolic notation has two component algorithms to get shape-string and feature-string which represent a sequence of symbols and sequence of linguistic terms respectively for each segment of membership functions. The similarity measure used the linguistic terms from extremely short to extremely long to obtain an overall similarity at unit interval. In a group decision making environment, the grouping membership functions by shape similarity aims to gather groups using similarity matrix.

2.4 Three-level similarity method (TLSM)

Jun ma et al. [8] stated a method to reduce the chance of applying inappropriate decisions in the multi-criteria group decision making (MCGDM). In this aspect, a Gradual Aggregation Algorithm (GAA) developed and established a TLSM. GAA faced two practical issues. First, how to handle missing values. Second, how to generate a decision dynamically in MCGDM?. To solve these issues, GAA is implemented in two ways, i.e., OGA (ordinary gradual algorithm) and WGA (weighted gradual algorithm). The OGA does not explicitly process the criteria weights and leaves it to the aggregation operator but the WGA does. TLSM will be measuring the similarity of two participants opinion at three sequence levels, i.e., assessment level, criterion level, and problem level. In Level-1, the term set will be divided into several semantic-equal groups for the criterion and then used HCFSM. In Leve-2, identifies an appropriate SUF for each criterion at PSA and PRW. In Level-3, each individual criterion provides a single perspective to observe similarity of two opinions.

2.5 Unsupervised dependency analysis-based approach

Xiaolin Zheng et al. [17] presented an unsupervised dependency analysis-based approach to extract AEP (Appraisal Expression Pattern) from reviews. The problem statement is defined at different terminologies such as domain, aspect word, sentiment word, background word, and review. Then the AEP is applied to represent the syntactic relationship between aspect and sentiment words by using Shortest Dependency Path (SDP), Confidence score (CS) and parameter inference. Finally, AEP-LDA (Latent Dirichlet Allocation) model is employed to jointly identify the aspect and sentiment words. It outperforms the base line method when compared to other supervised methods such as LDA, Local-LDA, Standard LDA, AEP-LDA (no-AEP).

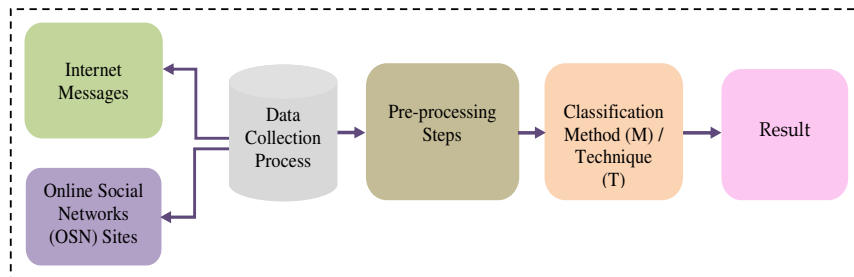


Fig. 1 General System Architecture of OMSA approach

2.6 SuperedgeRank algorithm

Ning Ma et al. [11] introduced a SuperedgeRank algorithm to identify opinion leaders in online public opinion supernetwork model. This model is built up into three stages. First, data processing which identifies the main information from the public comments collected from internet and then applies the ICTCLAS for splitting sentence into words. Second, online public opinion supernetwork modeling includes four types of elements i.e., Social Subnetwork, Environment Subnetwork, Psychological Subnetwork, and Viewpoint Subnetwork to form four layers of supernetwork. Third, Indexes of online public opinion supernetwork includes Node Superdegree, Superedge Degree to identify opinion leaders and Superedge-superedge Distance, and Superedge Overlap to evaluate and verify the results. The aforesaid algorithm ranks the superedges in supernetwork by using the indexes: the influential degree of information dissemination, the transformation likelihood between different psychological types, and the similarity between keywords of viewpoints.

2.7 Hybrid opinion mining framework for e-commerce application

Vinodhini G et al. [16] introduced two frameworks by the combination of classifiers with principal component analysis (PCA) to reduce the dimension of feature set. First, PCA with Bagging which is used to construct each member of the ensemble, and to predict the combination over class labels. Second, Bayesian boosting model is employed using rapid miner tool. For data pre-processing, a word vector representation of review sentences is created for the aforesaid models. Finally, the results proved that the PCA is a suitable dimension reduction method for bagged SVM and Bayesian boosting methods.

2.8 Opinion mining model for ontologies

Isidro Penalver-Martinez et al. [7] presented an innovative method called ontology based opinion mining to improve the feature-based opinion mining by employing the ontologies in selection of features and to provide a new vector analysis-based method for sentiment analysis. The framework composed by four main modules namely, NLP module, Ontology-based feature identification module, polarity identification module, and opinion mining module. In module1, obtains the morphologic and syntactic structure of each sentence by including the pre-processing and POS. In module2, extract the features from the opinions expressed by users. In module3, provides the positive, negative and neutral values of nouns, adjectives and verbs. In module4, the vector analysis enables an effective feature sentiment classification. Each feature is represented using three coordinates. Finally, the results obtained in the movie review domain.

2.9 Sentiment extraction and change detection

Alvaro Ortigosa et al. [2] introduced a new method is called sentiment extraction and change detection. The method includes the operations for extracting sentiments from texts are pre-processing, segmentation into sentences, tokenization, emotion detection, interjection detection, token score assignation (building the lexicon, removing the repetitive letters, spell checking), syntactical analysis, polarity calculation, and for sentiment change detection are building the user regular patter and then comparing weeks. In this aspect, a Facebook application is implemented in SentBuk, which obtains the data from Facebook with the following permissions: Offline_access, Read_stream, and User_about_me and using the user interface performs user regular pattern and combining weeks for classification.

2.10 Semi-supervised Laplacian Eigenmap (SS-LE)

Kyoungok Kim et al. [9] presented SS-LE to reduce the dimensionality of the data points. The experimental process is taken into text cleaning, text refinement, vectorization, applying PCA, and applying SS-LE to reduce the dimension to 2 or 3. SS-LE constructs the graph by utilizing label information and without label information. From this two graphs, graph laplacian matrices are calculated separately, and then dimensionality reduction process used to minimize the distance between two data points and its neighbors by weights.

2.11 Three ensemble methods with five learners in Facebook application

Gang Wang et al. [6] conducted the comparative assessment to measure the performance of three ensemble methods i.e., Bagging, Boosting, and Random Subspace with five learners: NB which is a simple probabilistic classification method, MaxEnt doesn't make any assumptions in relations between features, Decision Tree is a sequence model for a sequence of simple tests, K-Nearest Neighbor classifies majority of its vote of its neighbors, and SVM has ability to model non-linearity. For the above mentioned three ensemble methods, the base learners are constructed from the training data set using random independent, weighted versions, and random subspaces of the feature space respectively.

2.12 VIKOR and Sentiment Analysis framework

Daekook Kang et al. [4] presented a new framework in two stages by combining the VIKOR approach and sentiment analysis for measurement of customer satisfaction in mobile services. VIKOR is a compromising ranking method for MCDM. First, data collection and pre-processing stages involves into the operations like preprocessing data from relevant website. Second, compiling dictionaries of service attributes and sentiment words, it combines the dictionary of attributes and dictionary of sentiment words, and then expressed in verb phrases, adjective phrases and adverbial phrases for sentiment words into positive, negative and neural polarity, at the last assigns score with WordNet. Third, the constructed keyword vectors of customer's opinions with reference to the dictionaries. Fourth, the customer satisfaction is measured with respect to each service attribute. Finally, evaluates the customer satisfaction by considering all attributes.

2.13 Fuzzy deep belief network

Shusen Zhou et al. [14] constructed a two step semi-supervised learning method for the sentiment classification. In this sense, the general DBN is trained by using abundant unlabeled and labeled reviews, design a fuzzy membership function for each class of reviews, and then DBN maps each review into the output space for constructing the FDBN. In the first step, all unlabeled and labeled reviews trained by using the general DBN and estimate the parameters based on mapping results of all reviews. In second step, all unlabeled and labeled reviews trained based on membership functions. Also, AFD proposed by combining active learning with FDBN for labeling and uses them for training.

2.14 Construction of constrained domain-specific sentiment lexicon

Sheng Huang et al. [13] proposed an automatic construction strategy of domain specific sentiment lexicon based on constrained label propagation. In this method, each steps presented sequentially as follows. In step1, sentiment term extraction, which is used to detect and extract candidate term from the corpus. In this step, nouns and noun phrases are expressing objective states, adjectives,

verbs and their phrases are used for reviewed objects, adverbs and their phrases are used to enhance or weaken the adjectives and verbs opinions. In step2, sentiments seeds extraction, which maintains consistent sentiment polarities across multiple domains and it extracted from semi-structured format i.e., Title, Pros, Cons, and Text. It describes rated aspect, positive and negative aspect (nouns and adjective-noun phrases) respectively. In step3, association similarity graph construction, which constructs similarity graph to propagate sentiment information. In step4, constraints definition and extraction focused two types of constraints called contextual constraint and morphological constraint. In contextual constraint, coherence or incoherence relations are used to maintain sentiment polarity directly or reversely. In morphological constraint, coherence or incoherence relations between sentiment terms and also maintains sentiment polarity directly or reversely. In step5, constraint propagation defines a matrix and encodes the direct and reverse constraints into pair-wise constraints. In step6, constrained label propagation in which each sentiment term receives sentiment information from its neighbors and retains its initial polarity label.

Table 1. Main characteristics of OMSA approach published in 2014.

Ref. No.	OMSA approach year	Discipline	Model	Task	Method (M) / Technique (T)	Algorithm (A)/Equation(E)/Frame (F)
[2]	Alvaro Ortigosa et al., 2014	Machine Learning	Descriptive	Classification	M: SentBuk	E: Weekly user vector, Euclidean distance vector, Distance between weekly profiles
[3]	Arturo Montejo-Kaer et al., 2014	Machine Learning	Descriptive	Classification	M: Personalized Page Rank Vectors (PPVs)	A: Random walk algorithm, E: final estimate
[4]	Daekook Kang et al., 2014	Machine Learning	Descriptive	Classification	M: VIKOR and Sentiment Analysis	F: Data collection and preprocessing, and measurement of customer satisfaction
[5]	Farhan Hassan Khan et al. [5], 2014	Machine Learning	Descriptive	Classification	M: EEC, IPC & SWNC	A: Polarity Classification Algorithm
[6]	Gang Wang et al., 2014	Machine Learning	Descriptive	Classification	M: Bagging, Boosting, and Random Subspace, NB, MaxEnt, Decision Tree, K-Nearest neighbor, and SVM	A: Bagging Algorithm, AdaBoost algorithm, Random Subspace algorithm
[7]	Isidro Penabaz-Martinez et al., 2014	Machine Learning	Descriptive	Classification	M: N-Gram	E: Neutral Sense
[8]	Jun Ma et al., 2014	Machine Learning	Descriptive	Similarity	T: Hierarchical clustering M: TLSM (Assessment level, Problem level and criterion level)	A: Gradual Aggregation Algorithm (GAA)
[9]	Kyoungok Kim et al., 2014	Machine Learning	Descriptive	Classification	M: Semi-supervised dimensionality reduction	A: Semi-supervised Laplacian eigenmaps,
[10]	Malhar Anjaria et al., 2014	Machine Learning	Predictive	Classification	M: NB, MaxEnt, SVM, ANN, SVM with PCA	A: Predicting outcome of an Electoral Poll with Modified extended features, and feed forward network training by back propagation learning
[11]	Ning Ma et al., 2014	Machine Learning	Descriptive	Classification	T: Correlation analysis among superedges	A: SuperedgeRank algorithm
[13]	Shang Huang et al., 2014	Machine Learning	Descriptive	Classification	T: Association similarity graph construction	A: Constraint propagation algorithm
[14]	Simson Zhou et al., 2014	Machine Learning	Descriptive	Classification	M: Fuzzy deep belief networks (FDBN)	A: FDBN Algorithm, AFD Algorithm
[15]	Tapia-Rosero A et al., 2014	Machine Learning	Descriptive	Similarity	M: Symbolic notation and Similarity measure	A: getShapeString Algorithm and getFeatureString Algorithm
[16]	Vinodhini G et al., 2014	Machine Learning, Statistical	Predictive	Classification	T: PCA with Bagging and PCA with Boosting	A: Bayesian Boosting algorithm
[17]	Xiaolin Zheng et al., 2014	Machine Learning, Probability	Descriptive	Classification	M: Nearest method and Mutual information method	E: Topic assignments for all sentence F: Dependency graph of sentence

2.15 Ranked WordNet graph for sentiment polarity

Arturo Montejo-Raez et al. [3] employed a method for sentiment classification by using weights of WordNet graph. In this method, the polarity of measurement consider in the interval $[-1, 1]$ and defines a function where values over zero refers positive polarity, values below zero refers negative polarity and values to closer to zero refers neutral. This function is computed by expanding senses and final estimation. Expanding senses intends to expand few concepts in order to calculate the global polarity of the tweet by using the graph of WordNet according to random walk algorithm. The final estimation i.e., final polarity score is evaluated by the combination of SentiWordNet score and random walk weights.

3. EVALUATION RESULTS OF OMSA APPROACH

The OMSA approaches are demonstrated in particular resources with different datasets and its volume as stated in Table 2. Also, the evaluated classification performance was analyzed by using the feature selection process with the different types of metrics such as confusion matrices, precision, recall and F-measure, etc., and their key findings in all the OMSA approach as shown in Table 3. In this paper, we have used the Polarity Classification Algorithm (PCA) and evaluation procedure to verify the accuracy for the above mentioned approach by using the Sanders-Twitter Sentiment Corpus [18]. The corpus contained 5513 hand-classified tweets which are focused on the topic of the companies (Apple, Google, Microsoft and Twitter) and products. The tweet sentiments are labeled as positive, neutral, negative and irrelevant. The datasets count is given in Table 4. These datasets has been measured with 43 trained tweets by using the confusion matrices (Table 5), precision, recall, F-measure and accuracy. The results are shown in Table 6 and Figure 2. In this analysis, the overall system accuracy is only considered and shown.

Table 4. Dataset count

Datasets	No. of. Tweets
Apple	1313
Google	1381
Microsoft	1415
Twitter	1404

Table 5. Confusion matrix

	P	Q	R	S
P	tpP	ePQ	ePR	ePS
Q	eQP	tpQ	eQR	eQS
R	eRP	eRQ	tpR	eRS
S	eSP	eSQ	eSR	tpS

Precision $P = tpP / (tpP + eQP + eRP + eSP)$, Recall $P = tpP / (tpP + ePQ + ePR + ePS)$, F-measure $= 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$, Accuracy $= (\text{True Positive} + \text{True Negative} + \text{True Neutrals} + \text{True irrelevant}) / (\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative} + \text{True Neutrals} + \text{False Neutrals} + \text{True Irrelevant} + \text{False Irrelevant})$. Based on this results and observations, the accuracy level of the OMSA approach remains the same for the resources.

Table 2. Results of OMSA approach with datasets and key findings

Ref. No.	OMSA approach year	Resources	Datasets	Volume	Key Findings
[2]	Alvaro Ortigosa et al., 2014	Facebook	Facebook	1000 for each class (positive, negative or neutral)	Users' sentiment polarity extracted and modeled to detect emotional changes
[3]	Arturo Montejó-Ras et al., 2014	Twitter	-	376,296 tweets - 181,492 positive, 194,804 negative	Sentiment polarity classification in Twitter posts by using weighted nodes and WordNet graph
[4]	Daekook Kang et al., 2014	App Store – AppStore HQ	Bump, Facebook, Foursquare, Fring, Google+, Skype, Twitter, Word Press	1487 Reviews from 8 mobile app services, 166, 303, 144, 112, 152, 231, 271, 108	Customer satisfaction measured for mobile services by combining VIKOR and Sentiment Analysis.
[5]	Farhan Hassan Khan et al., 2014	Twitter	Twitter	6 (2116 – Random Tweets)	Improved the accuracy of classification using Twitter Opinion Mining (TOM) framework
[6]	Gang Wang et al., 2014	10 public S.A datasets	Camera, Camp, Doctor, Drug, Laptop, Lawyer, Movie, Music, Radio, TV	250:248, 402:402, 739:739, 401:401, 88:88, 110:110, 1000:1000, 291:291, 502:502, 235:235	Comparative assessment conducted for three ensemble methods with five learners
[7]	Isidro Penálver-Martínez et al., 2014	Internet messages	Movie Ontology	Movie reviews	Ontology based opinion mining
[8]	Jun Ma et al., 2014	Social Policy, Energy Policy	-	6 Social actors for 7 possible policies 3 Energy policies for 6 domain experts	Developed a gradual aggregation algorithm and Similarity measured at assessment level, problem level and criterion level Redundant features removed by applying Semi-supervised Laplacian eigenmap (SS-LE) and enabled visualization of documents
[9]	Kyoungok Kim et al., 2014	Internet messages	Book, DVD, Electronics, and Kitchen	1000 positive and 1000 negative reviews for each domain	Outcomes of an election result predicted based on User influence factor and extracted opinions using direct and indirect feature
[10]	Malhar Anjaria et al., 2014	Twitter	US Presidential Election 2012 result Tweets, Karnataka State Assembly Election 2013 result tweets	25000	Opinion leaders identified in supernetwork analysis
[11]	Ning Ma et al., 2014	Tianya Club	Japan's nuclear leakage accident	1019 posts by 671 participants	Automatic construction strategy for domain-specific sentiment level lexicon
[13]	Sheng Huang et al., 2014	Epinions.com TripAdvisor and Edmunds.	Cars & Hotels	972,988 & 42,288, 254,539	Constructed FDBN architecture for improved classification performance
[14]	Shusen Zhou et al., 2014	Internet messages	MOV, BOO, DVD, ELE, KIT	1000 positive and 1000 negative reviews for each dataset	Identified similarity shaped membership functions from the large group of experts
[15]	Tapia-Rosero A et al., 2014	Symbolic Notations	Shape - symbolic notation	120	PCA with Bagging, PCA with Boosting
[16]	Vinodhini G et al., 2014	Internet messages	Product reviews	-	Appraisal Expression Patterns (AEP) extracted to represent syntactic relations. Aspect and sentiment words jointly identified by using AEP-LDA model
[17]	Xiaolin Zheng et al., 2014	TripAdvisor, Amazon.com	Restaurant, Hotel, MP3, Camera	2000 (from each domain)	

Table 6. Dataset calculation for confusion matrices, precision, recall and accuracy

Data Sets		Confusion Matrices				Results			
		Positive	Negative	Neutral	Irrelevant	Precision (%)	Recall (%)	F-measure (%)	Accuracy (%)
Apple	Positive	182	3	6	7	94.79	91.92	93.33	96.73
	Negative	7	367	5	3	97.09	96.07	96.58	
	Neutral	2	5	567	1	98.10	98.61	98.35	
	Irrelevant	1	3	0	154	93.33	97.47	95.36	
Google	Positive	209	3	6	7	95.43	92.89	94.14	96.89
	Negative	7	51	5	3	82.26	77.27	79.69	
	Neutral	2	5	590	1	98.17	98.66	98.42	
	Irrelevant	1	3	0	488	97.80	99.19	98.49	
Microsoft	Positive	84	3	6	7	89.36	84.00	86.60	96.96
	Negative	7	128	5	3	92.09	89.51	90.78	
	Neutral	2	5	657	1	98.35	98.80	98.57	
	Irrelevant	1	3	0	503	97.86	99.21	98.53	
Twitter	Positive	65	3	6	7	86.67	80.25	83.33	96.93
	Negative	7	68	5	3	86.08	81.93	83.95	
	Neutral	2	5	627	1	98.28	98.74	98.51	
	Irrelevant	1	3	0	601	98.20	99.34	98.77	

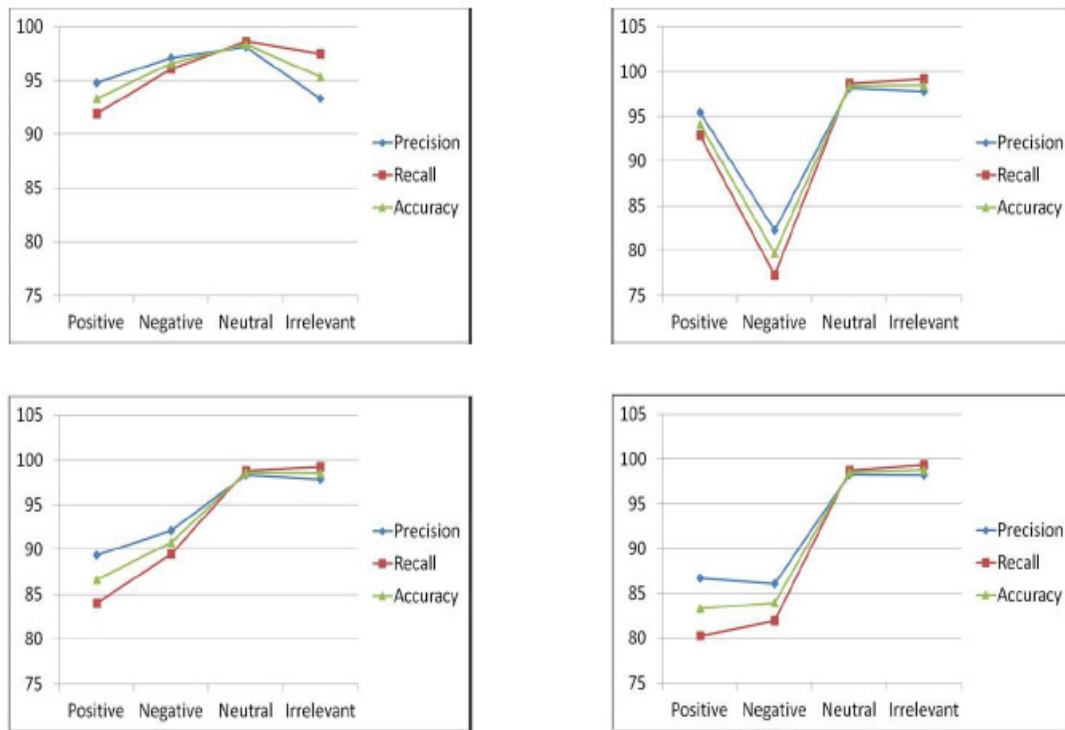


Figure 2. Comparing precision, recall, F-measure and accuracy with four different datasets

4. DISCUSSED CHALLENGES AND FUTURE DEVELOPMENTS IN OMSA

The challenges and future developments are discussed below for the above presented frameworks of OMSA approach published in 2014. It is one of the important tools to the aspirant researchers to focus on new innovative idea. Farhan Hassan Khan et al. [5] indicated that TOM framework faced challenges due to their short length and irregular structure of the content such as named entity recognition, anaphora resolution, parsing, sarcasm, sparsity, abbreviations, poor spellings, punctuation and grammar, incomplete sentences. Also, suggested to compare the proposed algorithm with TweetFeel and Sentiment 140 for further improvement of the accuracy. Malhar Anjaria et al. [10] mentioned that the presence of all entities in unbiased and equal manner was the biggest challenge to provide the accuracy in the standard election prediction model. But, there is a chance of increasing the accuracy level in future by including the more influential factors as age, educational background, employment, economic criterion, rural and urban and social development index.

Tapia-Rosero A et al. [15] discussed that the group decision making process might be difficult under a supervision of mediator. This work could be extended to final objectives i.e., strategic planning, suitability analysis, and applications like fuzzy control, fuzzy time series to find the similarities. Jun ma et al. [8] indicated the major issues that to reduce the risk of putting an inappropriate decision making and measuring opinion similarity between the participants. In GAA, the integrating information according to group of inputs and missing value and unclear answers need to be studied in future. Xiaolin Zheng et al. [17] stated that to jointly identify aspect and sentiment word with comparison of other models. In future work, AEP-LDA model to assume at single sentence and extend at clause level and into aspect-based review summarization, sentiment classification, and personalized recommendation systems.

Ning Ma et al. [11] discussed that online public opinion controlled by deleting long existing posts with rumor from negative opinion leaders. After identifying the opinion leaders, the further work should be focused on how to implement corresponding guidance and interference. Vinodhini G et al. [16] misclassification is reduced, and the classification accuracy of negative opinion to be improved. Isidro Penalver-Martinez et al. [7] addressed the present challenges that ontology-based feature identification, and feature polarity identification. Alvaro Ortigosa et al. [2] demonstrated to extract information from user messages and detect emotional changes. Further, tests to be conducted to determine the values in each case for sentiment changes, and the threshold to distinguish between small changes on user sentiment and significant changes.

Kyoungok Kim et al. [9] addressed the dimensionality reduction transformations into 2D or 3D by using term frequency matrices, and further work suggested that the weight of the edges in the label graph will be adjusted by using the sophisticated approach and to transform document into term frequency. Gang Wang et al. [6] evaluated the ensemble methods, and specified that large datasets need to be collected for validating the result in future i.e., improving the interpretability of ensembles is an important research direction. Daekook Kang et al. [4] indicated that the measurement of the customer satisfaction was conducted by surveys. It's taken more time and effort to collect useful information. Further studies suggested that to incorporate more advanced techniques of sentiment analysis and validate the empirical results presented.

Shusen Zhou et al. [14] discussed the issues that embed the fuzzy knowledge to improve the performance of semi-supervise based sentiment classification. Sheng Huang et al. [13] incorporated the contextual and morphological constraints between sentiment terms, and suggested the future works that incorporate more types of constraints knowledge between sentiment terms and to distinguish the aspect-specific polarities. Arturo Montejo-Raez et al. [3]

stated that how to deal with negation, and to study the context of a specific tweet among the time line of tweets from the particular user in order to identify publisher's mood and adjust final score.

Table 3. Performance Analysis of OMSA approach using feature extraction

Ref. No.	OMSA approach year	Performance Analysis by using feature extraction	
		Types of Metrics	Accuracy (%)
[2]	Alvaro Ortigosa et al., 2014	Evaluating lexicon-based approach with all the messages, Evaluating lexicon-based approach with status messages, Evaluating machine learning approaches and hybrid solutions	83.27
[3]	Arturo Montejó-Raez et al., 2014	Using Support Vector Machine (RW-SVM), Precision, Recall, F1	0.6429 0.6147 0.6285
[4]	Daekeok Kang et al., 2014	Maximum group utility and minimum individual regret	-
[5]	Farhan Hassan Khan et al., 2014	Confusion Matrices, Precision, Recall, F-Measure	87.5
[6]	Gang Wang et al., 2014	Average accuracy by confusion matrix, 10-fold cross validation, Five used base learners, NB – Simple Probabilistic classification method, Violated by real-world data, ME, DT, KNN, SVM, Unigram, Bigram, Term frequency and TF-IDF	RV-SVM
[7]	Isidro Penalver-Martínez et al., 2014	N_GRAM Before, N_GRAM After, N_GRAM Around, and All Phrase	-
[8]	Jun Ma et al., 2014	Similarity Matrix	-
[9]	Kyoungok Kim et al., 2014	Sentiment visualization – 2D scatter plots compared by 6 dimensionality reduction (PCA, Isomap, t-SNE, LE, SS-MMC, and SS-LE, Sentiment Classification – predicted accuracy of SVM and k-NN using seven dimensional reduction methods including SS-LE.	20 slightly better than SVM
[10]	Malhar Anjaria et al., 2014	Effects of Data Cleansing, Hybrid Model of Unigram and Bigram US-E, Gender Based Vote, Vote Shared Based KSAE, Sentiment Accuracy, Comparison for SVM with PCA, Twitter User-USA& India	88
[11]	Ning Ma et al., 2014	Newly evaluation metric formed by Superedge overlap (SO) and Superedge- superedge distance	-
[13]	Sheng Huang et al., 2014	Chi-square based polarity determination, PMI-IR, Label Propagation, Micro – and macro- averaged Precision, Recall, F-measure	-
[14]	Shusen Zhou et al., 2014	Performance of fuzzy deep belief networks, Test accuracy with 100 labeled reviews for active semi-supervised learning, Performance of active fuzzy deep belief networks, Active learning for 5 iterations, Experiments with a different number of labeled reviews, and AFD with AND	-
[15]	Tapia-Rosero A et al., 2014	Similarity matrix with dimensions 120 x 120, values ranging from 0 to 1	-
[16]	Vinodhini G et al., 2014	Compared SVM – Model I with SVM – Model II, SVM, Logistics regression, Bagged SVM, Bayesian boosting	-
[17]	Xiaolin Zheng et al., 2014	Experimental setup - Fleiss kappa, Analysis of appraisal expression pattern, Qualitative evaluation by using F-score, Aspect word identification by using precision, recall, F-score, Sentiment word identification by Nearest Method and MIM Domain adaptation of AEP by using Loss measure	-

Based on this observation, OMSA approach is not following any single algorithm or technique or method for data collection process, pre-processing and classification, and to solve all issues in social media for extracting the user's opinion. Therefore, the problem is defined with some specific application domain to estimate the polarity of the system. The OMSA approach is also dealt with various models as shown in Table 1 such as descriptive, predictive and statistical for the purpose of extracting opinions by using the applications of Fuzzy sets, grouping similarity measure, group decision making process, policy creation process, OSN sites, and symbolic notations. The analysis of result is very useful to overcome issues and to develop various new methods or techniques without duplication.

5. CONCLUSION

Opinion mining and sentiment analysis is emerging as a challenging field as part of text mining in social networks to an organization, Government and public. It has a lot of applications and developments that to predict people's polarity towards their decision making process. In this paper, we conclude that the frameworks and algorithms of OMSA are presented with the data collection process, preprocessing methods, classification and performance evaluation results as a review.

REFERENCES

- [1] Alejandro Pena-Ayala.: Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*. 41, 1432-1462 (2014)
- [2] Alvaro Ortigosa, Jose M. Martin, Rosa M. Carro.: Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*. 31, 527-541 (2014)
- [3] Arturo Montejo-Raez, Eugenio Martinez-Camara, M. Teresa Martin-Valdivia, L. Alfonso Urena-Lopez.: Ranked WordNet graph for Sentiment Polarity Classification in Twitter. *Computer Speech and Language*. 28, 93-107 (2014)
- [4] Daekook Kang, Yongtae Park.: Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications*. 41, 1041-1050 (2014)
- [5] Farhan Hassan Khan, Saba Bashir, Usman Qamar.: TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*. 57, 245-257 (2014)
- [6] Gang Wang, Jianshan Sun, Jian Ma, Kaiquan Xu, Jibao Gu.: Sentiment Classification: The contribution of ensemble learning. *Decision support systems*. 57, 77-93 (2014)
- [7] Isidro Penalver-Martinez, Francisco Garcia-Sanchez, Rafael Valencia-Garcia, Miguel Angel Rodriguez-Garcia, Valentin Moreno, Anabel Fraga, Jose Luis Sanchez-Cervantes.: Feature-based opinion mining through ontologies. *Expert Systems with Applications*. 41, 5995-6008 (2014)
- [8] Jun Ma, Jie Lu, Guangquan Zhang.: A three-level-similarity measuring method of participant opinions in multiple-criteria group decision supports. *Decision Support Systems*. 59, 74-83 (2014)
- [9] Kyoungok Kim, Jaewook Lee.: Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. *Pattern Recognition*. 47, 758-768 (2014)
- [10] Malhar Anjaria & Ram Mohana Reddy Guddeti.: Influence factor based opinion mining of twitter data using supervised learning. *Sixth IEEE International conference on communication systems and networks (COMSNETS)*. ISSN: 1409-5982, (2014)
- [11] Ning Ma, Yijun Liu.: SuperedgeRank algorithm and its application in identifying opinion leader of online public opinion supernetwork. *Expert Systems with Applications*. 41, 1357-1368 (2014)
- [12] R.V. Vidhu Bhala, S. Abirami.: Trends in word sense disambiguation. *Artificial Intelligence Review: An International Science and Engineering Journal*. DOI 10.1007/s10462-012-9331-5, Springer, (2012)
- [13] Sheng Huang, Zhendong Niu, Chongyang Shi.: Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. *Knowledge-Based Systems*. 56, 191-200 (2014)
- [14] Shusen Zhou, Qingcai Chen, Xiaolong Wang.: Fuzzy deep belief networks for semi-supervised sentiment classification. *Neurocomputing*. 131, 312-322 (2014)
- [15] Tapia-Rosero A, A. Bronselaer, G. De Tre.: A method based on shape-similarity for detecting similar opinions in group decision-making. *Information Sciences*. 258, 291-311 (2014)
- [16] Vinodhini G, Chandrasekaran RM.: Measuring quality of hybrid opinion mining model for e-commerce application. *Measurement*. 55, 101-109 (2014)
- [17] Xiaolin Zheng, Zhen Lin, Xiaowei Wang, Kwei-Jay Lin, Meina Song.: Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. *Knowledge-Based Systems*. 61, 29-47 (2014)
- [18] Niek J. Sanders.: Sanders-Twitter Sentiment Corpus, Version 2.0.1

OPTIMAL BUFFER ALLOCATION IN TANDEM CLOSED QUEUING NETWORK WITH MULTI SERVERS USING PSO

K.L.Narasimhamu¹ V.Venugopal Reddy² and C.S.P.Rao³

¹Department of Mechanical Engineering, AITS, Rajampet, INDIA

²Department of Mechanical Engineering, JNTUCE, Pulivendula, INDIA

³Department of Mechanical Engineering, NIT, Warangal, INDIA

¹klsimha@gmail.com, ²vgreddy7@rediffmail.com

ABSTRACT

Buffer Allocation Problem is an important research issue in manufacturing system design. Objective of this paper is to find optimum buffer allocation for closed queuing network with multi servers at each node. Sum of buffers in closed queuing network is constant. Attempt is made to find optimum number of pallets required to maximize throughput of manufacturing system which has pre specified space for allocating pallets. Expanded Mean Value Analysis is used to evaluate the performance of closed queuing network. Particle Swarm Optimization is used as generative technique to optimize the buffer allocation. Numerical experiments are shown to explain effectiveness of procedure.

KEYWORDS

Buffer Allocation Problem, Closed Queuing Network, Expanded Mean Value Analysis, Particle Swarm Optimization

1. INTRODUCTION

Buffer Allocation Problem deals with allocation of optimal buffer slots among intermediate buffer locations of a manufacturing system to achieve a specific objective. The primary reason for having storage buffers is to reduce the idle time because of starving and blocking. Less idle time increases throughput rate of manufacturing system. Buffer needs additional capital investment and floor space. Work-In-Process increases because of buffering in the space available. So total buffer space should be as minimum or with available buffer space total throughput rate should be as maximum as possible. Throughput rate of manufacturing system is a function of service rates of machines and buffer sizes at various machines.

Manufacturing system is combination of machines and queues. Manufacturing system can be shown as network of queues and it can be named as queuing network. A queuing network with constant work-in-process is known as closed queuing network.

2. BUFFER ALLOCATION PROBLEM

BAP is an NP-hard combinatorial optimization problem in design of manufacturing system. In general three types of buffer allocation problem models can be found in the literature.

Model 1: To find optimum buffer allocation in order to maximize throughput rate for a given fixed amount of buffers.

Objective function: Maximize (Throughput rate)

Subject to, Sum of buffers = Total space available.

Model 2: To find optimum buffer allocation in order to minimize total buffer size with desired throughput rate.

Objective function: Minimize (Total buffer size)

Subject to, Throughput rate \geq required (desired) throughput rate.

Model 3: To find optimum buffer allocation in order to minimize Work-in-process inventory with desired throughput rate and total space available.

Objective function: Minimize (Work-in-process inventory)

Subject to, Throughput rate \geq Desired throughput rate

Sum of buffers \leq Total buffer space available.

2.1 General procedure to solve BAP

Generative and evaluative methods can be used in cyclic manner to solve BAP as shown in Figure 1.

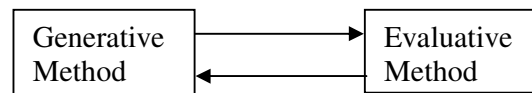


Figure 1. BAP solution process

Evaluative methods are used to obtain the value of objective function. Generative methods are used to search for optimal solution. Evaluative methods can be classified into analytical methods and simulation methods. Further analytical methods can be classified into exact methods and approximate methods. Exact methods are suitable only for small size buffers. Simulation methods are time consuming methods. Generative methods are used to search optimum buffer sizes to optimize system performance. These methods can be classified into traditional and heuristic search algorithms. Sometimes traditional search algorithms cannot jump over local optimum solution in order to find global optimum solution. Meta heuristic methods are strategies to explore search space in order to find optimal or near optimal solutions.

3. LITERATURE REVIEW

Daskalaki and Smith [1] attempted BAP, combined with routing in serial parallel queuing networks. An iterative 2-step method was used to solve BAP and routing problem. Expansion method was used as evaluative method and Powell's algorithm was used as generative method.

Maximization of throughput and minimization of buffer size are the objective functions of this problem.

Smith and Cruz [2] solved BAP for general queuing networks. Minimization of total buffer size is objective function of this problem. The Generalized Expansion Method was used as evaluative method and Powell's algorithm was used as generative technique. Similar work was done by Smith et al. [3] for multi-server queuing networks.

Cruz et al. [4] solved the BAP in an arbitrary queuing networks. Aim was to find minimum buffer size in order to achieve desired throughput. Generalized Expansion Method was used as evaluative method and Lagrangian relaxation method was used as generative method.

Yuzukirmizi et al. [5] considered optimum buffer allocation for closed queuing networks. This is the first procedure to find optimum buffer allocation in closed queuing networks with general topologies and multiple servers. Expanded Mean Value Analysis was used to evaluate throughput of closed queuing network and Powell's algorithm was used to find optimum buffer allocation.

Cruz et al. [6] applied generalized expansion method and multi objective genetic algorithm to optimize throughput and buffer sizes for single server queuing network. Similar work was extended to optimize buffer size, throughput and server rate by Cruz et al. [7].

Soe and Lee [8] presented solutions for tandem queuing networks. Explicit expression was developed as an evaluative method to BAP.

4. MODEL DESCRIPTION

In the present work manufacturing system is considered as a closed queuing network with single server (machine) at each node. Buffer size includes the part being operated on machine. All servers are reliable. Maximization of throughput is objective function subjected to sum of buffers is constant.

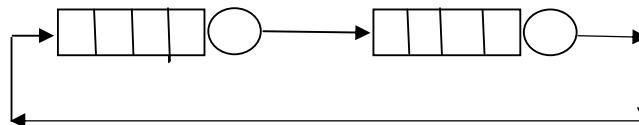


Figure 2. Two node closed queuing network

Two node closed queuing network is shown in Figure 2. Number of components in closed queuing network is constant. Therefore sum of buffers in manufacturing system i.e. work in process of system is constant. Expanded Mean Value Analysis proposed by Yuzukirmizi et al. [5] is used as evaluative method to calculate throughput of manufacturing system which is designed as closed queuing network. Particle Swarm Optimization is used as search method to optimize buffer sizes.

4.1. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is one of the metaheuristic search technique to find optimum solution. It is a nature inspired technique. Social sharing of information is main idea behind the implementation of PSO.

Pseudo code of general PSO algorithm is as follows:

```

Initialize parameters
Initialize population
Evaluate

Do {Find pbest (population best)
    Find gbest (global best)
    Update velocity
    Update position
    Evaluate
} While (Termination)

```

Less complexity and economic computational cost are the main advantages of PSO algorithm.

4.2 Proposed Algorithm

Let

M Number of machines (Number of buffers)
 B Total buffer space
 N Number of particles
 Mu Service rates of machines
 S Number of servers
 P Number of pallets
 Pbest Particle best
 Gbest Global best
 X Buffer size

Cp, Cg Coefficients (Constants)

Step 1. Generate initial population with n number of particles. Each particle with M number of dimensions.

Step 2. Generate initial velocity matrix (n X M) randomly.

Step 3. Calculate throughput rate for each particle using EMVA algorithm as follows.

Step 3.1. for P=1 to B

 Compute throughput rate using Mu,S

Step 3.2. Find maximum throughput and corresponding number of pallets

Step 4. Find particle best and global best.

Step 5. Calculate change in velocity.

 Change in velocity = Cp * rand ()*(Pbest – X) + Cg * rand ()*(Gbest – X).

Step 6. New velocity = old velocity + change in velocity.

Step 7. Prepare new population as follows.

 If new velocity is positive then increase buffer size by one. If new velocity is negative then decrease buffer size by one. Make adjustments to satisfy total buffer size.

Step 8. Assign new velocity to old velocity.

Step 9. If termination condition is satisfied go to step 10. Otherwise go to step 3.

Step 10. Finalize the best buffer allocation and total number of pallets.

5. NUMERICAL EXPERIMENTS AND RESULTS

MATLAB code is written to find maximum throughput, optimum buffer allocation and optimum number of pallets in the queuing network. Total buffer size, number of machines, service rates of machines, Number of servers at each node and number of particles are inputs to the program. By executing this program optimum buffer sizes can be obtained. PSO parameters are considered as follows.

Table 1. PSO parameters.

Parameter	Value
Velocity	[-4,4]
Cp	0.5
Cg	0.5

Experiment 1: Two nodes, $\mu = [9,1]$, Total buffer size= 6, Number of servers = [1,2]

Two machines are considered with service rates [9,1] and total buffer size 6. Different possibilities are verified for throughput calculation using Expanded Mean Value Analysis. Experiments were conducted and tabulated in table 2. Optimum buffer allocation using proposed algorithm is shown in table 3. It is coinciding with maximum throughput buffer allocation of complete enumeration calculation. Similar experiments with total buffer size 8 and 16 are shown tables from 4 to 9.

Experiment 1: Total number of pallets=6, $\mu = [9 \ 1]$, Servers= [1,2]

Table 2. Complete Enumeration for total number of pallets=6, $\mu = [9 \ 1]$, Servers= [1,2]

Number of pallets	Buffer allocation				
	(1,5)	(2,4)	(3,3)	(4,2)	(5,1)
1	0.9	0.9	0.9	0.9	0.9
2	1.7822	1.7822	1.7822	1.7822	1.7822
3	1.946	1.9527	1.9527	1.9527	1.9229
4	1.9734	1.9881	1.9896	1.9829	1.919
5	1.9779	1.9941	1.9959	1.9816	1.91
6	1.9786	1.9947	1.9948	1.9767	1.9008

Table 3. Solution using proposed algorithm for total number of pallets=6, $\mu = [9 \ 1]$, Servers= [1, 2]

Optimum Buffer Allocation	Number of Pallets	Maximum Throughput
(3,3)	5	1.9959

Experiment 2: Total number of pallets=8, $\mu = [1, 2]$, Servers [1, 2]

Table 4. Complete Enumeration for total number of pallets=8, Mu= [1, 2], Servers [1, 2]

Number of pallets	Buffer allocation						
	(1,7)	(2,6)	(3,5)	(4,4)	(5,3)	(6,2)	(7,1)
1	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667
2	0.9231	0.9231	0.9231	0.9231	0.9231	0.9231	0.9231
3	0.9376	0.9811	0.9811	0.9811	0.9811	0.9811	0.9699
4	0.9349	0.9841	0.9953	0.9953	0.9953	0.9925	0.976
5	0.9317	0.9822	0.996	0.9988	0.9981	0.9939	0.9766
6	0.9289	0.9791	0.9955	0.9988	0.9985	0.994	0.9766
7	0.9265	0.9757	0.9938	0.9984	0.9983	0.9939	0.9766
8	0.9245	0.9687	0.9915	0.9979	0.9982	0.9939	0.9766

Table 5. Solution using proposed algorithm for total number of pallets=8, Mu= [1 2], Servers= [1, 2]

Optimum Buffer Allocation	Number of Pallets	Maximum Throughput
(4,4)	6	0.9988

Experiment 3: Total number of pallets=8, Mu= [2, 1], Servers [1, 3]

Table 6. Complete Enumeration for total number of pallets=8, Mu= [2, 1], Servers [1, 3]

Number of pallets	Buffer allocation						
	(1,7)	(2,6)	(3,5)	(4,4)	(5,3)	(6,2)	(7,1)
1	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667
2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
3	1.4851	1.5789	1.5789	1.5789	1.5789	1.5789	1.5152
4	1.5548	1.7042	1.7538	1.7538	1.7538	1.7204	1.5624
5	1.5663	1.7267	1.8197	1.8483	1.8291	1.7395	1.5373
6	1.5667	1.7087	1.8279	1.8749	1.8383	1.7095	1.5207
7	1.5657	1.6826	1.7961	1.8498	1.8044	1.6812	1.5156
8	1.5651	1.6486	1.7362	1.7971	1.7613	1.6611	1.5149

Table 7. Solution using proposed algorithm for total number of pallets=8, Mu= [2 1], Servers= [1, 3]

Optimum Buffer Allocation	Number of Pallets	Maximum Throughput
(4,4)	6	1.8749

Experiment 4: total number of pallets=16, Mu= [0.2, 0.5], Servers [5, 3]

Table 8. Complete Enumeration for total number of pallets=16, Mu= [0.2, 0.5], Servers [5, 3]

Number of pallets	Buffer allocation							
	(1,15)	(2,14)	(3,13)	(4,12)	(5,11)	(6,10)	(7,9)	(8,8)
1	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429
2	0.2857	0.2857	0.2857	0.2857	0.2857	0.2857	0.2857	0.2857
3	0.4162	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286
4	0.5216	0.5667	0.5702	0.5702	0.5702	0.5702	0.5702	0.5702
5	0.594	0.6912	0.7074	0.7086	0.7086	0.7086	0.7086	0.7086
6	0.6256	0.7694	0.8135	0.82	0.8205	0.8205	0.8205	0.8205
7	0.636	0.7887	0.8703	0.8898	0.8928	0.8931	0.8931	0.8931
8	0.6449	0.7713	0.8819	0.9222	0.9317	0.9333	0.9335	0.9335
9	0.6515	0.7484	0.8667	0.9308	0.9512	0.9565	0.9574	0.9575
10	0.6547	0.7293	0.8392	0.9224	0.9571	0.9687	0.9718	0.9724
11	0.656	0.7173	0.8086	0.9025	0.9522	0.9726	0.9795	0.9812
12	0.6566	0.7114	0.7821	0.8749	0.9389	0.9695	0.9818	0.9854
13	0.6568	0.7095	0.7636	0.844	0.9177	0.9599	0.9788	0.9852
14	0.6568	0.7096	0.754	0.8148	0.8887	0.9423	0.9702	0.9808
15	0.6568	0.7104	0.7516	0.7912	0.8531	0.9152	0.9551	0.9726
16	0.6568	0.7111	0.7532	0.7752	0.8158	0.8792	0.9327	0.9608

Number of pallets	Buffer allocation							
	(9,7)	(10,6)	(11,5)	(12,4)	(13,3)	(14,2)	(15,1)	(9,7)
1	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429
2	0.2857	0.2857	0.2857	0.2857	0.2857	0.2857	0.2857	0.2857
3	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286	0.4235	0.4286
4	0.5702	0.5702	0.5702	0.5702	0.5702	0.5683	0.5494	0.5702
5	0.7086	0.7086	0.7086	0.7086	0.7078	0.6996	0.6582	0.7086
6	0.8205	0.8205	0.8205	0.8201	0.8162	0.7952	0.7281	0.8205
7	0.8931	0.8931	0.8929	0.8909	0.8799	0.8416	0.7559	0.8931
8	0.9335	0.9334	0.9322	0.9261	0.9042	0.8504	0.76	0.9335
9	0.9575	0.9568	0.9531	0.94	0.9066	0.8449	0.7589	0.9575
10	0.972	0.9697	0.9616	0.9404	0.8989	0.8379	0.758	0.972
11	0.9802	0.975	0.9615	0.9339	0.89	0.8338	0.7577	0.9802
12	0.9833	0.9748	0.9566	0.926	0.8838	0.8325	0.7576	0.9833
13	0.9824	0.9712	0.9505	0.9196	0.8809	0.8326	0.7576	0.9824
14	0.9786	0.9663	0.945	0.9157	0.8803	0.833	0.7577	0.9786
15	0.9728	0.9613	0.9409	0.914	0.8808	0.8333	0.7577	0.9728
16	0.9659	0.9566	0.9381	0.9136	0.8816	0.8334	0.7577	0.9659

Table 9. Solution using proposed algorithm for total number of pallets=16, Mu= [0.2, 0.5], Servers [5, 3]

Optimum Buffer Allocation	Number of Pallets	Maximum Throughput
(8,8)	12	0.9854

Experiment 5: Total number of pallets=7, Mu= [2, 0.1, 1], Servers [1, 2, 1]

Table 10. Complete Enumeration for total number of pallets=7, Mu= [2, 0.1, 1], Servers [1, 2, 1]

Number of pallets	Buffer allocation							
	(1,1,5)	(1,2,4)	(1,3,3)	(1,4,2)	(1,5,1)	(2,1,4)	(2,2,3)	(2,3,2)
1	0.087	0.087	0.087	0.087	0.087	0.087	0.087	0.087
2	0.1723	0.1723	0.1723	0.1723	0.1723	0.1723	0.1723	0.1723
3	0.1933	0.1938	0.1938	0.1938	0.1934	0.1933	0.1938	0.1938
4	0.1974	0.1986	0.1987	0.1986	0.1976	0.1974	0.1986	0.1986
5	0.1982	0.1996	0.1997	0.1995	0.1983	0.1982	0.1996	0.1995
6	0.1982	0.1998	0.1999	0.1997	0.1984	0.1982	0.1998	0.1996
7	0.1981	0.1998	0.1999	0.1997	0.1984	0.1981	0.1998	0.1997

Number of pallets	Buffer allocation						
	(2,4,1)	(3,1,3)	(3,2,2)	(3,3,1)	(4,1,2)	(4,2,1)	(5,1,1)
1	0.087	0.087	0.087	0.087	0.087	0.087	0.087
2	0.1723	0.1723	0.1723	0.1723	0.1723	0.1723	0.1723
3	0.1934	0.1933	0.1938	0.1934	0.1933	0.1934	0.1929
4	0.1976	0.1974	0.1986	0.1976	0.1974	0.1976	0.1964
5	0.1983	0.1982	0.1994	0.1983	0.198	0.1982	0.1968
6	0.1984	0.1982	0.1995	0.1984	0.198	0.1983	0.1968
7	0.1984	0.1981	0.1995	0.1984	0.1979	0.1983	0.1967

Table 11. Solution using proposed algorithm for total number of pallets=7, Mu= [2, 0.1, 1], Servers [1, 2, 1]

Optimum Buffer Allocation	Number of Pallets	Maximum Throughput
(1, 3, 3)	7	0.1999

Experiment 6: Experiments were conducted for 3 node closed queuing network with various total buffer sizes. Results are shown in table 12.

Table 12. Optimum solutions for 3 node network

Total buffer space	Service rates	Number of servers	Optimum Buffer Allocation	Optimum number of pallets	Maximum Throughput
12	(0.3333,1,1)	(3,1,1)	(4,4,4)	10	0.7743
15	(0.3333,0.5,0.3333)	(3,2,3)	(5,5,5)	13	0.8047
15	(1,1.5,3)	(2,2,1)	(6,6,3)	11	1.9269

Experiment 7: Experiments were conducted for 5 node closed queuing network with various total buffer sizes. Results are shown in table 13.

Table 13. Optimum solutions for 5 node network

Total buffer space	Service rates	Number of servers	Optimum Buffer Allocation	Optimum number of pallets	Maximum Throughput
25	(0.8,0.8,0.8,0.8,0.8)	(3,2,1,2,3)	(1,1,10,11,2)	20	0.7998
33	(4,1,3,2,1.5)	(1,5,2,2,3)	(8,9,3,4,9)	31	3.6758

Experiment 8: Experiments were conducted for 8 node closed queuing network with various total buffer sizes. Results are shown in table 14.

Table 14. Optimum solutions for 8 node network

Total buffer space	Service rates	Number of servers	Optimum Buffer Allocation	Optimum number of pallets	Maximum Throughput
40	(1,0.5,1,0.5,1,0.5,1,0.5)	(1,3,1,3,1,3,1,3)	(1,1,5,7,5,7,6,8)	26	0.7916
50	(1.2,1,.8,1.2,1,.8,1.2,1)	(1,2,3,1,2,3,1,2)	(6,12,1,4,11,2,5,9)	50	1.1434

Present work is focused on multi server reliable machines. Work can be extended to solve merge, split, unreliable systems. Extension of this work is under progress by the authors.

REFERENCES

- [1] Daskalaki, S., & Smith, J. M. (2004). Combining routing and buffer allocation problems in serial-parallel queuing networks. *Annals of Operations Research*, 125, 47–68.
- [2] Smith, J. M., & Cruz, F. R. B. (2005). The buffer allocation problem for general finite buffer queuing networks. *IIE Transactions*, 37(4), 343–365.
- [3] Smith, J. M., Cruz, F. R. B., & Van Woensel, T. (2010). Topological networks design of general, finite, multi-server queuing networks. *European Journal of Operational Research*, 201(2), 427–441.
- [4] Cruz, F. R. B., Duarte, A. R., & Van Woensel, T. (2008). Buffer Allocation in general single-server queuing networks. *Computers, Operations Research* 35(11), 3581-3598.
- [5] Yuzukirmizi, M., & Smith, J. M. (2008). Optimal buffer allocation in finite closed networks with multiple servers. *Computers, Operations Research*, 35, 2579-2598.
- [6] Cruz, F. R. B., Van Woensel, T., & Smith, J. M. (2010). Buffer and throughput trade-offs in M/G/1/K queuing networks: A bicriteria approach. *International Journal of Production Economics*, 125, 224–234.
- [7] Cruz, F. R. B., Kendall, G., While, L., Duarte, A. R., Brito, N.L.C.(2012). Throughput maximization of queuing networks with simultaneous minimization of service rates and buffers. *Mathematical Problems in Engineering*. Volume 2012, Article ID 692593.
- [8] Seo, D-W., & Lee, H. (2011). Stationary waiting times in m-node tandem queues with production blocking. *IEEE Transactions on Automatic Control*, 56(4), 958-961.

INTENTIONAL BLANK

AN IMPROVED TEACHING-LEARNING BASED OPTIMIZATION APPROACH FOR FUZZY CLUSTERING

Parastou Shahsamandi E.¹ and Soheil Sadi-nezhad²

^{1,2}Department of Industrial Engineering, Science & Research Branch,
Islamic Azad University, Tehran, Iran

¹p.shahsamandi@mau.ac.ir, ²sadinejad@hotmail.com

ABSTRACT

Fuzzy clustering has been widely studied and applied in a variety of key areas of science and engineering. In this paper the Improved Teaching Learning Based Optimization (ITLBO) algorithm is used for data clustering, in which the objects in the same cluster are similar. This algorithm has been tested on several datasets and compared with some other popular algorithm in clustering. Results have been shown that the proposed method improves the output of clustering and can be efficiently used for fuzzy clustering.

KEYWORDS

Meta-heuristic algorithm, Improved teaching-learning-based optimization, Fuzzy clustering

1. INTRODUCTION

Data clustering is an important problem in data mining and knowledge discovery. The main objective of any clustering technique is grouping a set of objects in a number of clusters; in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct.[1,2] One of the proper measures of similarity for data in K cluster is the distance between data and their cluster center (e.g. the Euclidean distance in fuzzy c-means algorithm proposed by [3]). In fact this unsupervised classification produces a $K \times P$ optimum partition matrix $U^*(x)$ of the given dataset X , consisting of m data samples, $X = \{x_1, x_2, x_3, \dots, x_m\}$, where each X_i in the universe X is an p -dimensional vector of m elements or m features; and $i=1, 2, \dots, m$. The partition matrix can be represented as $U = \{u_{ki}\}$, $K=1, 2, \dots, k$, and u_{ki} is the membership of object X_i to k th cluster. For fuzzy clustering of data, $0 < u_{ki} < 1$, (i.e. u_{ki} denotes the degree of belonging of object X_i to the k th cluster.)

Finding the optimum matrix, U^* , is difficult for practical problems, so the application of advanced optimization techniques is required. Considering that, the clustering problem is NP-hard problem (when the number of data and number of cluster are exceeded), the application of metaheuristic method is necessary for partitioning data. [4]

The metaheuristic algorithms can be classified into different groups depending on the criteria being considered; the evolutionary algorithms (ES), (such as Genetic algorithm, Differential

Evolution) and Swarm intelligence algorithm (such as Particle Swarm Optimization, Ant Colony Optimization, and Artificial bee colony) are based on population criteria. Beside these algorithms,[5] there are some other algorithms which work on the principles of different natural phenomena, such as Harmony Search [6], Gravitational Search algorithm,[7] Teaching-Learning Based optimization. [8,9]

The metaheuristic algorithms can solve large problem faster and can obtain robust algorithms, more over these algorithms are simple to design and implement [4]. Many of these algorithms have been introduced to solve clustering problems. [10-12] have been used Genetic Algorithm(GA) to optimizing a specified objective function related to solve the clustering problem. [13,14] proposed a Tabu Search based heuristic for clustering. [15] considered the problem of clustering and Simulated Annealing approach has been proposed for solving problem. [16] presented an Ant Colony clustering algorithm, which simulate the way real ants find the shortest path from their nest to food source and back. They compared the performance of this algorithm with other algorithms. [17] hybridized Particle Swarm Optimization with K-means and Nelder-Mead simplex search method to improve the performance of algorithm for clustering problem.[18] presented an Artificial Bee Colony to optimally solve the clustering problem. The performance of this algorithm has been compared with other popular heuristics algorithm in clustering. [8,9] implemented Teaching-Learning-Based Optimization (TLBO) for automatic clustering and named it Auto-TLBO. This algorithm is evaluated on benchmark datasets and performance comparisons are made with some well-known clustering algorithms.

Different algorithm requires common controlling parameters such as population size, number of generations and its own algorithm specific control parameters (e.g. mutation rate and cross over rate in GA algorithm; inertia weight and social and cognitive parameters in PSO). Among of all algorithms, [8,9] showed that TLBO algorithm does not requires any algorithm-specific parameters. TLBO algorithm simulates the teaching-learning phenomenon of a classroom, where a group of learners are considered the population and different subjects offered to the learners are similar to different design variables of optimization problem. The best solution in the entire population is considered as teacher.

This algorithm has been improved by introducing more than one teacher for learners (i.e. increased the collective knowledge) and some other modifications. [19]

Thereby, in this paper the Improved TLBO is proposed for fuzzy clustering problem; the objective function of fuzzy c-means algorithm is used as fitness function and Euclidian distance metric as a distance metric. The minimum amount of this objective shows the better clustering. Clustering results are reported for a number of real-life and artificial datasets. The performance of algorithm compared with several other proposed clustering algorithms. This paper is organized as follows: the next section discussed the fuzzy c-mean algorithm. In section 3, the improved TLBO algorithm for data clustering is described. Section 4 presents the experimental results conducted on several data sets .Finally, section 5 concludes the article.

2. FUZZY CLUSTERING ALGORITHM

Let $X_{m \times p}$ be the profile data matrix, with m rows (set of m objects) and p columns (p -dimensional), where each X_{ij} corresponds to the j th real value feature ($j=1, 2, \dots, p$) of i th object ($i=1, 2, \dots, m$). Given $X_{m \times p}$ the goal of partitioned clustering algorithm is to find grouping or structures; such that objects which are assigned to the same cluster should be similar (i.e. Homogeneity), while objects which are assigned to different clusters should be different (i.e. Heterogeneity).

In most cases the data is in the form of real value vector. The Euclidean distance is a suitable measure of similarity for these datasets. The Euclidean id derived from Minkowski metric.(Eq.1)

$$d(x, y) = (\sum_{i=1}^p |x_i - y_i|^r)^{\frac{1}{r}} \xrightarrow{r=2} d(x, y) = (\sum_{i=1}^p |x_i - y_i|^2)^{\frac{1}{2}} \quad \text{Eq.1}$$

Fuzzy c-mean (FCM) is a widely used technique which allows a datum to belong to more than one cluster. (Eq.2), It is based on minimization of the following measure:

$$J_m = \sum_{i=1}^p \sum_{k=1}^K u_{ij}^{m'} d_{ij} \quad \text{Eq.2}$$

Where p is the number of data objects, K represent number of clusters, u is the fuzzy membership matrix; m' ($m' > 1$) is the weighting exponent and controls the fuzziness of resulting clusters and d_{ij} is euclidian distance from data X_i to cluster center. This criterion is based on the compactness of data in clusters.[3]

3. ITLBO ALGORITHM BASED FUZZY CLUSTERING

TLBO algorithm simulates the teaching-learning process that every individual tries to learn something from other individual to improve themselves. The algorithm simulates two fundamental modes of learning: Teacher phase and Learner phase. A group of learner is considered the population of algorithm and the results of learner are the fitness value of optimization problem, which indicates its quality. [8,9]

In teacher phase, the learning of the learner through teacher is simulated. During this phase, a teacher conveys knowledge among the learners and makes an effort to increase the mean result of the class. At any iteration of algorithm, there are n number of learners (population size) and m number of subjects. $M_{j,i}$ be the mean result of learners iv subject j th. The best overall result $X_{total-kbest,i}$ is the result of the best learner. A teacher is the most experience person (the best learner) in the algorithm. The difference between the result of the teacher and the mean result of the learner in each subject in given by (Eq.3),

$$Difference_Mean_{j,i} = r_i (X_{j,kbest,i} - T_F M_{j,i}) \quad \text{Eq.3}$$

Where $X_{j,kbest,i}$ is the result of the teacher (best learner) in subject j . T_F is the teaching factor which decides the value of mean to be changed, and r_i is the random number in range $[0,1]$. Based on the $Difference_Mean_{j,i}$, the existing solution is updated in the teacher phase according to the following expression:

$$X'_{j,k,i} = X_{j,k,i} + Difference_Mean_{j,i}$$

Where $X'_{j,k,i}$ is the update value; accept it if it gives a better function value. These accepted values become the input to the learner phase.

The learner phase of the algorithm simulates the learning of the learners; through interaction among themselves. The learners can also gain knowledge by interacting with other learners. A learner interacts randomly with other learners and learns new things if the other learner has more knowledge than him or her. Randomly select two learners P and Q , such that

$X'_{total-P,i} \neq X'_{total-Q,i}$, (where these values are the updated value at the end of teacher phase). The following equations are for maximization problem.

$$X''_{j,P,i} = X'_{j,P,i} + r_i (X'_{j,P,i} - X'_{j,Q,i}) \quad , \\ \text{if } X'_{total-P,i} > X'_{total-Q,i}$$

$$X''_{j,P,i} = X'_{j,P,i} + r_i (X'_{j,Q,i} - X'_{j,P,i}) \quad , \\ \text{if } X'_{total-Q,i} > X'_{total-P,i}$$

Accept $X''_{j,P,i}$, if it gives a better function value.

The algorithm stops with the criteria such as the maximum iteration and the minimum change of objective function.

[19] improved the algorithm by introducing more than one teacher for learners and some modifications to adaptive teaching factor and self motivated learning; named it ITLBO. In this paper the ITBO proposed for clustering.

Based on the mentioned statements, the steps of ITLBO algorithm with detailed description of each are as follows:

Step 1: objective function; define the optimization problem as minimizing overall deviation of a partitioning or maximize compactness inside the clusters. This is simply computed as the overall summed distances between data items and their corresponding cluster center; (i.e. the objective function of fuzzy c- mean algorithm; Eq.2) the weighting exponent m' in Eq.2 is set to 2, which is a common choice for fuzzy clustering. Considering that, TLBO algorithm does not require any algorithm-specific parameter; so, setting the control parameter value is not necessary.

Step 2: Initialization; Initialize the population (N learner), in order to solve clustering, each candidate solution in the population consist of N number $U_{k \times m}$ matrix, where each elements of this matrix represents the degree of belonging objects to k th clusters. The fuzzy matrix U is generated randomly according to population size, then the center of each cluster is compute to find the distance between each data and the centroids of clusters (as Eq.2). In the experiment, we set the population size or the number of learner as 100.

Step 3: Evaluation; evaluate the population, then select and assign the best solution as chief teacher to first rank (i.e. $f(x)_{best}$). $(X_{teacher})_1 = f(x)_1$, where $f(x)_1 = f(x)_{best}$. Select the other teachers based on the chief teacher and rank them. $f(x)_s = f(x)_1 - rand \times f(x)_1$, where s is the number of teacher as selected; (if the equality is not met, select the $f(x)_s$ closet to the value calculated above). We select four teachers in this algorithm.

Step 4: Assignment; assign the learners to the teachers according to their fitness value as: $(X_{teacher})_s = f(x)_s$, where $s = 1, 2, \dots, T$ (in this paper $T=4$);

For $k=1: (n-s)$

if $f(x)_1 \leq f(x)_k < f(x)_2$,

assign the learner $f(x)_k$ to teacher 1

else, if $f(x)_2 \leq f(x)_k < f(x)_3$,
 assign the learner $f(x)_k$ to teacher 2
 else, if $f(x)_3 \leq f(x)_k < f(x)_4$,
 assign the learner $f(x)_k$ to teacher 3
 else, assign the learner $f(x)_k$ to teacher 4
 end

Step 5: keep the elite solutions of each group; elitism is a mechanism to preserve the best individual from generation to generation. Therefore the system never loses the best individuals found during the optimization process. It can be done by placing one or more of the best individual directly in to the population for the next generation.

Step 6: Updating; calculate the mean result of each group of learners in each subject (i.e. $(M_j)_s$) and evaluate the difference given by Eq. 3. For each teacher, the adaptive teaching factor is as:

$$(T_F)_i = \left(\frac{X_{total-k}}{X_{total-kbest}} \right), \quad k=1, \dots, n,$$

$$(T_F)_i = 1, \quad \begin{array}{l} \text{if } X_{total-kbest,i} \neq 0 \\ \text{if } X_{total-kbest,i} = 0 \end{array}$$

Where $X_{total-k}$ is the result of any learner, $X_{total-kbest}$ is the result of teacher at the same iteration, i. for each group, update the learners' knowledge with the help of teacher's knowledge, along with the knowledge acquired by the learners during the tutorial hours, according to: (where $hh \neq s$)

$$(X'_{j,k})_s = (X_{j,k} + \text{Difference}_{Mean_j})_s + \text{rand}(X_{hh} - X_k)_s, \quad \text{If } f(X)_{hh} < f(X)_s$$

$$(X'_{j,k})_s = (X_{j,k} + \text{Difference}_{Mean_j})_s + \text{rand}(X_k - X_{hh})_s, \quad \text{If } f(X)_s < f(X)_{hh}$$

For each group, update the learner's knowledge by utilizing the knowledge of some other learners, as well as by self learning, according to:

$$(X''_{j,k})_s = X'_{j,k,i} + \text{rand}(X'_{j,k} - X'_{j,p})_s + \text{rand}(X_{teacher} - E_f X'_{j,k})_s,$$

$$\text{if } f(X')_k < f(X')_p$$

$$(X''_{j,k})_s = X'_{j,k,i} + \text{rand}(X'_{j,p} - X'_{j,k})_s + \text{rand}(X_{teacher} - E_f X'_{j,k})_s,$$

$$\text{if } f(X')_p < f(X')_k$$

Where E_f = Exploration Factor = round (1+rand)

Step 7: Replace the worst solution of each group with an elite solution

Step 8: Eliminate the duplicate solutions; it is necessary to modify the duplicate solutions in order to avoid trapping in the local optima. These solutions are modified by randomly selected.

Step 9: combine all groups.

Step 10: check the termination criteria; if the termination criterion is not satisfied, repeat the step 3 to last step, otherwise stop the algorithm. In this experiment, the maximum iteration is 300 and the minimum improvement of objective function is 10^{-6} .

4. EXPERIMENTAL RESULTS

We tested the performance of the ITLBO algorithm for fuzzy clustering on three different real-life datasets (Iris, Thyroid and wine datasets) and two artificial datasets;¹ then the ability of algorithm has been compared with FCM [3], SA [15], TS [14], GA [10], ACO [16], PSO [17]. The well-known datasets are described below:

The Iris dataset: This dataset contains 3 clusters of 50 objects; where each cluster refers to a type of Iris plant as Setosa, Viginia and Versicolor. The data is four dimensional space(sepal length, sepal width, petal length and petal width). There is no missing attribute value.

The wine dataset: this dataset contains 178 data points along with 13 continuous features derived from chemical analysis (e.g. Alcohol, Malicacid, Ash) and it divided in to 3 clusters.

The Thyroid dataset: this data set contains 215 samples of patients suffering from three human thyroid disease; each individual was characterized by five features of laboratory tests.

Artificial dataset: this is a two dimensional data set consisting of 900 points. The data set has 9 classes. The data set is shown in fig. 1.

```
class1:[-3.3,-0.7]x[0.7,3.3]
class2:[-1.3,1.3]x[0.7,3.3]
class3:[0.7,3.3]x[0.7,3.3]
class4:[-3.3,-0.7]x[-1.3,1.3]
class5:[-1.3,1.3]x[-1.3,1.3]
class6:[0.7,3.3]x[-1.3,1.3]
class7:[-3.3,-0.7]x[-3.3,-0.7]
class8:[-1.3,1.3]x[-3.3,-0.7]
class9:[0.7,3.3]x[-3.3,-0.7]
```

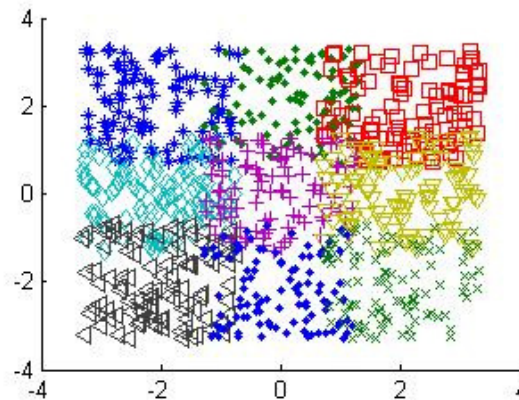


Figure 1. Artificial dataset

The algorithm is implemented in MATLAB Version 12 and the controlling parameters of other algorithms are set the same as their parameters on mentioned references. For maximize compactness inside the clusters, according to FCM algorithm, can calculate the overall summed distances between data and cluster centers, as defined in Eq. 1. Obviously the smaller objective function is, the higher quality of data clustering.

From table 1 we can see that the ITLBO algorithm has achieved the best performance in terms of average compactness criteria.

¹ These datasets are taken from: (<http://www.ics.uci.edu/~mllearn/MLRespository.html>) (<ftp://ftp.ics.edu/pub/machine-learning-databases>)

Table1. Average of compactness index for some of popular algorithms

Algorithms name	Artificial dataset	Iris dataset	Wine dataset	Thyroid dataset
fuzzy c-mean (FCM)	253.32	605.58	17960.84	20642.59
simulated Annealing (SA)	101.82	97.13	16530.53	10114.04
Tabu Search (TS)	102.91	97.86	16785.46	10354.31
Genetic Algorithm (GA)	100.76	125.19	16530.53	10128.82
Ant Colony optimization (ACO)	98.90	97.17	16530.53	10112.13
Particle Swarm Optimization (PSO)	97.83	96.67	16293.00	10109.70
Artificial Bee Colony (ABC)	93.51	78.94	16260.52	10104.03
Improved Teaching Learning Based Optimization (ITLBO)	90.13	77.04	16070.63	10001.18

From the above results, we can obtain that ITLBO algorithm performed better than other mentioned algorithms in terms of intra-cluster distance.

5. CONCLUSIONS

This paper proposed an approach for data clustering based on Improved TLBO algorithm; this algorithm model the process of teaching-learning that every individual to learn something from other individuals to improve themselves. TLBO algorithm does not require any algorithm-specific parameter; because of this advantage the application of algorithm is easier than other meta-heuristics algorithm. To evaluate the performance of this algorithm, it is compared with genetic algorithm, simulated annealing, ant colony, tabu search, artificial bee colony and particle swarm optimization. This algorithm was tested on several datasets.

The experimental results over 4 dataset show that the proposed algorithm is efficient. The ITLBO algorithm performed better than other compared algorithms in terms of intera-cluster distance.

REFERENCES

- [1] Gan, Gujun,Chaoqun Ma & Jianhong Wu, (2007) "Data Clustering: Theory, Algorithms, and Applications", ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA,
- [2] Everitt, B.S., (1993) "Cluster analysis", 3rd edition. New York, Toronto: Halsted Press,
- [3] J.C. Bezdek, (1981) "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York,.
- [4] Brucker, P. (1978) "On the complexity of clustering problems", In M. Beckmenn & H. P. Kunzi (Eds.), Optimisation and operations research. Lecture notes in economics and mathematical systems , Berlin: Springer, Vol. 157, pp. 45–54,.
- [5] El-ghazali Talbi.,(2009) "Metaheuristics : from design to implementation",John Wiley & Sons, Inc.
- [6] Geem, Z. W., Kim, J.H. & Loganathan G.V. (2001) "A new heuristic optimization algorithm: harmony search", Simulation, vol.76, pp 60-70.
- [7] Rashedi, E., Nezamabadi-pour, H. & Saryazdi, S. (2009) "GSA: A gravitational search algorithm", Information Sciences, vol. 179, pp 2232-2248
- [8] Rao, R.V., Savsani, V.J. & Vakharia, D.P. (2011) "Teaching-learning-based optimization: A novel method for constrained mechanical design optimization problems", Computer-Aided Design, Vol. 43, No.3, pp 303-315.
- [9] Rao, R.V., Savsani, V.J. and Vakharia, D.P. (2011) "Teaching-learning-based optimization: a novel optimization method for continuous non-linear large scale problems", Inform. Sci., Vol. 183, No.1, pp. 1–15
- [10] Murthy, C. A., & Chowdhury, N. (1996) "In search of optimal clusters using genetic algorithms", Pattern Recognition Letters, vol.17, pp 825–832.

- [11] Mualik, U., & Bandyopadhyay, S. (2002) "Genetic algorithm based clustering technique", Pattern Recognition, vol.33, pp 1455–1465.
- [12] Krishna, K., & Murty (1999) "Genetic K-means Algorithm", IEEE Transactions on Systems Man and Cybernetics B Cybernetics, Vol.29, pp 433–439.
- [13] Sung, C. S., & Jin, H. W. (2000) "A tabu-search-based heuristic for clustering", Pattern Recognition, Vol.33, pp 849–858.
- [14] Al-Sultan, K. S. (1995) "A tabu search approach to the clustering problem" Pattern Recognition, Vol.28, No.9, pp 1443–1451.
- [15] Selim, S. Z., & Al-Sultan, K. (1991) "A simulated annealing algorithm for the clustering problem", Pattern Recognition, Vol.24, No.10, pp 1003–1008.
- [16] Shelokar, P. S., Jayaraman, V. K., & Kulkarni, B. D. (2004) "An ant colony approach for clustering", Analytica Chimica Acta, Vol. 509, pp 187–195.
- [17] Kao, Y.-T., Zahara, E., & Kao, I.-W. (2008) "A hybridized approach to data clustering", Expert Systems with Applications, Vol. 34, No.3, pp 1754–1762.
- [18] Zhang C. , Ouyang D., & Ning J., (2010) "An artificial bee colony approach for clustering", Expert Systems with Applications, Vol.37, pp 4761-4767
- [19] Rao, R.V., Patel V., (2013) "An improved teaching-learning-based optimization algorithm for solving unconstrained optimization problems", Scientia Iranica, Vol.20, No.3, pp 710-720

AUTHORS

Parastou Shahsamandi E. Received the B.Sc.degree in electrical engineering from Isfahan University of technology, in 1997 and the M.Sc. degree in industrial engineering from the Tarbiyat modares University, in 2007. She is currently pursuing her Ph.D. degree at Science and research branch of Islamic Azad University, Tehran, Iran, where her work focused on fuzzy clustering and optimization.



Soheil Sadi-Nezhad is Assistant Professor at Science and research branch of Islamic Azad University Received the M.sc and B.sc in Industrial engineering (1989 and 1987) from University of Science & Technology, Tehran, Iran. He got his PhD on Industrial Engineering, In 1999. His research Interests mostly are in Soft computing, Fuzzy Multiple Criteria Decision Making, System thinking, Fuzzy Ranking, FIS and Fuzzy clustering.



A MODIFIED INVASIVE WEED OPTIMIZATION ALGORITHM FOR MULTIOBJECTIVE FLEXIBLE JOB SHOP SCHEDULING PROBLEMS

Souad Mekni and Bisma Char Fayeck

National School of Engineering of Tunis, Tunisia, LR-ACS-ENIT
msouadpop@gmail.com, Bisma.fayeckchaar@insat.rnu.tn

ABSTRACT

In this paper, a modified invasive weed optimization (IWO) algorithm is presented for optimization of multiobjective flexible job shop scheduling problems (FJSSPs) with the criteria to minimize the maximum completion time (makespan), the total workload of machines and the workload of the critical machine. IWO is a bio-inspired metaheuristic that mimics the ecological behaviour of weeds in colonizing and finding suitable place for growth and reproduction. IWO is developed to solve continuous optimization problems that's why the heuristic rule the Smallest Position Value (SPV) is used to convert the continuous position values to the discrete job sequences. The computational experiments show that the proposed algorithm is highly competitive to the state-of-the-art methods in the literature since it is able to find the optimal and best-known solutions on the instances studied.

KEYWORDS

Flexible job shop scheduling problem, Multiobjective optimization, Metaheuristics, Smallest Position Value, Invasive Weed Optimization.

1. INTRODUCTION

Solving a NP-hard scheduling problem with only one objective is a difficult task. Adding more objectives obviously makes this problem more difficult to solve. In fact, while in single objective optimization the optimal solution is usually clearly defined, this does not hold for multiobjective optimization problems. Instead of a single optimum, there is rather a set of good compromises solutions, generally known as Pareto optimal solutions from which the decision maker will select one. These solutions are optimal in the wider sense that no other solution in the search space is superior when all objectives are considered. Recently, it was recognized that Invasive Weed Optimization (IWO) was well suited to multiobjective optimization.

The invasive Weed Optimization algorithm developed by Mehrabian and Lucas [1] in 2006 is a newly stochastic optimization approach inspired from a common phenomenon in agriculture: colonization of invasive weeds. IWO is an appropriate competitor for other evolutionary algorithms. In fact, it is simple and easy to understand and program. It has strong robustness and fast global searching ability.

Some of the distinctive properties of IWO in comparison with other numerical search algorithms are the way of reproduction, spatial dispersal, and competitive exclusion. These properties are presented in details in section 3. Section 2 introduces and formulates the flexible job shop scheduling problem. The experiments are provided in section 4. Finally, brief conclusions and future perspectives are discussed in section 5.

2. MATHEMATICAL FORMULATION

The problem of flexible job shop scheduling (FJSSP) belongs to the NP-hard family [2]. It presents two difficulties. The first one is the assignment of each operation to a machine, and the second one is the scheduling of this set of operations in order to optimize our criteria. The result of a scheduling algorithm must be a schedule that contains the start times and allocation of resources to each operation. The data, constraints and objective of our problem are as follows:

2.1. Data

- M represents a set of m machines. A machine is called M_k ($k = 1, \dots, m$), each M_k has a Workload called W_k .
- N represents a set of n jobs. A job is called j_i ($i = 1, \dots, n$), each job has a linear sequence of n_i operations.
- $O_{i,j}$ represents the operation number j of the job number i . The realization of each operation $O_{i,j}$ requires a machine M_k and a processing time $p_{i,j,k}$. The starting time of $O_{i,j}$ is $t_{i,j}$ and the ending time is $t_{f_{i,j}}$.

2.2. Constraints

- Machines are independent of one another.
- A machine can be unavailable during the scheduling (case of machine breakdown).
- Jobs are independent of one another.
- In our work, we suppose that: each job j_i can start at the date $t = 0$ and the total number of operations to perform is greater than the number of machines.

2.3. Criteria

We have to minimize Cr_1 , Cr_2 and Cr_3 :

- The makespan: Cr_1
- The total workload of machines: Cr_2
- The workload of the most loaded machine: Cr_3

In this paper, the objective is to find a schedule which has a minimum makespan, a minimum total workload of machines and a minimum workload of the critical machine. The sum of these three objectives is taken as the objective function. To measure the quality of solutions found, we use the lower bounds (BCr_1 for makespan, BCr_2 for total machine workload, and BCr_3 for the workload of the most loaded machine) proposed in [3].

3. INVASIVE WEED OPTIMIZATION ALGORITHM FOR FJSSP

The IWO algorithm was proposed by Mehrabian and Lucas [1] in 2006, and since then, it has been successfully utilized in different practical optimization problems such as optimal positioning of piezoelectric actuators [4], demanding a recommender system [5], Studying electricity market dynamics [6], Design of an E shaped MIMO antenna [7] and encoding sequences for DNA computing [8].

3.1. Invasive Weed Optimization Algorithm

A weed is any plant growing where it is not wanted. Weeds have shown very robust and adaptive nature which turns them to undesirable plants in agriculture. A common belief in agronomy is that “The Weeds Always Win”. The harder people try, the better they get [1]. Recently, many studies are carried out with inspirations from ecological phenomena for developing optimization techniques. The new algorithm that is motivated by a common phenomenon in agriculture is colonization of invasive weeds. The flow chart of this algorithm is shown in Figure 1 and the details of IWO are addressed as follows:

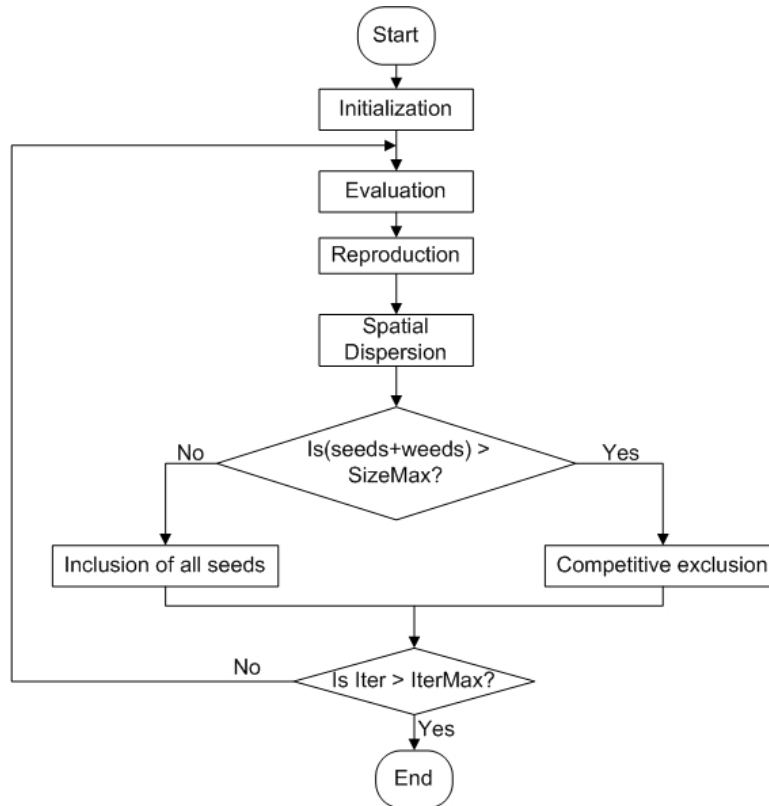


Figure 1. Flow Chart of IWO

3.1.1. Initialization

A population of initial solutions (weeds) is randomly generated over the search space.

3.1.2. Evaluation

The fitness of each weed in the population is calculated.

3.1.3. Reproduction

Each weed in the population is allowed to produce seeds depending on its comparative fitness in the population. In other words, a weed will produce seeds based on its fitness, the worst fitness and the best fitness in the population. In such way, the increase of number of seeds produced is linear. The number of seeds for each weed varies linearly between S_{\min} for the worst plant and S_{\max} for the best plant. Figure 2 illustrates the procedure of reproduction.

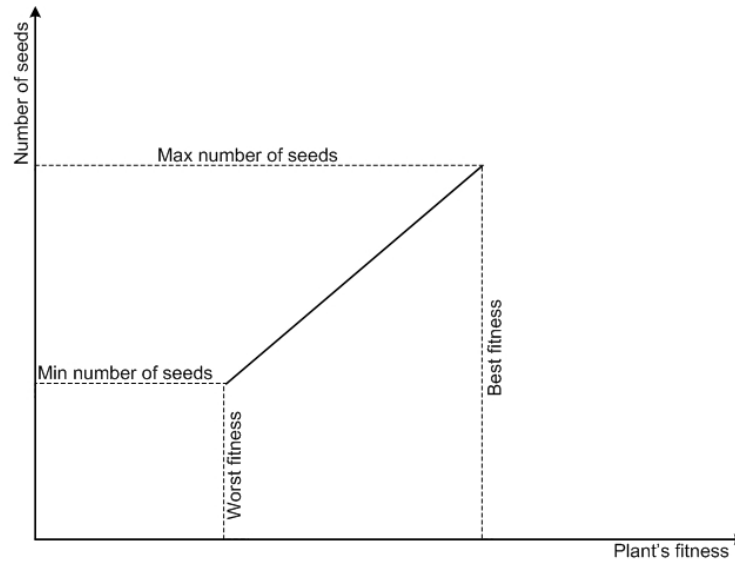


Figure 2. Procedure of reproduction

The equation for determining $Weed_{num}$ the number of seeds produced by each weed is presented in equation (1):

$$Weed_{num} = S_{\min} + (S_{\max} - S_{\min}) \frac{f - f_{worst}}{f_{best} - f_{worst}} \quad \text{Equation (1)}$$

Where f is the fitness of the weed considered, f_{worst} and f_{best} are respectively the worst and the best fitness in the population. For better clarification, the application of equation (1) is shown in Figure 3. In this figure, it is assumed that weed₅ and weed₁ are the worst and best weeds between a population containing five weeds. So, the number of seeds around Weed₅ is equal to S_{\min} and the number of seeds around Weed₁ is equal to S_{\max} .

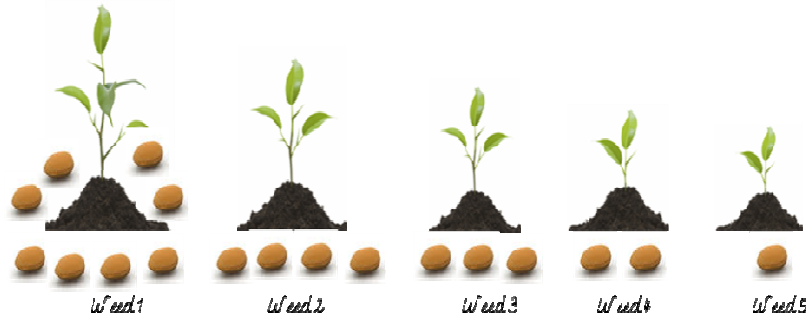


Figure 3. Schematic reproduction procedure for a problem with 5 weeds

3.1.4. Spatial Dispersal

This step ensures that the produced seeds will be generated around the parent weed, leading to a local search around each plant. The generated seeds are randomly spread out around the parent weeds according to a normal distribution with mean equal to zero and variance σ^2 . The standard deviation of the seed dispersion σ decreases as a function of the number of iterations $iter$. The equation for determining the standard deviation for each generation is presented in equation (2):

$$\sigma_{iter} = \frac{(iter_{max} - iter)^n}{(iter_{max})^n} (\sigma_{initial} - \sigma_{final}) + \sigma_{final} \quad \text{Equation (2)}$$

Where $iter_{max}$ is the maximum number of iterations. σ_{iter} is the standard deviation at the current iteration and n is the nonlinear modulation index. Obviously, the value of σ defines the exploration ability of the weeds. Therefore, as $iter$ increases, the exploration ability of all weeds is gradually reduced. At the end of the optimization process, the exploration ability has diminished so much that every weed can only fine its position [9].

3.1.5. Competitive exclusion

After a number of iterations, the population reaches its maximum, and an elimination mechanism is adopted: The seeds and their parents are ranked together and only those with better fitness can survive and become reproductive. Others are being eliminated.

3.2. Weed representation of FJSSP

The original IWO is developed to solve continuous optimization problems, but it can not be applied to discrete problems directly: individuals must be encoded appropriately to solve scheduling problems. In this paper, we implement a coding that takes into account all the constraints and the specificities of the problem. For the (n jobs, m machines, O operations) FJSSP, each plant is represented by four components: each component contains $2 \times O$ number of dimensions. Figure 5, Figure 6 and Figure 7 illustrate the solution representation of a weed corresponding to (3 jobs, 5 machines, 8 operations) FJSSP described in Figure 4. The 1st and 2nd halves of the 1st row of the weed (Figure 6 and Figure 7) represent operations as repetition of jobs (Figure 5). For example (J_1, J_1, J_1) represents $(O_{1,1}, O_{1,2}, O_{1,3})$, (J_2, J_2, J_2) represents $(O_{2,1}, O_{2,2}, O_{2,3})$, and so on. The 2nd row of (Figure 6 and Figure 7) represents weed's position. Each dimension of this row in Figure 6 maps one operation and each dimension of this row in

Figure 7 maps one machine. At this step, we use the Smallest Position Value (SPV) rule [10] to find the permutation of jobs. The smallest component of the weed's position in Figure 6 is -8 which corresponds to job number 1, thus J_1 (or the first operation of J_1) is scheduled first. The second smallest component of the weed's position is -5,2 which corresponds to job number 2, therefore J_2 (or the first operation of J_2) is the second in ordering, etc. The 2nd row of Figure 6 contains a random number in the interval $[0, m]$ that indicates after being rounded to its nearest integer the machine to which an operation is assigned during the course of IWO. The 3rd row of Figure 6 indicates the sequence of jobs in the ordering and the 3rd row of Figure 7 indicates the corresponding machines. Finally, the last row of Figure 6 indicates operations in the order and the last row of Figure 7 indicates starting times. In conclusion, the weed itself presents a solution as it shown in 3rd and 4th row of Figure 6 and Figure 7: First, the operation $O_{1,1}$ of job J_1 is executed by the machine M_1 at time $t = 0$, and then the operation $O_{2,1}$ of job J_2 is executed by the machine M_1 at time $t = 1$, and so on.

		M_1	M_2	M_3	M_4	M_5
J_1	$O_{1,1}$	1	9	3	7	5
	$O_{1,2}$	3	5	2	6	4
	$O_{1,3}$	6	7	1	4	3
J_2	$O_{2,1}$	1	4	5	3	8
	$O_{2,2}$	2	8	4	9	3
	$O_{2,3}$	9	5	1	2	4
J_3	$O_{3,1}$	1	8	9	3	2
	$O_{3,2}$	5	9	2	4	3

Figure 4. Example of (3 J, 5 M, 8 O) FJSSP

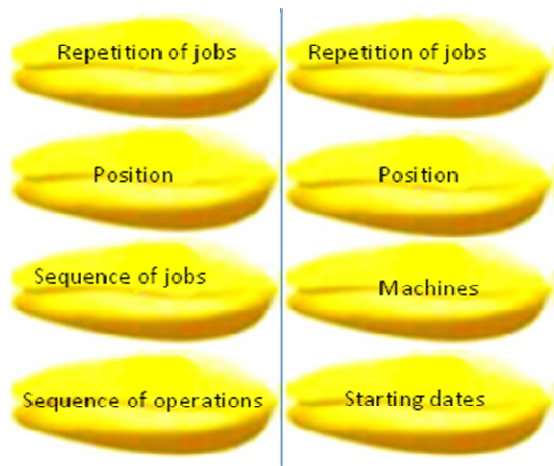


Figure 5. Weed representation

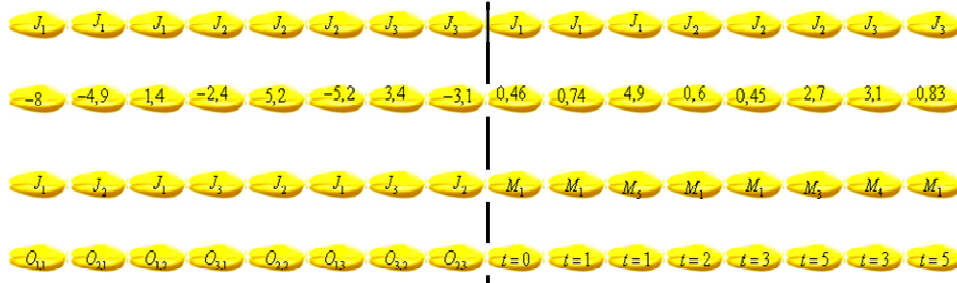


Figure 6. The first half of the weed

Figure 7. The second half of the weed

3.3. Pseudo-code of solving FJSSP by IWO algorithm

```

Begin{
  • Initialize population of weeds, set parameters;
  • Current_iteration=1;
  While (Current_iteration< Max_iteration)do
  {
    • Compute the best and worst fitness in the population
    • Compute the standard deviation std depending on iteration
    For each weed w in the population W
    {
      • Compute the number of seeds for w depending on its fitness
      • Select the seeds from the feasible solutions around the parent weed w in a neighborhood with normal distribution having mean=0 and standard deviation=std;
      • Add seeds produced to the population W
      If (|W|>Max_SizePopulation)
      {
        • Sort the population W according to their fitness
        • W=SelectBetter (weed, seed, Max_SizePopulation)
      }End if
    }End for
    Current_iteration=Current_iteration+1;
  }End while
}End

```

Figure 8. Pseudo code of IWO

4. EXPERIMENTAL RESULTS

Our approach is implemented in C++ on an Intel(R) Core(TM) i3 CPU M370@2,40 GHz machine. The non deterministic nature of IWO algorithm makes it necessary to carry out multiple runs on the same problem instance in order to obtain meaningful results. We run our algorithm twenty times from different starting solutions and tested it on a number of instances from

literature. The convergence of IWO depends on the selection of three parameters: the initial standard deviation $\sigma_{initial}$, the final standard deviation σ_{final} and the non linear modulation index n . The chosen parameters for IWO are given in table 1.

Table 1. Parameters of IWO.

Parameters	values
Number of initial population	50
Maximum number of population	200
Maximum number of iterations: $iter_{max}$	5000
Maximum number of seeds	5
Minimum number of seeds	1
$\sigma_{initial}$	10
σ_{final}	0,5
Non linear modulation index: n	3

To illustrate the effectiveness and performance of the algorithm used in this paper, we choose different instances of the problem of flexible job shop scheduling problem taken from Kacem [11]. Solutions in the literature to the instances presented in table 2 are presented in table3.

Table 2. Instances of Kacem.

Instances	n(jobs)	m(machines)
Instance 1	3	5
Instance 2	4	5
Instance 3	10	7
Instance 4	10	10
Instance 5	15	10
Instance 6	8	8

From table 3, we conclude that the obtained solutions are generally of a good quality. This is noted while comparing them with the existing approaches in the literature (for example Xia approach[12]) and also while comparing obtained values of the criteria with the computed lower bounds [3]. In fact, for instance 1, instance 2 and instance 3 our value of makespan Cr_1 is near the lower bound BCr_1 , our value of total machine workload Cr_2 is near the lower bound BCr_2 and our value of the workload of the critical machine Cr_3 is near the lower bound BCr_3 .

For instance 4, instance 5 and instance 6 our values of criteria are near lower bounds and similar or better (instance 4) than solutions found in [12].

Table 3. Solutions in Literature.

Instances	Lower Bounds	Xia et al [12]	IWO
Instance 1	$BCr_1 = 4$	-	$Cr_1 = 5$
	$BCr_2 = 11$	-	$Cr_2 = 13$
	$BCr_3 = 2$	-	$Cr_3 = 5$
Instance 2	$BCr_1 = 11$	-	$Cr_1 = 11$
	$BCr_2 = 32$	-	$Cr_2 = 32$
	$BCr_3 = 6$	-	$Cr_3 = 10$
Instance 3	$BCr_1 = 11$	-	$Cr_1 = 11$
	$BCr_2 = 60$	-	$Cr_2 = 61$
	$BCr_3 = 8$	-	$Cr_3 = 11$
Instance 4	$BCr_1 = 7$	$Cr_1 = 7$	$Cr_1 = 7$
	$BCr_2 = 41$	$Cr_2 = 44$	$Cr_2 = 42$
	$BCr_3 = 4$	$Cr_3 = 6$	$Cr_3 = 6$
Instance 5	$BCr_1 = 10$	$Cr_1 = 12$	$Cr_1 = 12$
	$BCr_2 = 91$	$Cr_2 = 91$	$Cr_2 = 91$
	$BCr_3 = 9$	$Cr_3 = 11$	$Cr_3 = 11$
Instance 6	$BCr_1 = 12$	$Cr_1 = 15$	$Cr_1 = 14$
	$BCr_2 = 73$	$Cr_2 = 75$	$Cr_2 = 77$
	$BCr_3 = 9$	$Cr_3 = 12$	$Cr_3 = 12$

5. CONCLUSIONS

In this paper, the performance of the Invasive Weed Optimization technique is investigated for solving the multiobjective flexible job shop scheduling problem. The main highlighting features in IWO are: it is simple and easy to understand and program and it has strong robustness and fast global searching ability.

Experimental results are encouraging since that the proposed algorithm is able to find relevant solutions minimizing makespan, total machine workload and the biggest machine workload on the studied instances. A more comprehensive study on a large number of instances should be made to test the efficiency of the proposed solution technique. Further investigation is needed to fully reveal the ability of IWO in tackling scheduling problems and solving other optimization problems. Future research should pay more attention to the hybridization of IWO and other metaheuristics in order to benefit from advantages of each algorithm.

REFERENCES

- [1] A R, Mehrabian. & C, Lucas, (2006) "A novel numerical optimization algorithm inspired from weed colonization", Ecological Informatics, Vol.1, pp355-366.
- [2] M , Sakarovitch, (1984) "Optimisation combiantoire. Méthodes mathématiques et algorithmiques. Hermann, Editeurs des sciences et des arts, Paris.

- [3] R, Dupas, (2004) “ amelioration de performances des systems de production: apport des algorithms évolutionnistes aux problems d’ordonnancement cycliques et flexibles, Habilitation , Artois university.
- [4] A R, Mehrabian & A, Yousefi-Koma (2007) “Optimal Positioning of Piezoelectric actuators on a smart fin using bio-inspired algorithms”, Aerospace Science and technology, Vol 11, pp 174-182.
- [5] H, Sepehri Rad & C, Lucas (2007) “ A recommender system based on invasive weed optimization algorithm”, IEEE Congress on Evolutionary Computation, CEC 2007, pp 4297-4304.
- [6] M, Sahaheri-Ardakani & M.Rshanaei & A, Rahimi-Kian & C, Lucas (2008) “ A study of electricity market dynamics using invasive weed colonization optimization”, in Proc.IEEE Symp. Comput.Intell. Games, pp 276-282.
- [7] A, R, Mallahzadeh & S, Es’haghi & A, Alipour (2009) “Design of an E shaped MIMO antenna using IWO algorithm for wireless application at 5.8 Ghz”, Progress in Electromagnetic Research, PIER 90, pp 187-203.
- [8] X, Zhang & Y,Wang & G, Cui & Y, Niu & J, Xu (2009) “Application of a novel IWO to the design of encoding sequence for DNA Computing, Comput. Math. Appl. 57, pp 2001-2008.
- [9] Z, D, Zaharis & C, Skeberis & T, D, Xenos (2012) “Improved antenna array adaptive beamforming with low side lobe level using a novel adaptive invasive weed optimization method” , Progress in Electromagnetics Research , Vol 124, pp 137-150.
- [10] S, Mekni & B, Châar Fayéçh & M, Ksouri (2010) “ TRIBES application to the flexible job shop scheduling problem”, IMS 2010 10th IFAC Workshop on Intelligent Manufacturing Systems, Lisbon, Portugal, July 1st -2nd 2010.
- [11] I, Kacem & S, Hammadi & P, Borne (2002) “ Approach by localization and multiobjective evolutionary optimization for flexible job shop scheduling problems. IEEE Trans Systems, Man and Cybernetics, Vol 32,pp 245-276.
- [12] W, Xia & Z, Wu (2005) “ An effective hybrid optimization approach for multiobjective flexible job shop scheduling problems, Journal of Computers and Industrial Engineering, Vol 48,pp 409-425.

AUTHORS

Souad Mekni: received the diploma of Engineer in Computer Science from the Faculty of Science of Tunis (Tunisia) in 2003 and the Master degree in Automatic and Signal Processing from the National Engineering School of Tunis (Tunisia) in 2005. She is currently pursuing the Ph.D.degree in Electrical Engineering at the National Engineering School of Tunis. Her research interests include production scheduling, genetic algorithms, particle swarm optimization, multiobjective optimization, Invasive Weed Optimization and artificial intelligence.

Besma Fayéçh Chaâr: received the diploma of Engineer in Industrial Engineering from the National Engineering School of Tunis (Tunisia) in 1999, the D.E.A degree and the Ph.D degree in Automatics and Industrial Computing from the University of Lille (France), in 2000, 2003, respectively. Currently, she is a teacher assistant in the Higher School of Sciences and Techniques of Tunis (Tunisia). Her research interests include scheduling, genetic algorithms, transportation systems, multiagent systems and decision-support systems.

IRANIAN CASHES RECOGNITION USING MOBILE

Ismail Nojavani¹, Amir Hassan Monadjemi² and Azade Rezaeezade³

^{1,2,3}Department of Computer Engineering, Isfahan University, Isfahan, Iran
¹e.nojavani@eng.ui.ac.ir, ²monadjemi@eng.ui.ac.ir
³azade.rezaeezade@eng.ui.ac.ir

ABSTRACT

In economical societies of today, using cash is an inseparable aspect of human life. People use cashes for marketing, services, entertainments, bank operations and so on. This huge amount of contact with cash and the necessity of knowing the monetary value of it caused one of the most challenging problems for visually impaired people. In this paper we propose a mobile phone based approach to identify monetary value of a picture taken from cashes using some image processing and machine vision techniques. While the developed approach is very fast, it can recognize the value of cash by average accuracy of about 95% and can overcome different challenges like rotation, scaling, collision, illumination changes, perspective, and some others.

KEYWORDS

Cash Identification Using Mobile, Visually Impaired Assistant, Iranian Cash Identification, Mobile Phone Extra Usage

1. INTRODUCTION

From the beginning of human appearance on the earth, we always gain lots of information about our surrounding environment visually. As result, our developed technologies mostly are based on vision. Hence, human beings who have some kind of visually impairment always have suffered from that. One of the most common innovations of humans that proof this assertion is money. Visually impaired people have lots of difficulties to use money in daily transactions, unless they use a person or device as an assistant to help them. World Health Organization approximates 285 million people are visually impaired worldwide, that 39 million of them are blind. According to the Health Organization of Iran, about 600 thousands Iranian have moderate or severe visual impairments that about a third of them are totally blind. This statistics show the importance of developing some efficient methods to assist this people in cash recognition. In this paper we develop a mobile phone based cash recognition system that helps visually impaired people to identify the current Iranian cash easily and accurately.

Banknotes in different countries have texture, colour, size, and appearance differences [1], so feature extraction and identification approaches that is used in one country usually is not usable in other countries or does not work properly over there. Moreover, lots of previous work in concept of cash recognition is restricted to specific standard conditions. For example many of the developed approaches do not support rotation or noisy backgrounds [2-5], or in some other approaches the whole banknote must be visible in taken picture [1, 6]. By considering visually

impaired limitations to take pictures with special physical characteristics, these approaches are not user friendly enough.

In this study we consider many complexity and variety for taken pictures and as a result the developed system can support rotation, scaling, complex or noisy background, camera angle changes, collision and even the variation of illumination. For explication we define this concept below and show some of them in Figure 1.

- Rotation: the cash can have different rotation angle to the camera.
- Scaling: the cash can poses in different distances from the mobile camera.
- Complex background: the cash can be taken with any background except other cashes.
- Perspective: the cash can poses in any camera angle.
- Collision: the taken picture can be from a part of cash. It should only contain the digits of monetary value.
- Variation of illumination: cash picture can be taken in different illumination.



Figure1. Some possible varieties in taking picture (oldness,rotation, scaling and different illumination)

In the developed approach, for robust and flexible cash recognition, we use value digits exist on cash. However, regarding the different conditions of taking pictures that was shown in Figure 1, these digits are not in a constant place for all the pictures. So, first of all we should find the place of these digits on the cash image. In this study, to localize the value digits on the image we apply a zero-finding algorithm on it. After that, we proceed to find the remained digit that can be 1, 2 or 5, and try to identify this digit using a neural network. Besides that, we should count the number of zeroes. After these two steps we can identify the monetary value of that cash.

The contributions of this paper include: section 2 reviews the state of the art on banknote recognition, section 3 explains the proposed approach for Iranian cash recognition, section 4 evaluates the recognition results and section 5 conclude the whole paper.

2. RELATED WORK

The existing cash recognition approaches in literature, mainly use image processing and neural network techniques. Moreover, there is some developed devices which use physical characteristics of cashes like size or colour. The most considerable point about cash recognition is that many

developed devices or methods for cash recognition in one country are not usable in other countries. In this section we will introduce some of these devices and methods.

Money Talker [7] is a device that recognizes Australian bank notes electronically, using the reflection and transmission properties of light. This device uses the largely different colours and patterns on each Australian banknote. Different colour lights are transmitted through the inserted note and the corresponding sensors detect distinct ranges of values depending on the color of the note. Cash Test [8] is another device determines the value of banknote by a mechanical means relying on the different lengths of each notes. The downfall of the Cash Test is that it does not allow for the shrinking of notes with time nor the creases or rips that are common in our notes. Moreover, while Cash Test is cheap and extremely portable by the reports of users, it is inaccurate and difficult to use. Kurzweil reader [9], iCare [10] (uses a wearable camera for imaging and a CPU chip for computation) and some other devices have been developed too, each has their advantages and disadvantages, but the common property of them is that the user must carry a device everywhere. Some of them are bulky and expensive too.

In [1], Hassanpour and Farahabadi proposed a paper note recognition method which uses Hidden Markov Model (HMM). This method models cashes using texture characteristics including size, colour and a texture-based feature from banknote and compares extracted vector with the instances in a data base of paper cashes. But the main purpose of this method is to distinguish national banknotes from different countries. In [11], Lui proposed a background subtraction and perspective correction algorithm using Ada-boost framework and trained it with 32 pairs of pixels to identify the bill values. This system uses video recording and works on snapshots. This method does not support rotation and noisy background and strongly relies on the white and straight edges of bill, so it is not usable for Iranian paper cash recognition because they do not have this required margins usually.

In [12] Hassanuzzaman and Yang proposed a component based framework using SURF features. The proposed method achieved 100% accuracy for US dollars through various conditions. While the proposed framework is a robust and effective approach and is available to overcome challenges like image rotation, noisy background, collision, scaling and illumination changes, it is not a suitable method for Iranian banknote recognition. This disproportion refers to the texture difference of Iranian currency tissues and US dollars. For proving this unfitness we have applied this approach to Iranian cashes and checked it by extracting different components. But the output accuracy was extremely low which has convinced us that SURF is not a suitable approach for Iranian cash recognition.

As we can see, while all these methods have their advantages, we cannot use them in a user friendly and accurate manner for Iranian cash recognition. So we are going to develop a user friendly method which is portable with no difficulties, have enough accuracy, and considering impaired people limitation for taking pictures in the next sections.

3. THE PROPOSED APPROACH

In this section we will explain the developed system completely. As mentioned earlier, in this method we should provide a picture of the cash that contains the monetary digits. The taken picture must be from the side that contains Iranian digits. For satisfying this condition, photo taker should use face detection technology of his mobile phone. In front side of all Iranian cashes, there is a face of Imam Khomeini. So if the user is taking picture from the front side his mobile will detect Imam's face. But if he is taking picture from the back side his mobile would not detect any face, so he will understand that the cash must be turned back. We assume this picture is taken by a mobile phone camera, so it does not have high resolution and is taken by a visually impaired

person, so can have any position. Then this picture will be analysed by machine vision and image processing tools to extract the individual features of the cash and to recognize the value of that cash. After that, the identified monetary value will be revealed to the user.

Feature extraction should be done by approaches that are robust to different environmental conditions that reviewed earlier including rotation, scaling, noisy background, collision, perspective, camera different angle, variance in illumination and any other possible situation. For this process we will have two main steps. The first one is localization of monetary value digits. This step is carried out by means of some image processing methods and some innovation in interpreting appeared concepts. Second step is recognition of the monetary value of cash. In this step we will use neural network to identify the non-zero digit and identify the value of that cash. In the following we will explain these two steps.

The mentioned face detection process, localization step for finding zeroes and counting them and identifying non-zero digits are shown in Figure 2.

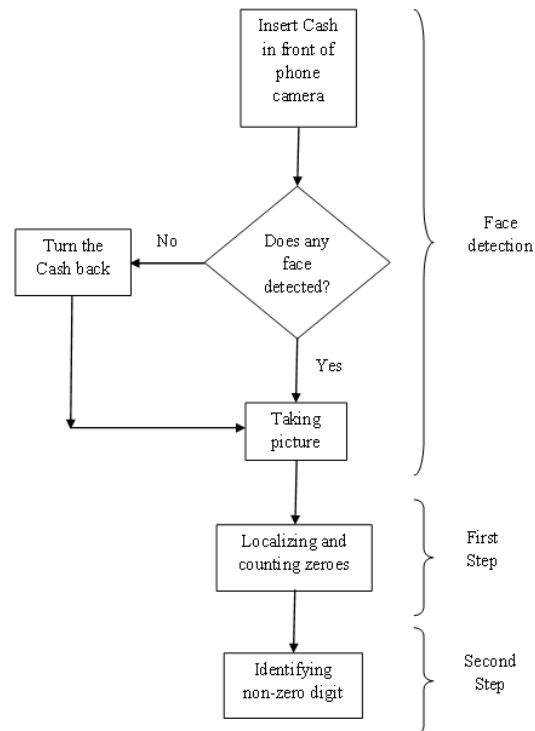


Figure 2. Summary of main steps in proposed approach

3.1. Localization of the Monetary Value

In the first step, to find the place of monetary value digits, the RGB image of the cash should be changed to a grayscale one; it is because all the next processes for feature extraction and cash identification are supposed to be done on a mobile device, so it should not be time and CPU consuming. We apply a Wiener filter on the grayscale image to reduce the effect of oldness of cash or some noises on the image. This grayscale image then will be converted to a binary image. Since photo taker can use any mobile phone camera with any quality, and regarding to that the user can take picture in variant illumination and with different background, for conversion of grayscale image to a valid binary one we use a local adaptive approach for determining the

optimum threshold. The used approach is explained in [8]. In Figure 3(a) a sample of original picture taken by a mobile phone camera is shown. The binary image after grayscale conversion and using local adaptive threshold is shown in Figure 3 (b). As you see, in this picture we have a lot of black and white regions. We refer to these white regions as components.

In some Iranian cashes, like 20000, 50000 and 100000 Ryials, because of existence of some huge white components near the zeroes, there is a possibility that after binarization, these zeroes and big components stick to each other. Moreover, we have this problem for zeroes in 100000 Ryails. For example in Figure 3 (c), a bigger region belongs to Imam's face is stuck to zeroes. So, for separation of these two regions, we use a 3×3 median filter to delete some small components on image and to fill some small holes created on zeroes.

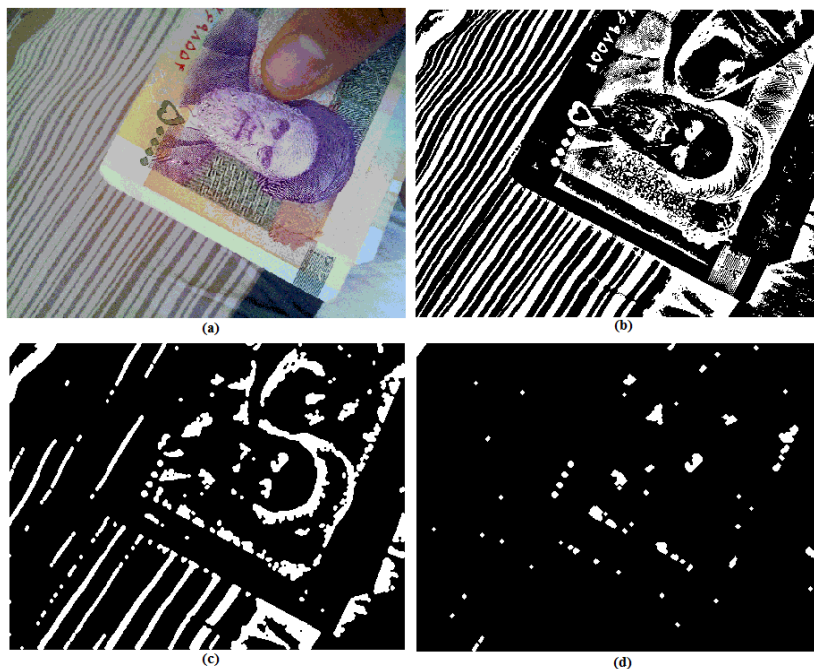


Figure3. (a) A sample of original taken picture, (b) Binary image after applying local adaptive threshold (c) Corresponding image after applying median filter, erosion and dilation operators. (d) Corresponding image after elimination of regions that have a length to width ratio greater than 2 or number of white pixel to boundary rectangle area pixel ratio greater than two.

In continuous, we are going to eliminate regions which cannot be a batch of zeroes because of their shapes. In Persian orthography the width and length of zeroes must be approximately the same. It means that the width and length of boundary rectangle of zeroes on cash or on the taken picture must be approximately equal. So we can omit regions which the ratio of their boundary rectangle length to its width is more than 2. So, components in a region with too different width and length can be omitted with a high reliability. Moreover, while the shape of component are irregular, there are some component that the ratio of length and width of their boundary rectangle is less than 2, but if we calculate the ratio of their white pixel to the boundary rectangle area it is greater than 2. It means that these components cannot be zero too. So we can omit them again. The resulted image is shown in Figure 3(d).

While the existing zeroes in the image of cash are near to each other, if we apply closing morphology operator to the image, these zeroes and some other components will stick to each other again. The result image is shown in Figure 4(a).

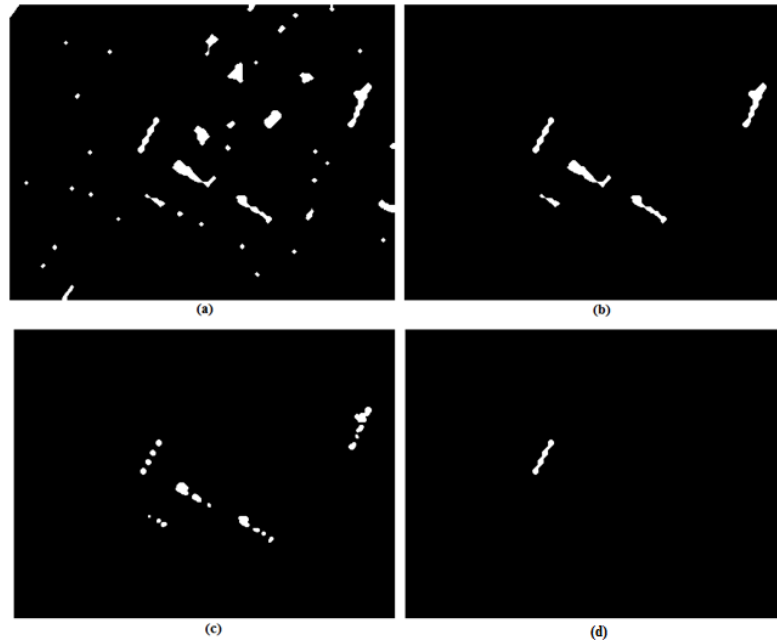


Figure4. (a) Corresponding image after using closing operator. (b) Elimination of regions with less than 3 components. (c) Corresponding image after elimination of regions and components which are not belong to an acceptable line. (d) Keeping the region with the least difference between its components.

The lowest number of zeroes on current Iranian cashes is 3. So after applying closing operator, we can delete connected regions which are made up of less than 3 components. This will limit the number of questioned regions for being a batch of zeroes seriously. Resulted image is shown in Figure 4(b). We should notice that it is not correct to omit regions with more than 5 components (5 is the number of zeroes in 100000 Riyal) because some nonzero components might be connected to the zeroes too.

According to the existence of zeroes on a unique line, we can find regions with more than 2 components on a communal line and omit other regions and then in remained regions omit components which are not on the corresponding line. For this purpose, we extract a line in each connected region crossing gravity centres of any two components of that region. For example, if we have a region with 3 connected components, we should establish equation for 2 lines. The first one must cross gravity centre of the first and second components. The second one must cross gravity centre of the second and third components. After that, for each line we should test if it crosses other components of the connected regions or not. In the previous example, firstly we should test that the first line crosses the third component or not and then we should check if the second line crosses the first component or not. Notice that in this step existence of only one pixel of the component on that line is sufficient. For each connected regions we keep the line that have the most components on it. In addition, for each connected regions any other components which are not on the established line should be omitted because they cannot be zero. The result is shown in Figure 4(c).

As it is shown in Figure 4 (c), it is possible that the algorithm keeps more than one line. Now we should expand a solution to find the right region that contains zeroes. For this purpose we notice the size of zeroes that should approximately be the same. So, for each region we get the size of all components. After comparing the size of all components in each region, we keep only the region that its components have the least difference and omit the other lines and regions. The result is shown in Figure 4 (d).

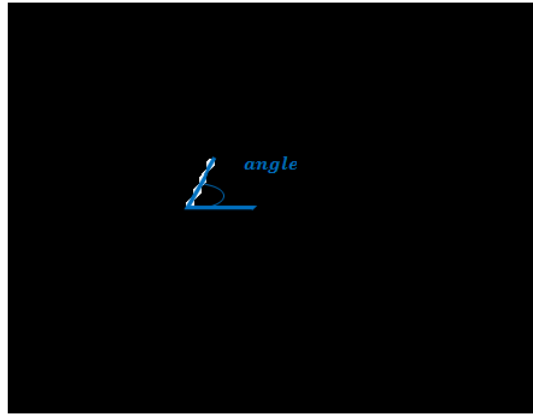


Figure 5. Obtaining the rotation angle

Surprisingly we can compute the rotation angle of the cash by means of corresponding line of the last remained region which should be the batch of zeroes. It is just enough to obtain the angle of the line that crosses the centre of first and last zeroes as it is shown in Figure 5. We can compute the angle from (1).

$$\text{angle} = \text{arctang} \left(\frac{y_1 - y_0}{x_1 - x_0} \right) \quad (1)$$

In this equation, y_0 is central width of first zero, y_1 is central width of last zero, x_0 is central length of first zero and x_1 is central length of last zero. Now if we rotate the image of Figure 3(b) by angle, the rotated image in Figure 6 will be generated. Then, we can easily count the number of zeroes. This count will be used for identifying the monetary value of image.

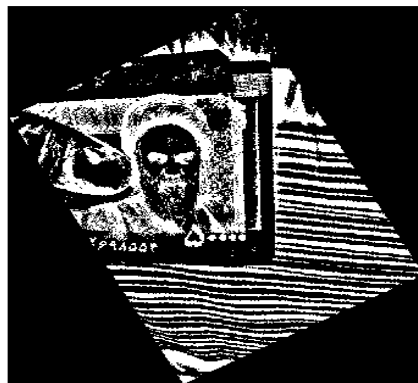


Figure 6. Corresponding image after rotation by angle

3.2. Recognition of Monetary Value

Till now, we could find place of zeroes and the rotation angle. Now we can find the nonzero digit using place of zeroes. In this section we will discuss the features extracted for remained digit. To find the nonzero digit, we start moving to right and left from the most right zero and the most left one equally. We will stop when we find the first component. This component certainly is our nonzero digit. For extracting this digit from the image, we use the length of crossed line from first zero to the last. In all cashes the nonzero digit will be fixed in a square by size of L that L is obtained from (2):

$$L = \frac{3 * \text{length of line}}{\text{number of zeroes}} \quad (2)$$

One of the vertical sides of square is very close to last zero and the centre of this side is intersecting by the corresponding line crosses zeroes. The other vertical side of square is parallel with the first one and has distance of size L from it. According to this we can extract the nonzero digit of monetary value. Figure 7 shows an extracted digit using this approach. Probably we will have some small components in the extracted square but we remove them by keeping the largest connected component and delete remainders. So we will maintain only the nonzero digit.



Figure7. Extracted nonzero digit

Now it is time to identify the number exist in extracted square. As we mentioned earlier, remained digit always is 1, 2 or 5. So for identification of any digit remained after finding it, we must extract features that can classify these digits accurately, so we choose features as below:

1. Ratio of digit pixels to boundary rectangle pixels.
2. Ratio of height of digit to its width (it is computed as ratio of length to height of boundary rectangle for each digit)
3. Ratio of left length to right length and ratio of top width to bottom width.
4. Horizontal and Vertical Symmetry that compare corresponding features in top and down half and right and left half.

For classification we used a Multi-Layer Perceptron (MLP) neural network with one hidden layer. For obtaining the best number of neurons in hidden layer, the network has been trained and tested with different number of hidden nodes and the best result has been achieved by 20 neurons in one hidden layer.

4. EVALUATION OF THE PROPOSED METHOD

In this section we will demonstrate the accuracy and performance of our method. As we mentioned earlier, proposed approach is a two-phase method:

1. Localization of monetary digits and counting the zeroes
2. Identifying nonzero digit of the monetary value

So for computing the accuracy of this method we must use conditional probability. It is because of dependence of second phase accuracy to the first phase. We suppose that the accuracy of correct localization and counting number of zeroes is a% and accuracy of identifying remained digit correctly is b%, so total accuracy of proposed method can be compute via the conditional probability of (3):

$$\text{accuracy} = a\% * b\% \quad (3)$$

For testing proposed approach we collect a data set containing 3500 image of current Iranian cashes in different states and by a vast amount of variety (different environmental conditions we mentioned in the introduction). We took these pictures by different mobile phone cameras which have different resolution from 3 to 8 mega pixels. The rate of accurate localization of zeroes and accurate count of them (first phase) for different cashes are shown in Table 1.

Table 1. Accuracy of localizing monetary value and counting number of zeroes

Cash	Accuracy
1000 Riyals	%96.8
2000 Riyals	%91.6
5000 Riyals	%96
10000 Riyals	%98.4
20000 Riyals	%93.6
50000 Riyals	%95.2
100000 Riyals	%89.4
Average	94.43%

As we mentioned earlier in 100000 Riyals cashes, the zeroes are near each other. Even in some cases they are stick to each other. Moreover, as you can see in Figure 8 in the case of 100000 Riyals cashes, there is an edge around zeroes. After binarization, this edge creates some noises around the zeroes that cause to incorrect localization of zeroes. The decrease in detection of zeroes place in the case of 100000 Riyals in Table 1 is because of this issues.

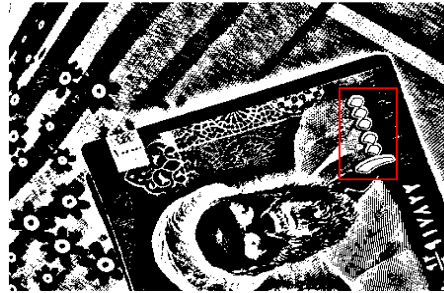


Figure 8: 100000 Riyal's zeroes has stuck to each other

As we mentioned earlier, we used MLP neural network with one hidden layer for identifying nonzero digit of monetary value. Number of neurons in the hidden layer is 20. Number of neurons in the input layer is 4 (the number of extracted features) and the number of neurons in the output layer is 3 (3 output including 1(1), 2(2) and 5(5) and we consider one neurons for each output). Transfer function for hidden layer is 'logistic' too.

Table 2. Accuracy of nonzero digits identification

Nonzero Digit	Accuracy in Training Set	Accuracy in Test Set
1(1)	99.40%	98.83%
2(2)	99.69%	99.28%
5(5)	100%	100%
Average	99.7%	99.37%

For training and test of neural network, we divide created data set into 2 distinct data sets randomly. 70% of samples should be in training data set and reminded 30% should be in test data set. Rate of correct identification or accuracy in this phase is shown in Table 2.

Now by means of Table 1 and Table 2 and by using multiplying rule in equation (3), total accuracy will be obtained as what is shown in Table 3.

Table 3 . Total accuracy for Iranian cash recognition

Cash	Total Accuracy
1000 Riyals	95.67%
2000 Riyals	90.94%
5000 Riyals	96%
10000 Riyals	97.25%
20000 Riyals	92.93%
50000 Riyals	95.2%
100000 Riyals	88.35%
Average	93.76%

5. CONCLUSIONS

In this paper we presented a fast and accurate method for recognizing Iranian monetary value of cashes. This proposed method is based on image processing and machine learning approaches. As input we need a picture taken by mobile phone camera. So the resolution can be very low. We take into account the limitation of visually impaired people for taking pictures. So this proposed approach is robust to rotation of input picture, different scale of the taken picture, different perspective of camera, variation of illumination, noisy background and collision. Moreover, this method is robust to different design of cashes. All this robustness is because of special component we focus on; all the processes in this method are on monetary value of cashes.

In the first step of our method we find the place of monetary value on the cash. This is done with a string of image processing procedures and operators. After finding the place of monetary value or more specially the place of zeroes we count the number of them, discover the rotation angle of input picture and extract the nonzero digit of monetary value. In second step we identify the nonzero digit by means of a MLP neural network with one hidden layer and 20 neurons in that layer. Outputs of this neural network are 1, 2 or 5 digits based on the input image. Based on accuracy of this step we could achieve total accuracy around 95%.

While the accuracy of this work is high in the field of Iranian cash recognition, and while it is the only implemented method of Iranian cash recognition for visually impaired people, the accuracy is not enough for a reliable detection. So one of the most important planes for future work is increasing the accuracy by means of machine vision approaches or testing other methods for nonzero digit identification.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Sayed Amir Hassan Monadjemi, all the survey and study participant.

REFERENCES

- [1] H.Hassanpour & P. Farahabadi,(2009), “Using hiddenMarkovmodels for paper currency recognition,” *Expert Syst. Appl.*, Vol. 36, pp. 10105–10111.
- [2] A. Frosini, M. Gori, & P. Priami, (1996), “A neural network-based model for paper currency recognition and verification,” *IEEE Trans. Neural Netw.*, Vol. 7, No. 6, pp. 1482–1490.
- [3] T. Kosaka & S. Omatu, (1999), “Bill money classification by competitive learning,” in *IEEE Midnight-Sun Workshop Soft Comput. Methods Ind. Appl.*, pp. 5–9.
- [4] T. Kosaka, S. Omatsu & T. Fujinaka, (2001), “Bill classification by using the LVQ method,” in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Vol. 3, pp. 1430–1435.
- [5] J. Lee, S. Jeon, and H. Kim, (2004), “Distinctive point extraction and recognition algorithm for various kinds of euro banknotes,” *Int. J. Control, Autom., Syst.*, Vol. 2, No. 2, pp. 201–206.
- [6] N. Jahangir & A. Chowdhury, (2007), “Bangladeshi banknote recognition by neural network with axis symmetrical masks,” in *Proc. IEEE 10th Int. Conf. Comput. Inf. Technol.*, pp. 1–5.
- [7] A. Hinwood, P. Preston, G. J. Suaning & N. H. Lovell, (2006), “Banknote recognition for the vision impaired,” in *Australasian Physical Engineering Sciences in Medicine*, Vol. 29, No. 2, pp. 229-233.
- [8] Reserve Bank of Australia, “How the RBA assists people with a vision impairment to differentiate notes,” http://www.rba.gov.au/CurrencyNotes/vision_impaired.html, accessed 7/6/2005.
- [9] “K-NFB Reader Website,” (2007), <http://www.knfbreader.com>.
- [10] S. Krishna, G. Little, J. Black & S. Panchanathan, (2005), “A wearable face recognition system for individuals with visual impairments,” *Proceedings of the 7th international ACM SIGACCESS conference on Computers & accessibility*, pp. 106–113.
- [11] X. Liu, (2008), “A camera phone based currency reader for the visually impaired,” in *ACM Proc. ASSETS’08*, Halifax, Nova Scotia, Canada, pp. 305-306.
- [12] F. M. Hasanuzzaman, X. Yang & Y. Tian, (2009), “Robust and Effective Component-based Banknote Recognition by SURF Features,” *Expert Syst. Appl.*, Vol. 36, pp. 10105–10111.
- [13] T.Romen Singh, Sudipta Roy, O.Imocha Singh, Tejmani Sinam, Kh.Manglem Singh, (2011), “A New Local Adaptive Thresholding Technique in Binarization,” *IJCSI International Journal of Computer Science Issues*, Vol. 8, Issue 6, No 2.
- [14] J. Serra, (1986), “Introduction to Mathematical Morphology,” *Computer Vision, Graphics and Image Processing*, Vol. 35, No. 3, pp. 283-305.

AUTHORS

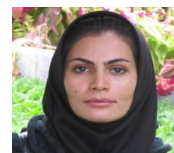
I.Nojavani born in 1987 in Bonab, Iran. He received his B.Sc. in computer software from Urmia University, Urmia, Iran, in 2010. Currently he is a M.Sc. student in Artificial Intelligence at the Department of Computer Engineering, University of Isfahan, Isfahan, Iran. His research interest contains OCR, DIP, and Artificial Intelligence.



S. A. Monadjemi is born in 1968 in Isfahan, Iran. He received his B.Sc. in electrical/computer engineering from Isfahan University of Technology, Isfahan, Iran in 1992, and his M.Sc. in computer engineering, machine intelligence from Shiraz University, Shiraz, Iran in 1995 and his PhD in computer engineering, processing and pattern recognition, from Bristol University, Bristol, England in 2004. His research interests are DIP, Machine Vision, Pattern Recognition, Artificial Intelligence, and Training through Computer. Dr. Monadjemi is currently an Asst. Professor in Department of Computer Engineering, University of Isfahan, Isfahan, Iran.



A. Rezaeezade born in 1988 in Aligudarz, Iran. She received her B.Sc. in computer hardware from Isfahan University of Technology, Isfahan, Iran, in 2010. Currently she is a M.Sc. student in Artificial Intelligence at the Department of Computer Engineering, University of Isfahan, Isfahan, Iran. Her research interest contains OCR, DIP, and Artificial Intelligence.



INTENTIONAL BLANK

QUBIT DATA STRUCTURES FOR ANALYZING COMPUTING SYSTEMS

Vladimir Hahanov¹, Wajeb Gharibi², Svetlana Chumachenko¹ and Eugenia Litvinova¹

¹Department of Computer Engineering, Kharkov National University of Radioelectronics, Kharkov, Ukraine

hahanov@kture.kharkov.ua

²Jazan University, Kingdom of Saudi Arabia.

gharibiw2002@yahoo.com

ABSTRACT

Qubit models and methods for improving the performance of software and hardware for analyzing digital devices through increasing the dimension of the data structures and memory are proposed. The basic concepts, terminology and definitions necessary for the implementation of quantum computing when analyzing virtual computers are introduced. The investigation results concerning design and modeling computer systems in a cyberspace based on the use of two-component structure <memory - transactions> are presented.

KEYWORDS

Quantum Computing, Data Structure, Qubit Model.

1. INTRODUCTION

Market feasibility of emulation of quantum computing methods to create virtual (cloud) computers (VC) in a cyberspace is based on the use of qubit data models, focused on parallel solving discrete optimization problems through significant increase in memory costs. We do not consider the physical basis of quantum computing, originally described in the works of scientists, focused on the use of non-deterministic quantum interactions within the atom. We do not address the physical foundations of quantum mechanics, concerning non-deterministic interactions of atomic particles [1-4], but we use the concept of qubit as a binary or multivalued vector for a concurrent definition of the power set (the set of all subsets) of the states for the discrete cyberspace area based on linear superposition of unitary codes, focused on parallel executing methods for analyzing and synthesizing cyberspace components.

In the market of electronic technologies there is competition between basics of idea implementation [2]: 1) Soft implementation of the project related to the synthesis of interpretative model of the device or hardware implementation of programmable logic devices based on FPGA, CPLD. There are advantages in manufacturability design modifications, the disadvantages - low performance in the operation of a digital system. The advantages lie in the manufacturability of design modifications, the disadvantages – in low performance of a digital system. 2) Hard implementation is focused to the use of compilation models in developing software applications or in the implementation of the project on the basis of VLSI chips. The advantages and

disadvantages of hard implementation are inverted with respect to soft projects: high performance and impossibility of modification. Given above mentioned four basic variants for the implementation of the idea below are represented quantum data structures, focused on raising the performance of flexible models of software or hardware project implementation.

2. QUANTUM STRUCTURES FOR DESCRIBING DIGITAL SYSTEMS

n-Qubit is a vector form of unitary encoding the universe of n primitives to specify the power set of states 2^{2^n} by using 2^n binary variables

For example, if n=2, then the 2-qubit sets 16 states by using 4 variables. If n=1, the qubit sets 4 states on the universe of two primitives by using 2 binary variables (00,01,10,11) [5]. Herewith, the superposition (simultaneous existence) of 2^n states in a vector is supposed. Qubit (n-qubit) allows using the logical operations instead of set-theoretic ones to significantly speed up the analysis of discrete systems. Further the qubit is identified by the n-qubit or vector if this does not prevent the understanding of presented material. As quantum computing is related to analysis of qubit data structures, further we use definition "quantum" for identifying technologies, based on two properties of quantum mechanics: concurrency of processing and superposition of states. When defining logical functionality the following synonym of qubit is used: Q-coverage (Q-vector) [5] is used as a unified vector form for the definition of superposed output states corresponding to unitary codes of addresses for input variables of any logical function.

Qubit of a digital system is a form for defining a structural primitive that is invariant to the technologies for implementing the functionality (hardware, software). Moreover, "quantum" synthesis of digital systems based on qubit structures is not rigidly tied to the Post's theorem, defining the conditions for the existence of a functionally complete basis. At the proposed abstraction level n-qubit gives exhaustive and wider opportunities for vector defining any function of the set $\beta(f) = 2^n$. Format of the structural qubit component of the digital circuit $Q^* = (X, Q, Y)$ involves interface (input and output variables), as well as qubit-vector Q, defining the functionality $Y = Q(X)$, the dimension of which is defined by the power function of the number of input lines $k = 2^n$.

Practically oriented novelty of qubit modeling lies in replacing truth tables of digital device components by vectors of output states. Such transformations can be simply demonstrated for the logic element. Let functional primitive has the following binary coverage:

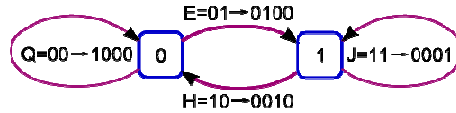
$$P = \begin{array}{|c|c|c|} \hline X_1 & X_2 & Y \\ \hline 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ \hline \end{array},$$

which can be transformed by unitary encoding input vectors based on the use of two-stroke alphabet [2, 5]. It was originally designed for a compact description of all possible transitions of automaton variables, as illustrated in Fig. 1 by corresponding graph and interpretation of symbols.

Here symbols, their binary and unitary codes are presented (for instance, $Q = 00 - 1000$) for describing the two adjacent states of automatic variables. Structurally alphabet is the set of all subsets of the states on the universe of four primitives $Y = \{Q, E, H, J\}$. A unitary code corresponds

to the format of the vector comprising two qubits by using which 16 symbols of two-stroke alphabet are generated. By using two-stroke alphabet any coverage of functional two-input logic primitive can be represented by two or even one cubes, given that they are mutually inverse:

$$P = \begin{matrix} 00 & 1 \\ 01 & 1 \\ 10 & 1 \\ 11 & 0 \end{matrix} = \begin{matrix} Q & 1 \\ E & 1 \\ H & 1 \\ J & 0 \end{matrix} = \begin{matrix} V & 1 \\ J & 0 \end{matrix} = \begin{matrix} 1110 & 1 \\ 0001 & 0 \end{matrix} \rightarrow \begin{matrix} 1 & 1 & 1 & 0 \end{matrix}$$



Q = 00 → 1000	E = 01 → 0100	H = 10 → 0010	J = 11 → 0001
O = {Q, H} = = {00,10} → 1010	I = {E, J} = = {01,11} → 0101	A = {Q, E} = = {00,01} → 1100	B = {H, J} = = {01,11} → 0101
S = {Q, J} = = {00,11} → 1001	P = {E, H} = = {01,10} → 0110	C = {E, H, J} = = {01,10,11} → 0111	F = {Q, H, J} = = {00,10,11} → 1011
L = {Q, E, J} = = {00,01,11} → 1101	V = {Q, E, H} = = {00,01,10} → 1110	Y = {Q, E, H, J} = = {00,01,10,11} → 1111	∅ = 00 → 0000

Figure 1. Two-stroke alphabet of automaton variables

At first all the pairs are encoded by the symbols of two-stroke alphabet and then the union of the first three cubes is made according with the rule “co-edge” operator [3]: vectors differing in one coordinate are minimized in one vector. Further the resulting coverage of two cubes is encoded by qubit vectors, corresponding to the given symbols. For modeling fault-free behavior it is enough to have only one cube (zero or unit one), since the second one is always a complement to the first. Consequently, for example, if we choose the unit cube determining 1 in the output, we can remove the bit of output primitive status, reducing the dimension of the cube or primitive model up to the number of addressable primitive states of the element, where address is the vector composed of the binary values of the input variables, by which output state of the primitive is determined. By analogy any truth table can be led to the qubit functionality in the form of vector of output states of logic element of n inputs.

The modeling procedure for Q-vector of the functionality is reduced to writing the bit status in the output variable Y; the bit address is formed by concatenation of values of input variables: $Y = Q(X) = Q(X_1 * X_2... * X_j... * X_k)$. For modeling digital systems components of which are Q-primitives interrelated on the basis of M-vector of equipotential lines, the processing procedure is defined by the equation: $M(Y) = Q[M(X)] = Q[M(X_1 * X_2... * X_j... * X_k)]$. Taken into account numbering of Q-primitives the universal procedure for modeling the current i-element has the following format: $M(Y_i) = Q_i[M(X_i)] = Q_i[M(X_{i1} * X_{i2}... * X_{ij}... * X_{ik})]$. In this case, the algorithm for analyzing digital system is greatly simplified and the performance of interpretative modeling is improved in 2^n times by increasing the amount of memory to describe the functionality of a circuit structure.

Synthesis of Q-coverage for a digital system is reduced to performing superposition of Q-vectors of the functionalities included in it. For example, for three primitives (and, and-not, and-not), which compose a circuit, the operation of superposition generates a Q-vector of whole functionality, where its dimension is greater than the sum of Q-coverages of the original primitives:

bottom row will be considered as one of $k = 2^{2^{16}}$ primitives of third hierarchy level. In each hierarchy level of qubits the number of states (the power set) is exponentially dependent on the number of primitives-vectors $k = 2^{2^n}$. If the vector Q includes all unit values $Q=1111111111111111$, it simultaneously determines the space containing 16 symbols of two-stroke alphabet, which correspond to the power set on the universe of four primitives [3,5].

The main innovative idea of quantum computation compared to the von Neumann machine is to move from the computational procedures of the byte operand, defining a single solution (point) in the discrete space to the quantum parallel processes of qubit operand, at the same time forming the Boolean of solutions. In this thesis the future of high-performance computers for parallel non-digit analysis and synthesis of structures and services for discrete cyberspace are formulated. In other words, the computational complexity of the procedure for processing a set of n elements in the "quantum" processor and a single element in von Neumann machine are equal due to respective n-fold increase in the hardware complexity of the quantum structure.

3. GRAPH STRUCTURES FOR DESCRIBING DIGITAL CIRCUITS

A somewhat different circuitry, not based directly on the transistors can be presented in the form of graph structures, where each node (arc) is identified with the functional transformation, which is given by a Q-vector. The arc (node) defines the relationship between the functional Q-coverages, as well as input and output variables. The implementation of such structures is based on the memory cells (LUT FPGA), which are capable to store information in the form of Q-vector, where each bit or digit has a unique address, identified with the input word. However, the software implementation of the structures is competitive in the EDA market in speed due to address implementation of modeling processes for the functional primitives. In addition, hardware support for EDA systems in the form of Hardware Embedded Simulator (HES, Aldec) [5] acquires a new motivation for system-level design of digital products, when software and hardware solutions have the same qubit format. Below (Fig. 2) a combinational circuit, containing 6 primitives and three different logic elements is proposed for consideration.

Three generic graph forms of digital functionality, which use Q-vectors to define the behavior of logical primitives, correspond to the circuit and are shown in Fig. 3. The structure shown in Fig. 3a contains 12 lines (arcs) associated with the quantum functionalities (1=0001, 7=0111, 14=1110). It is similar to the traditional structural-functional model of a combinational circuit. The graph in Fig. 3b like Sharshunov's register transfer model [2], which is reversed first structure. Here, blue horizontal arcs are identified with functionalities, and nodes – with the groups of input lines for functionalities, combined in register variables by green vertical arcs. States of these lines form a binary vector used as an address for calculating the state of logic element or more complicated functional. The variables used in forming an address for a Q-vector of the functionality can be combined into a single node, with showing all the identifiers of lines, which form vector-address. The register graph of the combinational circuit is ranked by levels of formation of input signals, enabling conditions for concurrent handling of the elements of the same level and performability of Seidel iterations [2], which improve the performance of algorithms for fault-free simulating digital systems.

The structure, represented in Fig 3b, is interesting by its register implementation that can be used to formalize the descriptions of both software and hardware models of gate, registry and system levels. This presentation is difficult for perceiving by a person, but it is technological and easily "understood" by computer to automatically create software systems for analyzing and synthesizing computing structures and cyberspace services. Thus, the quantum-register graph of a digital circuit is discontinuous (by galvanic connections) flexible system of interconnected

addressed primitives for creating the functional structure of any complexity, especially on PLD, where all combinational primitives are implemented by memory elements (LUTs), which provides high operational performance and online repairing the logic modules.

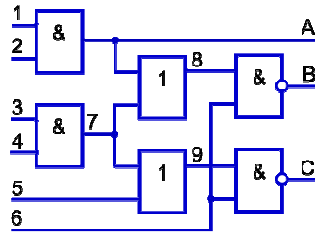


Figure 2. Combination structure of logical primitives

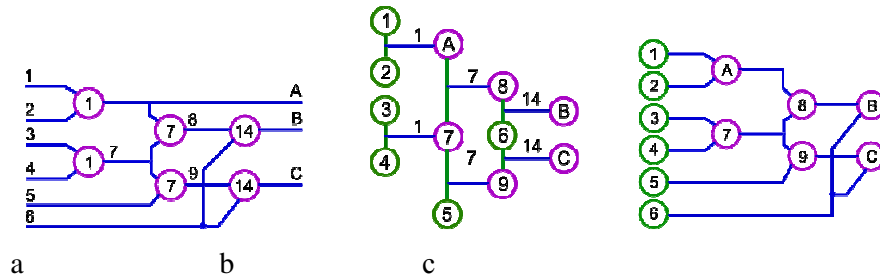


Figure 3. Graph forms of quantum functionalities

A one-dimensional Q-vector is a vector describing the functionality. It can be associated with the output (internal) line of the unit, which is formed in the process of simulation of considered Q-coverage. The register implementation of a combination unit can be represented by the modeling vector M. The functionality with arcs is associated with non-input lines going from the input variables, the values of which form an address of Q-vector bit that forms the state of non-input line under study (Fig. 3c). Otherwise, if the functionality is described by single output primitives, each of them can be identified by number or coordinate of non-input line associated with the element. If the functionality is multi-output, Q-coverage is represented by a matrix with the number of rows equal to the number of outputs. The preference of such primitive lies in the parallelism of concurrent computing the states of several outputs in one access to the matrix at the current address! This fact is an important argument in favor of synthesis of generalized qubits for fragments of a digital unit or whole circuit for their parallel processing in a single time frame.

Closed to the ideal data structure in terms of compactness and processing time, where Q-vectors of functionalities and numbers of input variables are associated with non-input lines of the unit, is the following table:

L	1	2	3	4	5	6	7	8	9	A	B	C
M	1	1	1	1	1	0	0	1	1	1	0	1
X	34	A7	75	12	86	96
Q	0	0	0	0	1	1
	0	0	1	1	1	1
	0	0	1	1	1	1
	1	1	1	1	0	0

It shows external variables of a digital circuit, how many functional primitives are available in the structure, and which inputs are associated with each Q-vector. The advantage of the table is the absence of the vector of output numbers for each primitive, but it is still a need to have numbers of input variables for generating addresses, processing of which is time-consuming. Model for analyzing the circuit structure is simplified to calculating two addresses (!) when forming the

modeling vector $M_i = Q_i[M(X_i)]$ by eliminating the complex address of the primitive output in writing output states to the coordinates of the vector M .

The qubit-register graph of Fig. 3c can be represented as matrix $\mu = \left| \mu_{ij} \right|$, $i = \overline{1, p}$; $j = \overline{1, q}$ for parallel-to-serial processing logic primitives:

μ_{ij}	1	2	3
1	$\begin{array}{ c c } \hline 1 & 1A \\ \hline 2 & \\ \hline \end{array}$	$\begin{array}{ c c } \hline A & 78 \\ \hline 7 & \\ \hline \end{array}$	$\begin{array}{ c c } \hline 8 & 14B \\ \hline 6 & \\ \hline \end{array}$
2	$\begin{array}{ c c } \hline 3 & 17 \\ \hline 4 & \\ \hline \end{array}$	$\begin{array}{ c c } \hline 7 & 79 \\ \hline 5 & \\ \hline \end{array}$	$\begin{array}{ c c } \hline 6 & 14C \\ \hline 9 & \\ \hline \end{array}$
3	$\begin{array}{ c c } \hline X & 1X \\ \hline X & \\ \hline \end{array}$	$\begin{array}{ c c } \hline X & 7X \\ \hline X & \\ \hline \end{array}$	$\begin{array}{ c c } \hline X & 14X \\ \hline X & \\ \hline \end{array}$

which shows the interaction of Q-coverages at three operation levels in accordance with the format (X–Q–Y) inputs-Q-vector-output for each primitive: [(1,2–1–A), (3,4–1–7)], [(A,7–7–8), (7,5–7–9)], [(8,6–14–B), (6,9–14–C)]. To provide the correctness of the functionality, it is necessary to generate all input variables till a given moment. Therefore quantum-register graph is split into operation levels, where all primitives within a single level can be processed in parallel, and the levels – in succession. Qubit matrix due to its regular structure is focused on solving the following problems: 1) Repair of logical primitives in the operation due to readdressing faulty elements on spare primitives (line 3) [2], just as is done in the memory matrix; 2) Index addressing each quantum of the matrix $\mu_{ij} \in \mu$, $\mu_{ij} = (X_{ij}, Q_{ij}, Y_{ij})$ for rapid repair of failed primitives (in the example we can replace three faulty primitives, one of each layer); 3) Providing high performance of combinational unit prototype based on quantum primitives implemented in PLD LUTs [2] due to parallel processing primitives of a single layer; 4) Developing a matrix quantum multi-processor, focused on synthesis of hardware prototypes of combinational units of large dimension to significantly speed up testing and verification of digital systems on chips like Aldec Hardware Embedded Simulator (HES) [2, 5]; 5) Developing methods of analysis and synthesis of combinational circuits, focused to matrix realizing quantum structures of logic elements by means of their implementation in PLD memory elements; 6) Developing a code generator for implementing the quantum matrix of the combinational circuit in the structure of PLD circuit primitives; 7) Designing a control automaton for functional processing and repairing the quantum matrix of combinational unit implemented in PLD structure.

The model of control automaton for simulating qubit structure of the combinational circuit involves three items:

1. The initiation of the next input action for combinational unit.
2. Selection of the next layer (matrix column) with the number i for parallel processing qubit primitives Q to form output states at the address of an input word represented by the vector $M(X_{ij})$, where X_{ij} is a vector of numbers of input variables for the primitive Q_{ij} , M is modeling vector for all lines of combinational unit:

$$M(Y_{ij}) = Q_{ij}[M(X_{ij})], j = \overline{1, q}.$$
3. Incrementing the index column $i=i+1$ and going to the item 2 for processing the next layer of qubit primitives. After the analysis of all the columns of the matrix $i = p$ incrementing index of the next input pattern $t=t+1$ is performed, and subsequent going to

the item 1. When reaching a finite number of input patterns $t = n_{\max}$ the loop for processing test of the qubit matrix ends.

The hardware implementation of the quantum structure of the digital devices, based on the use of memory elements, is shown in Fig. 4. The structure of the circuit contains the following variables and functional elements: input is designed for serial entering input values of vector M; rst – general reset of the system (in this case for counters); clk – sync input; counter of inputs – counter for filling the input coordinates of the vector M; counter of element – counter of processed primitive number, which provides two cycles for reading the input set of two coordinates of the vector M; Q[3:1] – bus for number of processed primitive; Q[0] – variable for mode of reading input value from the vector M or writing the result to M. Memory: Ram 8x4 output – stores the number of primitive output lines; Ram 8x4 input 1 and Ram 8x4 input 2 – store the numbers of primitive input lines. Ram 16x1d – dual-port memory for storing the modeling vector M, where addr0 – address of the input 1 when the value 00 appears on control inputs of the multiplexer, address of writing result when the value 01 appears on control inputs of the multiplexer, address for initializing input data when the value 1X appears on control inputs of the multiplexer; addr1 – address of input 2 for processed primitive; di0 – memory data input when processing primitive (MUX=1) or external input when initializing input data (MUX=0); we – permission of writing in the vector M; do0 – output, corresponding to the input addr0; do1 – output, corresponding to the input addr1. RAM 32x1 is designed for storing Q-vectors defined functionalities of combinational circuit: di – data input that can be used to initialize (write) the structure of quanta; addr – [4:0]; addr[4:2] – element number, addr[1:0] – input set for a primitive.

The complexity of hardware implementation of the combinational circuit is 150 gates, which include 20 LUTs of Xilinx Spartan 3E element system. The speed of the operation or generation of the modeling vector is 180 ns.

4. CONCLUSION

The qubit structure of combination devices provides an opportunity to make a simple automaton from the combinational circuit (integrating memory, functionality quanta, transaction operation) and move from the software simulation of digital systems for the hardware emulation of structures and processes, which are invariant with respect to implementation technologies. An analogue and prototype is Dr. Stanley Hyduke's hardware accelerator of simulation processes PRUS (Aldec Inc.) [2, 5] focused on reducing the time of the design and verification of digital systems on chips. But it is proposed to use the processor for soft (address based) hardware simulating qubit structures for the direct functional purpose as a computing product, delivering services to the consumer. This maintains the high speed operation of the device, supplemented by an opportunity of repairing in real time that is important for critical system.

1. Qubit models for describing digital systems and components are proposed, which are characterized by compactness of description of the truth tables in the form of Q-coverages due to the unitary encoding input states, which gives an opportunity to improve the performance of software and hardware applications for the interpretative simulation of computing devices due to address analysis of logic primitives.
2. The matrix model of qubit primitives for implementing combinatorial circuits is represented; it is characterized by address combining of Q-coverages by using memory elements, which are softly connected in the digital circuit through the state vector of lines that enables repairing faulty logic primitives in real-time by way of re-addressing them on spare components at sufficiently high speed operation of the computing device.

- An innovative idea of quantum computation is described, which is characterized by the transition from the computational procedures with byte operand, defining a single solution (point) in a discrete space, to the logical register parallel processes with the qubit operand, simultaneously generating a power set of solutions that enables to define new perspectives of creating high-performance computers for parallel analyzing and synthesizing discrete structures and services in cyberspace.

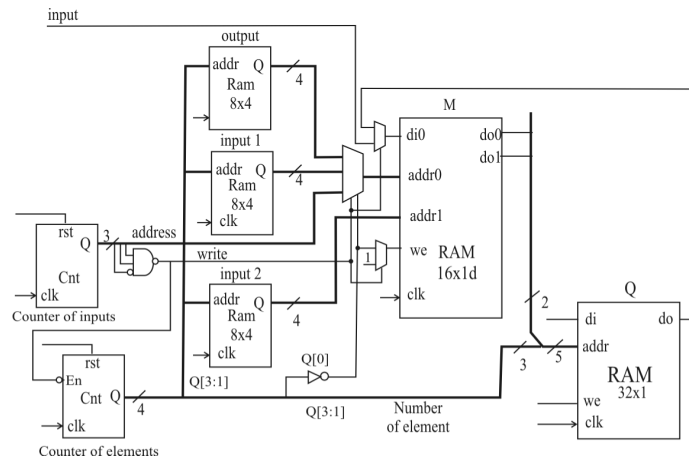


Figure 4. Hardware implementation of qubit structure of combinational circuit

REFERENCES

- [1] Michael A. Nielsen & Isaac L. Chuang, (2010) Quantum Computation and Quantum Information, Cambridge University Press.
- [2] Hahanov V.I., (2009) Digital System-on-Chip Design and Test, Kharkov: Novoye Slovo.
- [3] Hahanov V.I., (1995) Technical diagnosis of digital and microprocessor structures, K.: ICIO.
- [4] Gorbatov V.A. (1986) Fundamentals of Discrete Mathematics, M.: Vysshaya Shkola.
- [5] Hahanov V.I., Murad Ali Abbas, Litvinova E.I., Guz O.A., Hahanova I.V., (2011) "Quantum models of computing processes", Radioelectronics & Informatics, No. 3, pp35 40.

AUTHORS

Vladimir Hahanov – Professor, Doctor of Science, IEEE Senior Member, IEEE Computer Society Golden Core Member, Dean of Computer Engineering Faculty, Kharkov National University of Radioelectronics, Ukraine.



Wajeb Gharibi – PhD, Professor of Jazan University, Kingdom of Saudi Arabia.

Svetlana Chumachenko – Professor, Doctor of Science, IEEE Member, Head of Department "Computer Aided Design of Computers", Kharkov National University of Radioelectronics, Ukraine.



Eugenia Litvinova – Professor, Doctor of Science, IEEE Member, Vice Dean of Computer Engineering Faculty, Kharkov National University of Radioelectronics, Ukraine.



INTENTIONAL BLANK

ARABIC TWEETS CATEGORIZATION BASED ON ROUGH SET THEORY

Mohammed Bekkali and Abdelmonaime Lachkar

L.S.I.S, E.N.S.A, University Sidi Mohamed Ben Abdellah (USMBA),
Fez, Morocco

bekkalimohammed@gmail.com, abdelmonaime_lachkar@yahoo.fr

ABSTRACT

Twitter is a popular microblogging service where users create status messages (called "tweets"). These tweets sometimes express opinions about different topics; and are presented to the user in a chronological order. This format of presentation is useful to the user since the latest tweets from are rich on recent news which is generally more interesting than tweets about an event that occurred long time back. Merely, presenting tweets in a chronological order may be too embarrassing to the user, especially if he has many followers. Therefore, there is a need to separate the tweets into different categories and then present the categories to the user. Nowadays Text Categorization (TC) becomes more significant especially for the Arabic language which is one of the most complex languages.

In this paper, in order to improve the accuracy of tweets categorization a system based on Rough Set Theory is proposed for enrichment the document's representation. The effectiveness of our system was evaluated and compared in term of the F-measure of the Naïve Bayesian classifier and the Support Vector Machine classifier.

KEYWORDS

Arabic Language, Text Categorization, Rough Set Theory, Twitter, Tweets.

1. INTRODUCTION

Twitter is a popular micro-blogging service where users search for timely and social information. As in the rest of the world, users in Arab countries engage in social media applications for interacting and posting information, opinions, and ideas [1]. Users post short text messages called tweets, which are limited by 140 characters [2] [3] in length and can be viewed by user's followers. These tweets sometimes express opinions about different topics; and are presented to the user in a chronological order [4]. This format of presentation is useful to the user since the latest tweets are generally more interesting than tweets about an event that occurred long time back. Merely, presenting tweets in a chronological order may be too embarrassing to the user, especially if he has many followers [5] [6]. Therefore, there is a great need to separate the tweets into different categories and then present the categories to the user. Text Categorization (TC) is a good way to solve this problem.

Text Categorization Systems try to find a relation between a set of Texts and a set of categories (tags, classes). Machine learning is the tool that allows deciding whether a Text belongs to a set

of predefined categories [6]. Several Text Categorization Systems have been conducted for English and other European languages, yet very little researches have been done out for the Arabic Text Categorization [7]. Arabic language is a highly inflected language and it requires a set of pre-processing to be manipulated, it is a Semitic language that has a very complex morphology compared with English. In the process of Text Categorization the document must pass through a series of steps (Figure.1): transformation the different types of documents into brut text, removed the stop words which are considered irrelevant words (prepositions and particles); and finally all words must be stemmed. Stemming is the process consists to extract the root from the word by removing the affixes [8] [9] [10] [11] [12] [13] [14]. To represent the internal of each document, the document must passed by the indexing process after pre-processing. Indexing process consists of three phases [15]:

- a) All the terms appear in the documents corpus has been stocked in the super vector.
- b) Term selection is a kind of dimensionality reduction, it aims at proposing a new set of terms in the super vector to some criteria [16] [17] [18];
- c) Term weighting in which, for each term selected in phase (b) and for every document, a weight is calculated by TF-IDF which combine the definitions of term frequency and inverse document frequency [19].

Finally, the classifier is built by learning the characteristics of each category from a training set of documents. After building of classifier, its effectiveness of is tested by applying it to the test set and verifies the degree of correspondence between the obtained results and those encoded in the corpus.

Not that, one of the major problems in Text Categorization is the document's representation where we still limited only by the terms or words that occur in the document. In our work, we believe that Arabic Tweets (which are Short Text Messages) representation is challenge and crucial stage. It may impact positively or negatively on the accuracy of any Tweets Categorization system, and therefore the improvement of the representation step will lead by necessity to the improvement of any Text Categorization system very greatly.

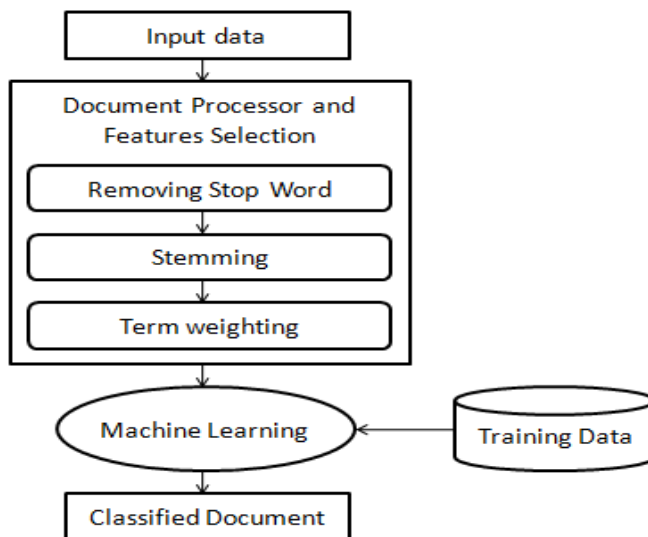


Figure .1 Architecture of TC System

To overcome this problem, in this paper we propose a system for Tweets Categorization based on Rough Set Theory (RST) [20] [21]. This latter is a mathematical tool to deal with vagueness and uncertainty. RST has been introduced by Pawlak in the early 1980s [20], it has been integrated in many Text mining applications such as for features selection [], in this work we proposed to use the Upper Approximation based RST to enrich the Tweet's Representation by using other terms in the corpus with which there is semantic links; it has been successful in many applications. In this theory each set in a universe is described by a pair of ordinary sets called Lower and Upper Approximations, determined by an equivalence relation in the Universe [20].

The remainder parts of this paper are organized as follows: we begin with a brief review on related work in Arabic Tweets Categorization in the next section. Section III presents introduction of the Rough Set Theory and his Tolerance Model; section IV presents two machine learning algorithms for Text Categorization (TC): Naïve Bayesian and Support Vector Machine classifiers used in our system; section V describes our proposed system for Arabic Tweets Categorization; section VI conducts the experiments results; finally, section VII concludes this paper and presents future work and some perspectives.

2. RELATED WORK

A number of recent papers have addressed the categorization of tweets most of them were tested against English Text [4] [30] [31]. Furthermore Categorization Systems that address Arabic Tweets are very rare in the literature [1]. This latter work realized by Rehab Nasser et al. presents a roadmap for understanding Arabic Tweets through two main objectives. The first is to predict tweet popularity in the Arab world. The second one is to analyze the use of Arabic proverbs in Tweets, The Arabic proverbs classification model was labeled "Category" with four class values sport, religious, political, and ideational.

On the other hand a wide range of Text Categorization based Rough Set Theory have been developed most of them were tested against English Text [39] [40]. Concerning Text Categorization Systems based on Rough Set that address Arabic Text is rare in the literature [41].

In Arabic Text Categorization we found Sawaf presented in [32] uses statistical methods such as maximum entropy to cluster Arabic news articles; the results derived by these methods were promising without morphological analysis. In [33], NB was applied to classify Arabic web data; the results showed that the average accuracy was 68.78%. The work of Duwairi [34] describes a distance-based classifier for Arabic text categorization. In [35] Laila et al. compared between Manhattan distance and Dice measures using N-gram frequency statistical technique against Arabic data sets collected from several online Arabic newspaper websites. The results showed that N-gram using Dice measure outperformed Manhattan distance.

Mesleh et al. [36] used three classification algorithms, namely SVM, KNN and NB, to classify 1445 texts taken from online Arabic newspaper archives. The compiled text Automated Arabic Text Categorization Using SVM and NB 125 were classified into nine classes: Computer, Economics, Education, Engineering, Law, Medicine, Politics, Religion and Sports. Chi Square statistics was used for features selection. [36] Discussed that "Compared to other classification methods, their system shows a high classification effectiveness for Arabic data set in terms of F measure (F=88.11)".

Thabtah et al. [37] investigate NB algorithm based on Chi Square feature selection method. The experimental results compared against different Arabic text categorization data sets provided evidence that features selection often increases classification accuracy by removing rare terms. In [38] NB and KNN were applied to classify Arabic text collected from online Arabic newspapers.

The results show that the NB classifier outperformed KNN base on Cosine coefficient with regards to macro F1, macro recall and macro precision measures.

Recently, Hadni et al. team [7] presents an Effective Arabic Stemmer Based Hybrid Approach for Arabic Text Categorization.

Note that, in any Text Categorization system the center point is the document and its representation that may impact positively or negatively on the accuracy of the system.

In the following section we present the Rough Set Theory, its mathematical background and also the Tolerance Rough Set Model which is proposed to deal with Text Representation.

3. ROUGH SET THEORY

3.1. Rough Set Theory

In this section we present Rough Set Theory that has been originally developed as a tool for data analysis and classification [20] [21]. It has been successfully applied in various tasks, such as features selection/extraction, rule synthesis and classification. The central point of Rough Set theory is the notion of set approximation: any set in U (a non-empty set of object called the universe) can be approximated by its lower and upper approximation. In order to define lower and upper approximation we need to introduce an indiscernibility relation that could be any equivalence relation R (reflexive, symmetric, transitive). For two objects $x, y \in U$, if xRy then we say that x and y are indiscernible from each other. The indiscernibility relation R induces a complete partition of universe U into equivalent classes $[x]_R, x \in U$ [22].

We define lower and upper approximation of set X , with regards to an approximation space denoted by $A = (U, R)$, respectively as:

$$L_R(X) = \{x \in U: [x]_R \subseteq X\} \quad (1)$$

$$U_R(X) = \{x \in U: [x]_R \cap X \neq \Phi\} \quad (2)$$

Approximations can also be defined by mean of rough membership function. Given rough membership function $\mu_X: U \rightarrow [0, 1]$ of a set $X \subseteq U$, the rough approximation is defined as:

$$L_R(X) = \{x \in U: \mu_X(x, X) = 1\} \quad (3)$$

$$U_R(X) = \{x \in U: \mu_X(x, X) > 0\} \quad (4)$$

Note that, given rough membership function as:

$$\mu_X(x, X) = \frac{|[x]_R \cap X|}{|[x]_R|} \quad (5)$$

Rough Set Theory is dedicated to any data type but when it comes with Documents Representation we use its Tolerance Model described in the next section.

3.2. Tolerance Rough Set Model

Let $D = \{d_1, d_2, \dots, d_n\}$ be a set of document and $T = \{t_1, t_2, \dots, t_m\}$ set of index terms for D . with the adoption of the vector space model, each document d_i is represented by a weight vector $\{w_{i1}, w_{i2}, \dots, w_{im}\}$ where w_{ij} denotes the weight of index term j in document i . The tolerance space is defined over a universe of all index terms $U = T = \{t_1, t_2, \dots, t_m\}$ [23].

Let $f_{d_i}(t_i)$ denotes the number of index terms t_i in document d_i ; $f_D(t_i, t_j)$ denotes the number of documents in D in which both index terms t_i and t_j occurs. The uncertainty function I with regards to threshold θ is defined as:

$$I_\theta = \{t_j \mid f_D(t_i, t_j) \geq \theta\} \cup \{t_i\} \quad (6)$$

Clearly, the above function satisfies conditions of being reflexive and symmetric. So $I_\theta(I_i)$ is the tolerance class of index term t_i . Thus we can define the membership function μ for $I_i \in T, X \subseteq T$ as [24]:

$$\mu_X(t_i, X) = v(I_\theta(t_i), X) = \frac{|I_\theta(t_i) \cap X|}{|I_\theta(t_i)|} \quad (7)$$

Finally, the lower and the upper approximation of any document $d_i \subseteq T$ can be determined as:

$$L_R(d_i) = \{t_i \in T: v(I_\theta(t_i), d_i) = 1\} \quad (8)$$

$$U_R(d_i) = \{t_i \in T: v(I_\theta(t_i), d_i) > 0\} \quad (9)$$

Once the documents handling is finished, the results will be the entry of any Text Categorization System. In the following section we present two of the most popular Machine Learning algorithms, Naïve Bayesian and the Vector Machine.

4. BASED MACHINE LEARNING

TC is the task of automatically sorting a set of documents into categories from a predefined set. This section covers two algorithms among the used known Machine Learning Algorithms for TC: Naïve Bayesian (NB) and Support Vector Machine (SVM).

4.1. Naïve Bayesian Classifier

The NB is a simple probabilistic classifier based on applying Baye's theorem, and its powerful, easy and language independent method. [25]

When the NB classifier is applied on the TC problem we use equation (10)

$$p(\text{class} \mid \text{document}) = \frac{p(\text{class}) \cdot p(\text{document} \mid \text{class})}{p(\text{document})} \quad (10)$$

where:

$P(\text{class} \mid \text{document})$: It's the probability that a given document D belongs to a given class C

$P(\text{document})$: The probability of a document, it's a constant that can be ignored

$P(\text{class})$: The probability of a class, it's calculated from the number of documents in the category divided by documents number in all categories

$P(\text{document} \mid \text{class})$: it's the probability of document given class, and documents can be represented by a set of words:

$$p(\text{document} \mid \text{class}) = \prod_i p(\text{word}_i \mid \text{class}) \quad (11)$$

so:

$$p(\text{class} \mid \text{document}) = p(\text{class}) \cdot \prod_i p(\text{word}_i \mid \text{class}) \quad (12)$$

where:

$p(\text{word}_i | \text{class})$: The probability that a given word occurs in all documents of class C , and this can be computed as follows:

$$p(\text{word}_i | \text{class}) = \frac{T_{ct} + \lambda}{N_c + V} \quad (13)$$

where:

T_{ct} : The number of times that the word occurs in that category C .

N_c : The number of words in category C .

V : The size of the vocabulary table.

λ : The positive constant, usually 1, or 0.5 to avoid zero probability.

4.2. Support Vector Machine Classifier

SVM introduced by Vapnik [26] and has been introduced in TC by Joachims [27]. Based on the structural risk minimization principle from the computational learning theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set.

Given a set of N linearly separable points $N = \{x_i \in \mathbb{R}^n \mid i = 1, 2, \dots, N\}$, each point x_i belongs to one of the two classes, labeled as $y_i \in \{-1, 1\}$. A separating hyper-plane divides S into 2 sides, each side containing points with the same class label only. The separating hyper-plane can be identified by the pair (w, b) that satisfies: $w \cdot x + b = 0$

and:

$$\begin{cases} w \cdot x_i + b \geq +1 & \text{if } y_i = +1 \\ w \cdot x_i + b \leq -1 & \text{if } y_i = -1 \end{cases} \quad (14)$$

For $i = 1, 2, \dots, N$; where the dot product operation (\cdot) is defined by:

$$w \cdot x = \sum w_i \cdot x_i$$

For vectors w and x , thus the goal of the SVM learning is to find the Optimal Separating Hyper plane (OSH) that has the maximal margin to both sides. This can be formularized as:

minimize:

$$\frac{1}{2} w \cdot w$$

subject to

$$\begin{cases} w \cdot x_i + b \geq +1 & \text{if } y_i = +1 \text{ for } i = 1, 2, \dots, N \\ w \cdot x_i + b \leq -1 & \text{if } y_i = -1 \end{cases} \quad (15)$$

Figure 2 shows the optimal separating hyper-plane

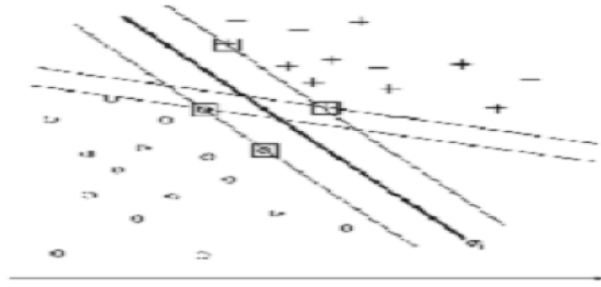


Figure .2 Learning Support Vector Classifier

The small crosses and circles represent positive and negative training examples, respectively, whereas lines represent decision surfaces. Decision surface σ_i (indicated by the thicker line) is, among those shown, the best possible one, as it is the middle element of the widest set of parallel decision surfaces (i.e., its minimum distance to any training example is maximum). Small boxes indicate the Support Vectors.

During classification, SVM makes decision based on the OSH instead of the whole training set. It simply finds out on which side of the OSH the test pattern is located. This property makes SVM highly competitive, compared with other traditional pattern recognition methods, in terms of computational efficiency and predictive accuracy [28].

The method described is applicable also to the case in which the positives and the negatives are not linearly separable. Yang and Liu [28] experimentally compared the linear case (namely, when the assumption is made that the categories are linearly separable) with the nonlinear case on a standard benchmark, and obtained slightly better results in the former case.

5. THE PROPOSED SYSTEM FOR ARABIC TWEETS CATEGORIZATION

In this section we present in detail our proposed system for Arabic Tweets Categorization. The proposed system contains two main components, the first component generates the Upper Approximation for each Tweet to extend the Tweet's Representation by taking into account not just their words but also the words with which there is semantic links (Figure 3); the second one does the categorization using the Naïve Bayesian and the Support Vector Machine.

The treatment goes through the following steps: each Tweet in the corpus will be cleaned by removing Arabic stop words, Latin words and special characters like (/, #, \$, ect...). After that for each word in corpus we make the following operations:

- Apply a stemmer algorithm to generate the root and eliminate the redundancy. too many algorithms have been proposed in this topic such as Khoja [11] and Light Stemmer [13] which are both the most known algorithms for Arabic text preprocessing; we used in this stage Khoja as a stemmer algorithm.
- Calculate the frequency in the document and also in the hole corpus.
- Determine the tolerance class of terms which contains all the words that occur with our word in the same document a number of times upper than θ . This tolerance class is defined using the formula (6).

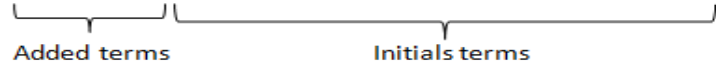
After these operations we can calculate the approximations for each document by using the formula (8) for the lower and the formula (9) for the upper approximation. Then term weighting in which, for each term a weight is calculated by TF-IDF which combine the definitions of term frequency and inverse document frequency.

Tweet's terms (Arabic/English) before using RS :

شقيقان, تم, تروي, فقير, الآخر, للمرة, أحدهما, عرض
Two brothers, were, rich, poor, the other,
for once, one of them, show

Tweet's terms (Arabic/English) after using Rough Set :

شقيقان, تم, تروي, فقير, الآخر, للمرة, أحدهما, عرض, السينما, الأولى
cinema, first Two brothers, were, rich, poor, the other,
for once, one of them, show, cinema, first



Added terms Initials terms

Figure .3 Example of Rough Set Application

To illustrate the semantics links discovered by the Upper Approximation, Figure 3 presents an example of Arabic Tweet after the pre-processing; the initial tweet contains the word show / عرض we saw that the word cinema / السينما was added to the generated Upper Approximation because the two words are semantically related and often found with each other.

Finally, the classifier is built by learning the characteristics of each category from a training set of Tweets. After building of classifier, its effectiveness is tested by applying it to the test set and verifies the degree of correspondence between the obtained results and those encoded in the corpus.

6. EXPERIMENTS RESULTS

To illustrate that our proposed method can improve the Tweet's representation by using the Upper Approximation of RST and therefore can enhance the performance of our Arabic Tweets Categorization System. In this section a series of experiments has been conducted.

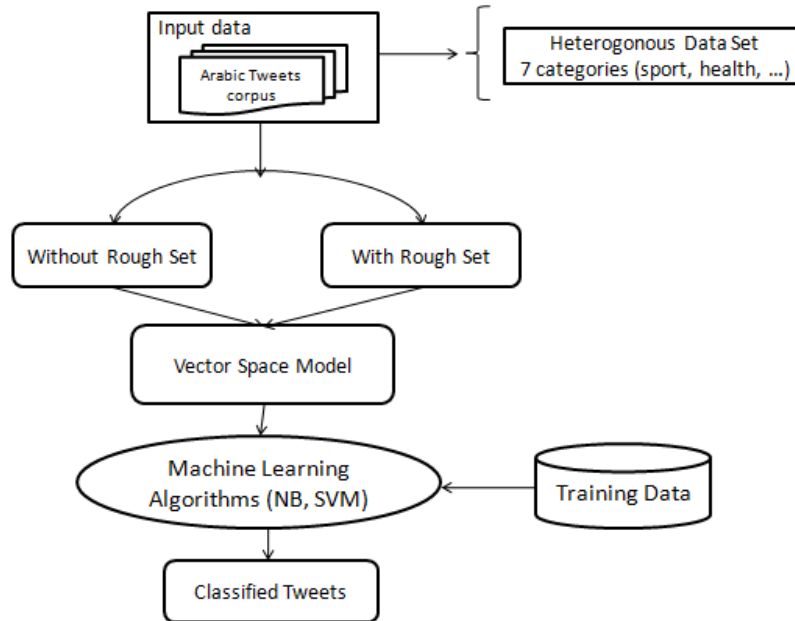


Figure .4 Descriptions of Our Experiments

Figure 4 describes our experiment with and without applying the Upper Approximation based Rough Set Theory.

The data set used in our experiments is collected from Twitter by using *NodeXL Excel Template* which is a freely Excel template that makes it super easier to collect Twitter network data [29]. This corpus is manually classified into six categories (Table 1). These categories are: Cinema/السينما, News/الأخبار, Documentary/وثائقي, Health/الصحة, Tourism/السياحة and Economics/الاقتصاد. Table 2 contains some examples of Tweets collected from Twitter.

The dataset has been divided into two parts: training and testing. The training data consist of 70% the documents per category. The testing data, on the other hand consist of 30% the documents of each category.

Table 1. The 6 categories and the number of Tweets in each one

Category	Number of Tweets
Cinema/السينما	79
Documentary/وثائقي	64
Economy/الاقتصاد	63
Health/الصحة	71
News/الأخبار	90
Tourism/السياحة	83
Total/المجموع	450

To assess the performance of the proposed system, a series of experiments has been conducted. The effectiveness of our system has been evaluated and compared in term of the F1-measure using the NB and the SVM classifiers used in our TC system.

F1-measure can be calculated using Recall and Precision measures as follow:

$$F1\text{-measure} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad (16)$$

Precision and Recall both are defined as follows:

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (17)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (18)$$

where:

- True Positive (TP) refers to the set of Documents which are correctly assigned to the given category.
- False Positive (FP) refers to the set of documents which are incorrectly assigned to the category.
- False Negative (FN) refers to the set of documents which are incorrectly not assigned to the category.

Table 2. Examples of the Tweets in the corpus

Category	Tweets in Arabic	Tweets in English
Cinema/السينما	تغرق في عالم يكتنفه الجنون و القتل و القوى الخارقة للطبيعة	Drowning in a world beset by madness, murder and supernatural powers
	القصة الحقيقية لمخطط اغتيال الرئيس ريتشارد نيكسون	The real story of the assassination's plot of the President Richard Nixon
Documentary/ وثائقي	تحت المجهر "حرب السدود": نهر النيل الآن على الجزيرة الوثائقية	Under the microscope "war dams": the Nile River is now on Al Jazeera Documentary
	بقي 6 أيام و يعرض البرنامج العالمي "الكون الكبير" على قناة ناشونال جيوغرافيك : أبو ظبي	6 days left and the global program the "big universe" will be displayed on the National Geographic Abu Dhabi Channel
Economy/ الاقتصاد	صعود الذهب إلى أعلى مستوياته ثلاثة أسابيع ونصف بعد موافقة واشنطن على توجيه جوية على العراق	The rise of gold to its highest level, three and a half weeks after the approval of Washington to direct flights on Iraq
	المبيعات الإجمالية لشركات الاسمنت تتراجع بنسبة 2% مقارنة بالشهر الماضي	Total sales of cement companies falls by 2% compared to last month
Health/الصحة	ذكرنا أعراض و علامات السكتة القلبية في تغريدتنا السابقة في تمام الساعة 10 صباحا	We mentioned symptoms and signs of heart attack in the previous Tweet at 10:00
	أخصائية تغذية لعمل نظام غذائي مرتفع السعرات ومناسب لك	Dietitian makes a high-calorie diet and suitable for you

الأخبار/News	سقوط قذيفة هاون افضى إلى إصابة محافظ الأنبار بجروح بليغة	A mortar shell fell led to the injury of the governor of Anbar by a seriously injuries
	حوارات و تقارير: في تونس سباق محموم للظفر بمنسب الرئيس	Dialogues and reports: in Tunisia a frantic race to win the post of president
السياحة/Tourism	الأماكن السياحية في تركيا: التفاصيل	Tourist places in Turkey: Details
	منتجع شاموني يعتبر أحد أكثر المنتجعات شهرة في فرنسا	Resort of Chamonix is one of the most famous resorts in France

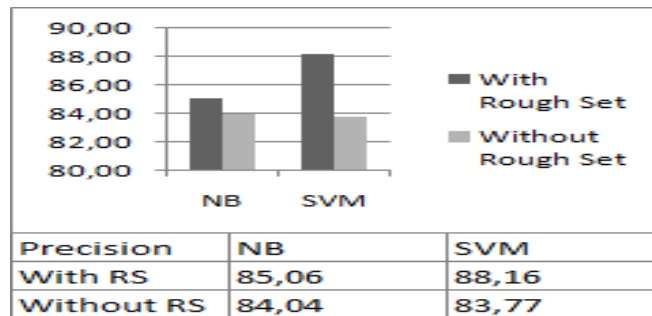


Figure .5 Representation of the Precision's Average

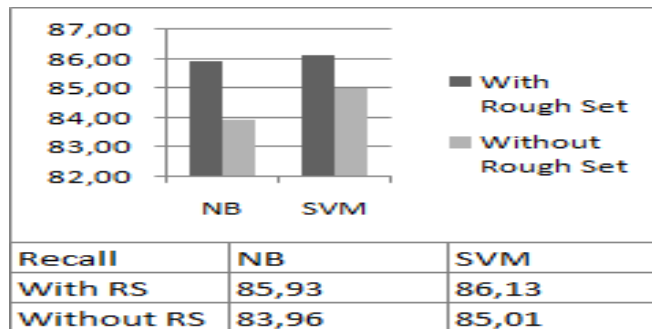


Figure .6 Representation of the Recall's Average

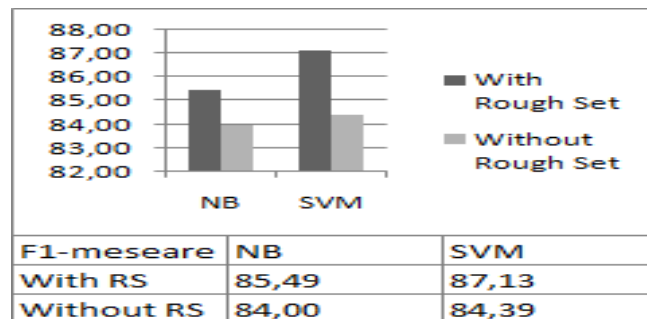


Figure 7. Representation of the F1-measure's Average

Figure 7 shows the obtained F1-measure results with and without using the Rough Set Theory in our Arabic Tweets Categorization System. These results illustrate that using the Rough Set Theory enhances greatly the performance of Arabic Tweets Categorization.

Using the SVM classifier, applying RST performed better results for the five classes: Cinema (86, 4), Documentary (91, 9), Economy (90, 98), News (87, 52) and Tourism (88, 51) compared to the Categorization without using the RST: Cinema (85, 6), Documentary (84, 4), Economy (85, 35), News (85, 45) and Tourism (85, 54). But an average F1-measure the 87, 13 % with using the RST compared to 84, 39 % without using the RST.

To validate our results, we used another classifier which is the NB and the results was 85, 49 % with applying the RST compared to 84 % without using the RST.

The precision and the recall results showed in Figure 5 respectively in Figure 6 illustrate also that using the RST influence positively in the process of Arabic Tweets Categorization by enriching the Tweet representation with others terms that not occurred in and they have some semantic links with the Tweet's terms.

7. CONCLUSION AND FUTURE WORK

Tweets Categorization becomes an interest topic in recent years especially for the Arabic Language. Tweets Representation plays a vital role and may impact positively or negatively on the performance of any Tweets Categorization System. In this paper, we have proposed an effective method for Tweets Representation based on Rough Set Theory. This latter enriches and adds other terms which are semantically related with the original terms existing in the original Tweets. The proposed method has been integrated and tested for Arabic Tweets Categorization using NB and SVM classifiers.

The obtained results show that using the Upper Approximation of the Rough Set Theory increases significantly the F1-measure of the Tweets Categorization Systems.

In our future work, we will focus on using an external resource like Arabic WordNet or Arabic Wikipedia to add more semantic links between terms in Tweets Representation step.

REFERENCES

- [1] Rehab Nasser, Al-Wehaibi*, Muhammad Badruddin, Khan "Understanding the Content of Arabic Tweets by Data and Text Mining Techniques", Symposium on Data Mining and Applications (SDMA2014)
- [2] K. Lee, D. Palsetia, R. Narayanan, Md Patwary, A. Agrawal, A. Choudhary. "Twitter Trending Topic Classification", 11th IEEE International Conference on Data Mining Workshops 2011, 978-0-7695-4409-0/11
- [3] A. Go, R. Bhayani, L. Huang. "Twitter Sentiment Classification using Distant Supervision", Processing (2009), S. 1—6
- [4] Bharath Sriram. "Short Text Classification In Twitter To Improve Information Filtering". Computer Science and Engineering. Ohio State University, 2010
- [5] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu. "Short Text Classification in Twitter to Improve Information Filtering", SIGIR'10, July 19–23, 2010, Geneva, Switzerland. ACM 978-1 60558-896-4/10/07.
- [6] Sebastiani F. "Machine learning in automated text categorization". ACM Computing Surveys, volume 34 number 1. PP 1-47. 2002.
- [7] M.Hadni, A.Lachkar, S. Alaoui Ouatik "Effective Arabic Stemmer Based Hybrid Approach for Arabic Text Categorization", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013

- [8] Al-Fedaghi S. and F. Al-Anzi. "A new algorithm to generate Arabic root-pattern forms". In proceedings of the 11th national Computer Conference and Exhibition. PP 391-400. March 1989.
- [9] Al-Shalabi R. and M. Evens. "A computational morphology system for Arabic". In Workshop on Computational Approaches to Semitic Languages, COLING-ACL98. August 1998.
- [10] Aljlal M. and O. Frieder. "On Arabic search: improving the retrieval effectiveness via a light stemming approach". Proceedings of ACM CIKM 2002 International Conference on Information and Knowledge Management, McLean, VA, USA, 2002, pp. 340-347.
- [11] Chen A. and F. Gey. "Building an Arabic Stemmer for Information Retrieval". In Proceedings of the 11th Text Retrieval Conference (TREC 2002), National Institute of Standards and Technology. 2002.
- [12] Khoja S.. "Stemming Arabic Text". Lancaster, U.K., Computing Department, Lancaster University. 1999.
- [13] Larkey L. and M. E. Connell. "Arabic information retrieval at UMass in TREC-10". Proceedings of TREC 2001, Gaithersburg: NIST. 2001.
- [14] Larkey L., L. Ballesteros, and M. E. Connell. "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co occurrence Analysis". Proceedings of SIGIR'02. PP 275-282. 2002.
- [15] Sebastiani F. "A Tutorial on Automated Text Categorisation". Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence. PP 7-35. 1999.
- [16] Liu T., S. Liu, Z. Chen and Wei-Ying Ma. "An Evaluation on Feature Selection for Text Clustering". Proceedings of the 12th International Conference (ICML 2003), Washington, DC, USA. PP 488-495. 2003.
- [17] Rogati M. and Y. Yang. "High-Performing Feature Selection for Text classification". CIKM'02, ACM. 2002.
- [18] Yang Y., and J. O. Pedersen. "A comparative study on feature selection in text categorization". Proceedings of ICML-97. PP 412-420. 1997.
- [19] Aas K. and L. Eikvil. "Text categorisation: A survey", Technical report, Norwegian Computing Center. 1999.
- [20] Pawlak, Z. Rough sets: Theoretical aspects of reasoning about data. Kluwer Dordrecht, 1991.
- [21] Jan Komorowski, Lech Polkowski, Andrzej Skowron "Rough Sets: A Tutorial"
- [22] Ngo Chi Lang "A tolerance rough set approach to clustering web search results"
- [23] Jin Zhang and Shuxuan Chen "A study on clustering algorithm of Web search results based on rough set", Software Engineering and Service Science (ICSESS), 2013
- [24] Ngo Chi Lang, "A tolerance rough set approach to clustering web search results", Poland: Warsaw University, 2003.
- [25] Saleh Alsaleem, "Automated Arabic Text Categorization Using SVM and NB", International Arab Journal of e-Technology, Vol. 2, No.2, June 2011.
- [26] Vapnik V. (1995). The Nature of Statistical Learning Theory, chapter 5. Springer-Verlag, New York.
- [27] Joachims T. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," In Proceedings of the European Conference on Machine Learning (ECML), 1998, pp.173-142, Berlin
- [28] Yang Y. and X. Liu, "A re-examination of text categorization methods," in 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), pp. 42-49, 1999.
- [29] <http://social-dynamics.org/twitter-network-data/>
- [30] B. Sriam, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, pp. 841-842.
- [31] D. Antenucci, G. Handy, A. Modi, M. Tinkerhess "Classification Of Tweets Via Clustering Of Hashtags", EECS 545 FINAL PROJECT, FALL, 2011
- [32] Sawaf, H. Zaplo, J. and Ney. H. "Statistical Classification Methods for Arabic News Articles". Arabic Natural Language Processing, Workshop on the ACL, 2001. Toulouse, France.
- [33] El-Kourdi, M., Bensaid, A., and Rachidi, T. "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," 20th International Conference on Computational Linguistics, 2004, Geneva
- [34] Duwairi R., "Machine Learning for Arabic Text Categorization," Journal of the American Society for Information Science and Technology (JASIST), vol. 57, no. 8, pp. 1005-1010, 2005.
- [35] Laila K. "Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study," DMIN, 2006, pp. 78-82.
- [36] Mesleh, A. A. "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System," Journal of Computer Science (3:6), 2007, pp. 430-435.

- [37] Thabtah F., Eljinini M., Zamzeer M., Hadi W. (2009) Naïve Bayesian based on Chi Square to Categorize Arabic Data. In proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt 4 - 6 January. (pp. 930-935).
- [38] Hadi W., Thabtah F., ALHawari S., Ababneh J. (2008b) Naive Bayesian and K-Nearest Neighbour to Categorize Arabic Text Data. Proceedings of the European Simulation and Modelling Conference. Le Havre, France, (pp. 196-200), 2008
- [39] Y. Li, S.C.K. Shiu, S.K. Pal, J.N.K. Liu, "A rough set-based case-based reasoned for text categorization", International Journal of Approximate Reasoning 41 (2006) 229–255, 2005
- [40] W. Zhao, Z. Zhang, "An Email Classification Model Based on Rough Set Theory", Active Media Technology 2005, IEEE
- [41] Yahia, M.E. "Arabic text categorization based on rough set classification", Computer Systems and Applications (AICCSA), 2011 IEEE

VARIABLE LENGTH KEY BASED VISUAL CRYPTOGRAPHY SCHEME FOR COLOR IMAGE

Akhil Anjekar¹, Prashant Dahiwal², Suchita Tarare³

¹Department of Information technology,
Rajiv Gandhi college of Engineering & Research, Nagpur, India.
akhil.anjekar09@gmail.com

²Department of Computer Sci. & Engg.,
Rajiv Gandhi college of Engineering & Research, Nagpur, India.
prashant.dahiwal@gmail.com

³Department of Computer Sci. & Engg.,
Rajiv Gandhi college of Engineering & Research, Nagpur, India.
suchitatarare@gmail.com

ABSTRACT

Visual Cryptography is a special encryption technique that encrypts the secret image into n number of shares to hide information in images in such a way that it can be decrypted by the human visual system. It is imperceptible to reveal the secret information unless a certain number of shares (k) or more are superimposed. Simple visual cryptography is very insecure.

Variable length key based visual cryptography for color image uses a variable length Symmetric Key based Visual Cryptographic Scheme for color images where a secret key is used to encrypt the image and division of the encrypted image is done using Random Number. Unless the secret key, the original image will not be decrypted. Here secret key ensures the security of image.

This paper describes the overall process of above scheme. Encryption process encrypts the Original Image using variable length Symmetric Key, gives encrypted image. Share generation process divides the encrypted image into n number of shares using random number. Decryption process stacks k number of shares out of n to reconstruct encrypted image and uses the same key for decryption.

KEYWORDS

Visual Cryptography, Secret Sharing, Random Number, Symmetric Key.

1. INTRODUCTION

Cryptography is study of mathematical technique to provide the methods for information security. It provides such services like authentication, data security, and confidentiality. Visual cryptography is one of the techniques used in modern world to maintain the secret message transmission. In this technique no need of any cryptographic algorithms like symmetric (DES, AES, TRIPLE DES etc) and asymmetric (RSA, Diffie- Hellman, Elliptic Curve Cryptographic) algorithms. Noar and Shamir introduce visual cryptography in 1994 [2]. This technique is used to

reduce complexity of encrypted and decrypted method and also two way communication can be achieved very securely. Traditional techniques use private and public key concepts. But it could be achieved only by the distribution of keys [7].

Until the year 1997 visual cryptography schemes were applied to only black and white images. First colored visual cryptography scheme was developed by Verheul and Van Tilborg. Image is a multimedia component sensed by human perception. A color digital image is composed of a finite number of elements called pixels. In a 24 bit digital image each pixel consists of 24 bits, which includes three parts, namely Red, Green and Blue, each with 8 bits [1][2].

Human visual system acts as an OR function. If shares are printed on transparencies and stack together then anyone can visualize the image. To make it more secure we are using variable length symmetric key. Fixed length key can be easily computed by combination of characters by the attacker. For variable length key, it is difficult to find the key as the length can be 0 to any number.

A Key is used to provide more security so that attacker cannot retrieve the secret information without the key. Original image is encrypted using key and produces cipher. Cipher is decrypted using key and the original image is retrieved. Same key is used for encryption and decryption called symmetric encryption.

2. LITERATURE REVIEW

Visual cryptography proposed by Naor and Shamir where encryption of image means the generation of shares without any cryptographic computation. Original image is divided into n number of shares by applying any k - n secret sharing visual cryptographic scheme. Decryption is done by human visual system means if shares are printed on transparencies and stack together then anyone can visualize the image. So, if anyone get some number of shares can easily decrypt the image. Simple visual cryptography is not very secure technique [1].

Watermarking using visual cryptography where original image is divided into shares, with k - n secret sharing visual cryptography scheme. An enveloping technique is proposed where the secret shares are enveloped within apparently innocent covers of digital pictures using LSB replacement digital watermarking. This adds security to visual cryptography technique from illicit attack as it befools the hacker's eye. K - n secret sharing process is simple as random number is used. Shares contain the original image contents, if anyone get shares then original image can be obtained [10].

The shares are enveloped into apparently innocent cover of digital pictures and can be sent through same or different communication channels. Invisible digital watermarking befools the hacker. Watermarking is a technique to put a signature of the owner within the creation. As shares are generated from the original image this scheme does not provide more security [2].

3. PROPOSED WORK

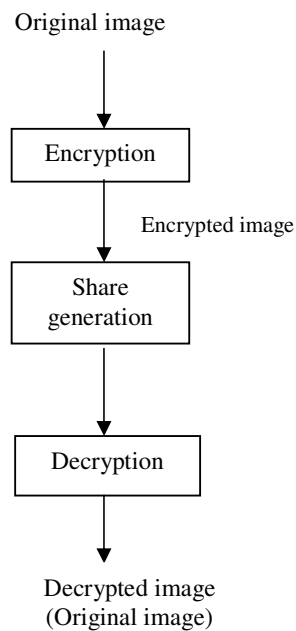


Figure 1. Overall process

3.1. Module – 1

Image encryption using secret key

Original image is encrypted using key. A user generated any combination of characters of varying length gives a key. Generated key and original image are taken as input. Pixel array is computed from original image and key is XOR ed with pixel array to give encrypted image. The contents of original image and encrypted image are totally different, this process makes encrypted image blur to some extent and provide security.

- Take original image as a input; calculate width and height

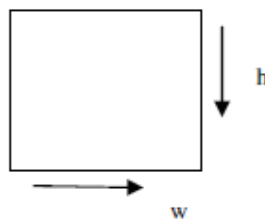


Figure 2. Original image

- Convert each pixel into 24-bit binary , so size of image is $(w*h*24)$

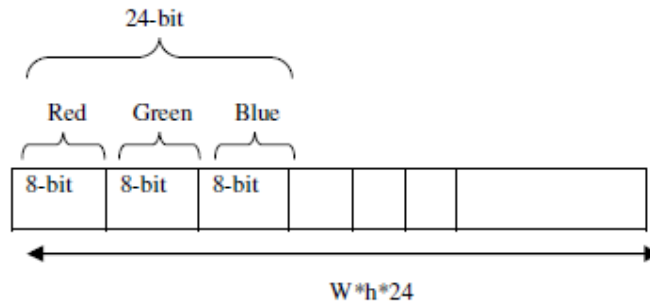


Figure 3. 24-bit converted image

- Enter key from user and calculate length also calculate 7 bit binary string
Let key is: **abcdef** length of key is: 6

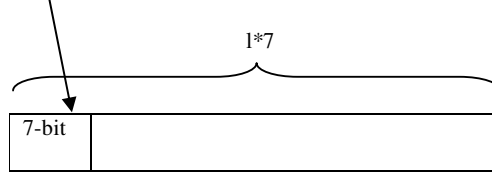


Figure 4. 7-bit binary String Key

- XOR 24-bit converted image and 7-bit binary string key

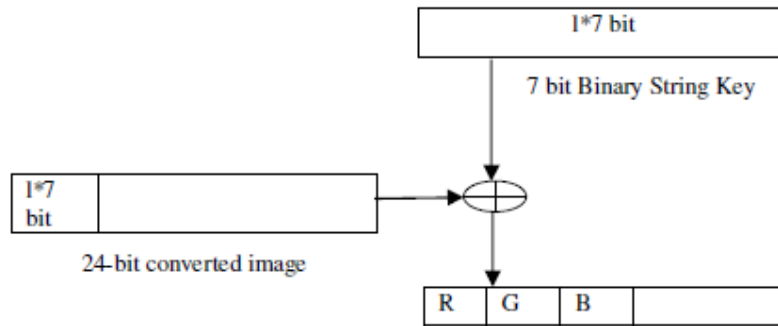


Figure 5. 24-bit encrypted image

- Now the 24-bit encrypted image is reconstructed to get Encrypted image of size equal to original image size

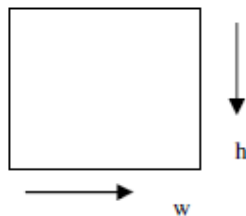


Figure 6. Encrypted image

3.2. Module - 2

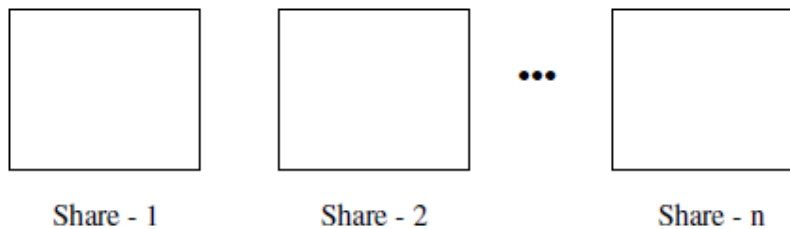
Division of encrypted image

The encrypted image is then divided into n number of shares using k-n secret sharing visual cryptography scheme i.e. using Random Number such that the size of shares equals original image size. K number of shares is sufficient to reconstruct the encrypted image. k number of shares produced is stacked together to reconstruct the encrypted image. Decryption is impossible until the k number of shares are available.

- Enter the number of shares you want to create, suppose n and shares required for reconstruction are k

Calculate $recons = (n-k)+1$

- N number of shares equal to 24-bit converted image size will be created.
- Take the encrypted image as input and convert it into 24-bit encrypted image.
- Scan each bit of 24-bit encrypted image and check for bit 1, if bit is 1 then Random number generator will generate different numbers in the range 1 to n, (numbers generated will be equal to recons).
- 1 is put in generated shares at the same position as in 24-bit encrypted image.
- The same procedure is followed until total bits are scanned.
- Then all the shares are reconstructed to make it equal to original image size.



3.3. Module – 3

Image decryption using secret key

The decryption process consists of two steps. First step is done by human visual system where at least k number of shares out of n number of shares is superimposed to give reconstructed image. Human visual system acts as an OR function. For computer generated process, OR function can be used for the case of stacking k number of shares out of n. Second step is decryption of reconstructed image, where pixel array is computed from reconstructed image and XOR ed with same key used for encryption. Decrypted image is exactly equal to original image.

- Input the number shares you have and the same key used for encryption. Shares should be equal to k or greater than k
- Perform the bitor operation on converted shares to ger reconstructed encrypted image.
- Now XOR the reconstructed encrypted image and converted key to get 24-bit decrypted image, it then reconstructed to give decrypted image equal to original image.

3. RESULT

3.1. Encryption Process:

Original Image: onion.png
Source image is



Figure 7. Original Image

Secret Key is : **testing**

The Encrypted Image :

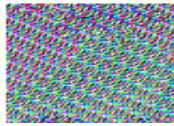


Figure 8. Encrypted Image

3.2. Division of image into number of shares

Number of Shares (n): 6

Numbers of shares for reconstruction (k): 4

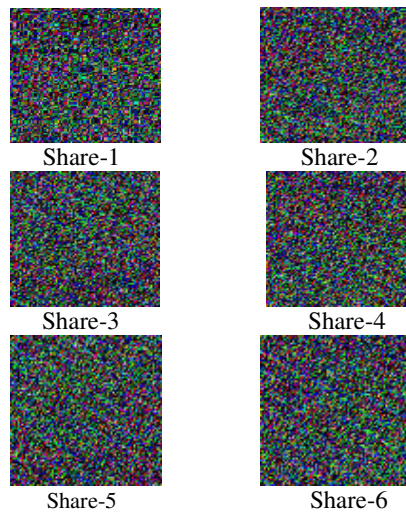


Figure 9. Image shares produced after applying k-n Visual Cryptography

3.3. Decryption Process

Number of shares taken for reconstruction: 4

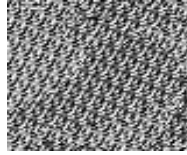


Figure 10. Reconstructed Image

Secret Key is: **testing**

Key is applied on reconstructed image. The Final image is:



Figure 11. Decrypted Image

4. CONCLUSION

In visual cryptography encryption means the generation of shares and decryption in is based on OR operation, so if a person gets sufficient k number of shares the image can be easily decrypted. So simple visual cryptography is not more secure.

In this paper we have proposed a variable length key based visual cryptography for color image with random number for share generation. In this scheme key adds robustness to the visual cryptography techniques and variable length of the key makes it more secure. Generated shares have totally different information regards to original image. For share generation we are using random number which needs very less mathematical calculation compare with other existing techniques of visual cryptography on color images [3][4][5]. As we are using variable length key for encryption and random number generator for share generation this process is more secure than other visual cryptography schemes [8].

Table 1. Comparison of other processes with Proposed Scheme

Other processes	Proposed scheme
Share generation process is applied directly on original image.	Share generation process is applied on encrypted image.
Generated shares contain the original image contents.	Generated shares have totally different contents.
Do not provide more security.	Use of key makes it more secure.
Share generation process is complex.	Share generation process is simple as random number is used.
Decryption is done by OR operation.	Decryption is done by OR as well as XOR operation.

REFERENCES

- [1] M. Naor and A. Shamir, "Visual cryptography," *Advances in Cryptology-Eurocrypt'94*, pp.1–12, 1995.
- [2] S. Craver, N. Memon, B. L. Yeo, and M. M. Yeung. Resolving Rightful Ownerships with Invisible Watermarking Techniques: Limitations, Attacks and Implications. *IEEE Journal on Selected Areas in Communications*, Vol16, No.4 May 1998, pp.573–586.
- [3] F. Liu¹, C.K. Wu¹, X.J. Lin, Colour visual cryptography schemes, *IET Information Security*, July 2008.
- [4] Kang InKoo et. al., Color Extended Visual Cryptography using Error Diffusion, *IEEE* 2010.
- [5] SaiChandana B., Anuradha S., A New Visual Cryptography Scheme for Color Images, *International Journal of Engineering Science and Technology*, Vol 2 (6), 2010.
- [6] Li Bai , A Reliable (k,n) Image Secret Sharing Scheme by, *IEEE*,2006.
- [7] M.Amarnath Reddy, P.Shanthi Bala, G.Aghila "visual cryptography schemes comparision", Vol. 3 No. 5 May 2011.
- [8] SaiChandana B., Anuradha S., A New Visual Cryptography Scheme for Color Images, *International Journal of Engineering Science and Technology*, Vol 2 (6), 2010.
- [9] Kang InKoo et. al., Color Extended Visual Cryptography using Error Diffusion, *IEEE* 2010.
- [10] JIM CAI, " A SHORT SURVEY ON VISUAL CRYPTOGRAPHY SCHEMES".

A BOOLEAN MODELING FOR IMPROVING THE ALGORITHM APRIORI

Abdelhak Mansoul¹ and Baghdad Atmani²

¹Computer Science Laboratory of Oran (LIO),
Department of Computer Sciences, University of Skikda, Algeria.
mansoul.abdelhak@yahoo.fr

²Computer Science Laboratory of Oran (LIO),
Department of Computer Science, University of Oran ES-Sénia, Algeria.
atmani.baghdad@gmail.com

ABSTRACT

Mining association rules is one of the most important data mining tasks. Its purpose is to generate intelligible relations between attributes in a database. However, its use in practice is difficult and still raises several challenges, in particular, the number of learned rules is often very large. Several techniques for reducing the number of rules have been proposed as measures of quality, syntactic filtering constraints, etc. However, these techniques do not limit the shortcomings of these methods. In this paper, we propose a new approach to mine association, assisted by a Boolean modeling of results in order to mitigate the shortcomings mentioned above and propose a cellular automaton based on a boolean process for mining, optimizing, managing and representing of the learned rules.

KEYWORDS

Cellular automaton, Data mining, Association Rules, Boolean modeling, Apriori-Cell

1. INTRODUCTION

Numerous studies on the association rules are made [2], [9], [11]. However, their uses in practice are difficult and still raises many challenges, especially the exorbitant number of rules learned, and the processing time. Recent studies have also proposed a series of solutions to improve the performance of the mining process [4], [5], [7], [15], [16], without eliminating the shortcomings of this method of search data. It became necessary to find adequate techniques and algorithmic solutions to minimize the cost for space and computing time. The Apriori algorithm introduced an approach called "test-and-generate" with pruning. However, this approach suffers from a number of candidates that generates, particularly for relatively small values of support. However, these approaches do not limit the shortcomings of these methods. Given this situation, it became necessary to invest in new methods to faces the following challenges:

- Find heuristics to prune the search space;
- Find technical or algorithmic solutions, specifically adequate data structures, to minimize the cost in space or in process time.

We will expose in our present article, the second part of our study (see 3.2 Step 4) and its experimentation that was performed with the basic Apriori in order to demonstrate relevance and efficiency of the approach that we have considered. Later (continuation of our study) we will present the first part (see 3.2, Steps 2 and 3), we adopt Apriori-Cell.

2. RELATED WORK

Recent studies have proposed a series of solutions to improve the performance of extraction process of frequent item sets, including cellular automata [10]. Solutions were oriented essentially on the Reduction of I / O and the minimization of the cost of the step of computing the support [9]. Other studies have been based on the discovery of "closed" item sets arising from the theory of formal concepts [11]. Others propose to generate a representative base or generic association rules [14] [16] and used techniques to reduce the number of rules with the use quality measurements [15], syntactic filtering by constraints [2].

3. THE PROPOSED APPROACH

We propose a new approach being located at the junction of two domains that are the Knowledge Discovery from Databases (KDD) on one hand and representation of knowledge from the other. Our approach proceeds in three steps:

1. Extraction of frequent patterns and generating association rules using the algorithm Apriori-Cell which operates on a cell basis;
2. Boolean modelling for association rules;
3. Rules management by the inference engine ICR of the cellular automaton.

A cellular automaton is a grid composed by cells which change states in discreet steps. After each step, the status of each cell is modified as states of its neighbors in the previous step [1].

Our approach is implemented by two modules:

- The module MAR (Mining Association Rules) ;
- The module ICR (Induction of Cellular Rules).

The dynamics of the cellular automaton. The inference engine of the cellular automaton simulates the basic operating principle of a classical inference engine using two finite layers of finite automata. The first layer CEL Fact/CEL Item for the basic facts/Items and the second layer, CEL Rule/CEL Transaction for the basic Rules/Transactions. Each cell at time $t+1$ depends only on the status of its neighbors and his own at time t .

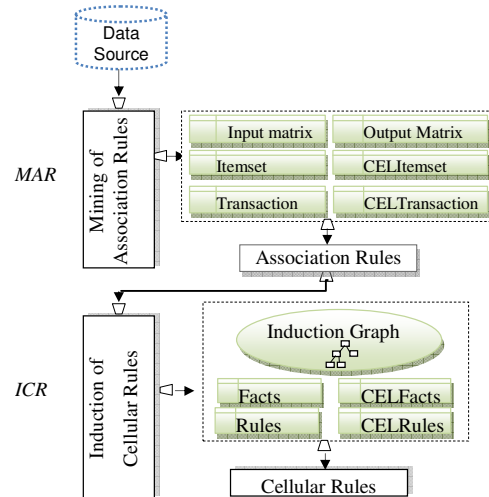


Figure 1. System Architecture

In each layer, the content of a cell determines if and how it participates in each inference step. At each step, a cell may be active (1) or passive (0) i.e whether or not participating in the inference. The principle adopted is simple:

- Any cell i of the first CEL Fact/CEL Item layer is considered an established fact if its value is 1, otherwise it is considered as fact to establish. It is presented in three states: input state (EF/EI), internal state (IF/II) and output state (SF/SI);
- Any cell j of the second layer CEL Rule/CEL Transaction layer is considered a Rule/Transaction candidate if its value is 1, otherwise it is considered as a Rule/Transaction which shall not participate in the inference. It is presented in three states: input state (ER/ET), internal state (IR /IT) and output state (SR/ST).
- Incidence matrix RE and RS represents the input / output relation of Facts/Items and are used in forward chaining and backward chaining by reversing their order.

Thus, the cellular automaton will help optimize the representation of extracted knowledge (association rules) by the boolean principle and their management by using its inference engine through the basic functions δ_{fact} and δ_{rule} which provide the dynamics of cellular automaton (See 4.1).

3.1.The Proposed Algorithm

The process adopted by our system is a succession of four major steps:

- Step 1:* Selection and data preprocessing;
- Step 2:* Cellular representation of preprocessed data;
- Step 3:* Data mining by the cellular automaton using the algorithm *Apriori-Cell*;
- Step 4:* Post-processing of results.

Algorithm : Apriori-Cell

Input : Transactional-data-base (D), minimum-support S , confidence C Output : lists-of-frequent-items (F_n), Association-rules-base (Br) $F_n \leftarrow \{ \}$

Begin

Input-matrix = data preprocessing for (D)

While we can make joints Do

For each line of input-matrix Do

calculate support for (item)

 If support(item) $\geq S$ Then $F_n \leftarrow F_n + \{ \text{item} \}$

EndIf

EndFor

EndWhile

Br = Generate-rules (F_n)End

3.2.Principle of the algorithm Apriori-Cell

This module applies the cellular principle on the basic Apriori algorithm for mining frequent itemsets. It simulates the basic operating of a join engine inspired by Apriori adapted to cellular automaton, on two finite layers of a finite automaton. the first layer *CELItems* for the items base, and the second layer *CELTransactions* for the transactions base.

The state of each cell at time $t + 1$ depends only on the state of its neighbors and his own at time t . In each layer, the contents of a cell determines whether and how it participates in each inference step: a cell can be active (1) or passive (0), i.e whether or not participating in the inference.

The principle is simple, we suppose that there are l cells in the layer *CELItems*, and r cells in the *CELTransactions* layer. The states of the cells are: *EI*, *II*, and *SI*, respectively *ET*, *IT* and *ST* are the input, the internal state and the output of a cell of *CELItems*, and respectively a cell of *CELTransactions*. The internal state *II* of a cell of *CELItems* indicates the status of the item: in the case of an item, $II = 1$ corresponds to a state type *support_item* \geq *minsupport_fixed*. For a cell from *CELTransactions*, the internal state *IT* will always be equal to 1 (the transactions are always established).

The join applied by Apriori-Cell. In the first step the join is made between items using the logical AND, line by line, i.e, it fixes line 1 for example, and it'll do its join to the rest of lines. Once completed, it will go to the second line without considering line 1, this time. And this process continues until the join between items become impossible. At the first iteration, the join is made unconditionally, but beyond 2 items, it applies the following rule: for the join of k -items we must have $(k-1)$ -items in antecedent of the rule to be common.

Generation rules. The module **Generate-rules** is used for the generation and validation of association rules from the lists of frequent n -itemset extracted by Apriori-Cell. This module allows to minimize the number of reading of the database by using to calculate the confidence of each rule, the data cubes, which help in the positioning of the lists of n -itemsets extracted on three dimensions. A dimension for transactions, another one for the antecedent and the last for the consequent of each rule.

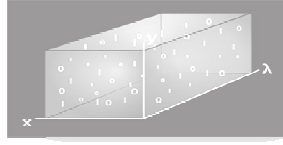


Figure 2. Representation of the data cube, with x : axis of transactions (1,2), y : axis of the list 1-frequent items (F1) (aac, acpP, aacA1) and λ : axis of the list 1-frequent items (F1)

Step 4 : post-processing of results

(a) Production of induction graph. An algorithm uses as input the association rules $\{R_i\}$, items of $Antecedent_i$ and $Consequent_i$, and it will give on output an induction graph, with a summit S_p and a node p on which we make a test with possible results binary or with multiple values.

(b) Generation of boolean rules from the induction graph. Induction graph is read to generate boolean rules in the following form:

$$Rb_k : \{ P_k \} \text{ Then } \{ C_k, S_p \}$$

(c) Representation of boolean rules. The generated rules (see step 4.b) are represented by cell layers where: $\{ Rb_k \}$ gives the set of rules $\{\text{Rules}\}$ and $\{ P_k, C_k, S_p \}$ gives the set of facts $\{\text{Facts}\}$

(d) Integration. The cellular automaton integrates the generated rules in the knowledge base for use through different inference strategies.

4. ILLUSTRATIVE EXAMPLE OF THE REPRESENTATION OF RULES BY CELLULAR AUTOMATA

We suppose have obtained the following two rules of association with genes aceA-2, pstS-3, argC and phhB, using the Apriori-Cell algorithm:

$$R_1 : \{ aceA-2=1 \}, \{ pstS-3=1 \}, 45, 77$$

$$R_2 : \{ aceA-2=1, phhB=1 \}, \{ argC=1 \}, 45, 70$$

and that these two rules have generated the following boolean rules from the induction graph :

$$Rb_1 : Si \{ S_0 \} \text{ Alors } \{ pstS-3=1, S_1 \}$$

$$Rb_2 : Si \{ S_1 \} \text{ Alors } \{ argC=1, S_2 \}$$

These rules will be represented (step 4, b) in layers CELFacts, CELRules, input matrix (R_E) and output matrix (R_S).

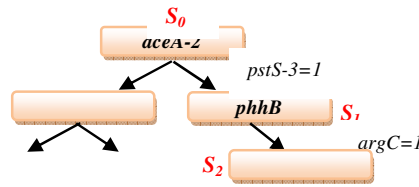
Step 3 : Data mining by the cellular automaton

From the sample test data base (see 5. Table1), we proceed to data mining by Apriori-Cell. We suppose have obtained two association rules with the following genes: aceA-2, pstS-3, argC and phhB.

Rule	antecedent	consequent	Support %	Confidence %
R_1	$aceA-2=1$	$pstS-3=1$	45	77
R_2	$aceA-2=1,$ $phhB=1$	$argC=1$	45	70

Step 4 : post-processing of results

a) Production of the graph induction



b) Generation of boolean rules from the graph induction

$$Rb_1: \text{If } \{ S_0 \} \text{ then } \{ pstS-3=1, S_1 \} \quad Rb_2: \text{If } \{ S_1 \} \text{ then } \{ argC=1, S_2 \}$$

c) Representation of boolean rules

The boolean rules Rb_1 and Rb_2 produced are represented by the layers *CELFacts* (Facts + CELFacts) and *CELRules* (Rules + CELRules) and input matrix (R_E) and output matrix (R_S).

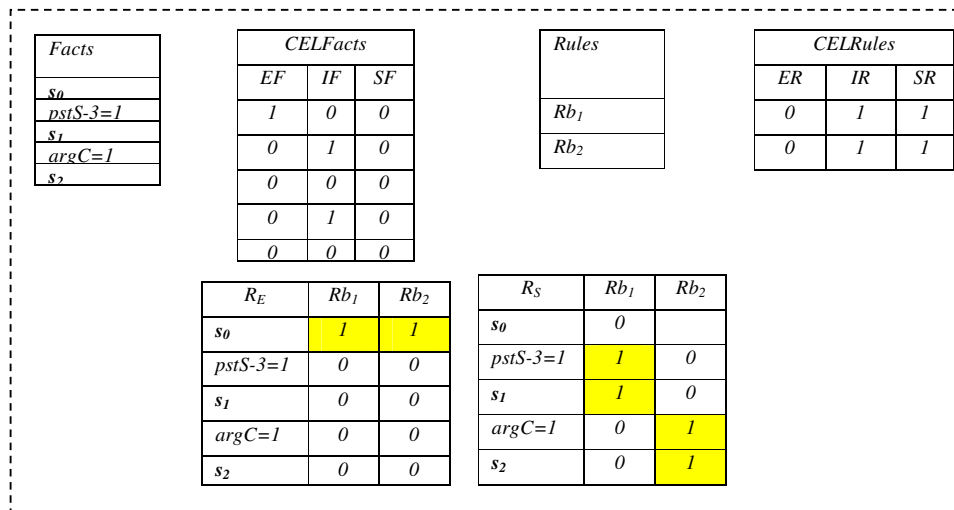


Figure 3. Cell layers of the cellular automaton with input state (EF), internal state (IF) and output state (SF), input state (ER), internal state (IR) and output state (SR)

4.1. Simulation of The Cellular Inference Engine

The cellular automaton simulates the operating of an inference engine by using the transition functions previously mentioned, δ_{fact} and δ_{rule} .

We show that simulation starting from *CELFacts* and *CELRules* of the example shown previously (Figure 3) by considering that G_0 is the initial configuration of the cellular automaton, and $\Delta = \delta_{rule} \circ \delta_{fact}$ the global transition function: $\Delta(G_0) = G_1$.

The cellular automaton change state from G_0 to the state G_1 with

$$\Delta(G_0) = G_1 \text{ if } G_0 \xrightarrow{\delta_{fact}} G'_0 \text{ and } G'_0 \xrightarrow{\delta_{rule}} G_1$$

After application of the law of global transition Δ we obtain the configurations G_1 , G_2 and finally G_3 .

1. G_0 is the initial configuration of the cellular automaton

G_0

Facts
s_0
$dstS-3=1$
s_1
$argC=1$
s_2

CELFacts		
EF	IF	SF
1	0	0
0	1	0
0	0	0
0	1	0
0	0	0

Rules
Rb_1
Rb_2

CELRules		
ER	IR	SR
0	1	1
0	1	1

2. Evaluation, selection and filtering (application of δ_{fact})

Facts
s_0
$dstS-3=1$
s_1
$argC=1$
s_2

CELFacts		
EF	IF	SF
1	0	1
0	1	0
0	0	0
0	1	0
0	0	0

Rules
Rb_1
Rb_2

CELRules		
ER	IR	SR
1	1	0
0	1	1

3. Execution (application of δ_{rule}) $\Delta(G_0) = G_1$

G_1

Facts
s_0
$dstS-3=1$
s_1
$argC=1$
s_2

CELFacts		
EF	IF	SF
1	0	1
1	1	0
1	0	0
0	1	0
0	0	0

Rules
Rb_1
Rb_2

CELRules		
ER	IR	SR
1	1	0
0	1	1

4. Application of the global transition function : $\Delta = \delta_{rule} \circ \delta_{fact}$ $\Delta(G_1) = G_2$

G_2

Facts
s_0
$dstS-3=1$
s_1
$argC=1$
s_2

CELFacts		
EF	IF	SF
1	0	1
1	1	1
1	0	1
0	1	0
0	0	0

Rules
Rb_1
Rb_2

CELRules		
ER	IR	SR
1	1	0
0	1	1

5. Evaluation, selection and filtering (application of $\delta fact$)

<i>Facts</i>	<i>CELFacts</i>	<i>Rules</i>	<i>CELRules</i>
<i>s_n</i>	<i>EF</i> <i>IF</i> <i>SF</i>		<i>ER</i> <i>IR</i> <i>SR</i>
<i>dstS-3=1</i>	1 0 1	<i>Rb₁</i>	1 1 0
<i>s₁</i>	1 1 1	<i>Rb₂</i>	1 1 0
<i>argC=1</i>	1 0 1		
<i>s₂</i>	0 1 0		
	0 0 0		

6. Application of the global transition function

G₃

<i>Facts</i>	<i>CELFacts</i>	<i>Rules</i>	<i>CELRules</i>
<i>s_n</i>	<i>EF</i> <i>IF</i> <i>SF</i>		<i>ER</i> <i>IR</i> <i>SR</i>
<i>dstS-3=1</i>	1 0 1	<i>Rb₁</i>	1 1 0
<i>s₁</i>	1 1 1	<i>Rb₂</i>	1 1 0
<i>argC=1</i>	1 0 1		
<i>s₂</i>	1 1 1		
	1 0 1		

Final configuration **G₃** obtained after four iterations.

5. EXPERIMENTATION

To examine the effectiveness in practice of our system, we have implemented the engine, and we conducted experimental tests on a machine (Intel Celeron 540 CPU frequency 186 GHz, 512 MB RAM) with a sample test data base (Table 1) representing the genomic sequences mycobacterium tuberculosis) with the first 12 genes of each strain, and the assumption that these genes are sufficiently distinctive and representative of each strain taken separately.

Table 1. Test Data Base ¹ (12 genes of each strain)

Strain	Genes
<i>Mt CDC155</i>	<i>aac accD aceA-1 aceA-2 aceB aceE ackA acnA acp-1 acp-2 acpP acpS</i>
<i>Mt F11</i>	<i>aceE acpP acpS adk alaS alr argC argD argJ argS aroB aroE</i>
<i>Mt H37Ra</i>	<i>aac aao accA1 accA2 accA3 accD1 ccD2 accD3 accD4 accD5 accD6 aceAa</i>
<i>Mt H37Rv</i>	<i>35kd_a aac aao accA1 accA2 accA3 accD1 accD2 accD3 accD4 accD5 accD6</i>

5.2. Discuss of The Results

Processing time. We observe that the Apriori algorithm takes an important part in execution time of the system in whole, ie in its most important phases as the generation of association rules by Apriori and the generation of boolean rules.

Table 2. Evolution of execution time (Basic Apriori and global)

Confidence %	Support %	Number of Genes	Generated items	Number of rules	Execution of Apriori	Global Execution
10	30	12	37	69	0.00 s	0.00 s
50	50	12	37	125874	0.67 s	1.69 s
70	60	12	37	786756	3.56 s	6.17 s

Storage space. We find that cell representation is more interesting, and it will be much most prominently with a more substantial sample.

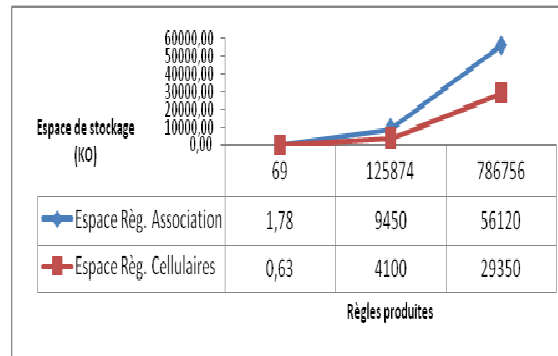


Figure 4 . Evolution of storage space

6. CONCLUSION

After describing the disadvantages of rule-based methods in data mining, we have proposed an extraction rules guided by a Boolean modeling based on the Boolean principle of cellular automata in order to have a base rules optimized and reduced processing time enough, and thus make a contribution to the construction of knowledge-based systems by adopting a new cellular technic Thus, the advantages of our method based on the cellular automaton can be summarized as follows:

- Simple and minimal preprocessing of association rules base, for its transformation into binary matrix according to the principle of cell layers,
- Ease of implementation functions δ_{fact} and δ_{rule} that are low complexity and well adapted to situations with many attributes of rules.

REFERENCES

- [1] Atmani, B., Beldjilali, B. (2007) "Knowledge Discovery in Database : induction graph and cellular automaton", Computing and Informatics Journal, Vol. 26 N°2 171-197
- [2] Besson, J., Robardet, c., Boulicaut, J.F. (2004) "Constraint-based mining of formal concepts in transactional data", Conference on Knowledge Discovery and Data Mining (PAKDD'04) volume 3056 of LNCS, Sydney- Australia, pp. 615–624
- [3] Pasquier, N., Bastide, Y., Taouil, R., Lakhil, L., Stumme, G. (2000) "Mining minimal non-redundant association rules using frequent closed itemsets", Proceedings of the Intl. Conference DOOD'2000, LNCS, Springer-verlag, July, pp. 972-986
- [4] Hajek, P., Havel, I., Chytil, M. (1966) "The GUHA method of automatic hypotheses determination", Computing 1, pp. 293-308

- [5] Boullé, M. (2007) “Recherche d’une représentation des données efficace pour la fouille des grandes bases de données”, Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications Paris.
- [6] Abdelouhab, F., Atmani, B. (2008) “Intégration automatique des données semi-structurées dans un entrepôt cellulaire”, Troisième atelier sur les systèmes décisionnels, Mohammadia– Maroc, 10-11 octobre, pp. 109-120
- [7] Agrawal, R., Imielinski, T., Swami, A. (1993) “Mining associations between sets of items in large databases”, Proc. of the ACM SIGMOD Conf., Washington DC, USA
- [8] Agrawal, R., Srikant, R. (1994) “Fast Algorithms for Mining Association Rules”, Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, pp 487-499
- [9] Bykowski, A., Rigotti, C. (2001) “A condensed representation to find frequent patterns”, Proceedings of the Twentieth ACM SIGACTSIGMOD SIGART Symposium on Principles of Database Systems, ACM Press, pp. 267-273
- [10] Fawcett, T. (2008) “ Data mining with cellular automata”, ACM SIGKDD Explorations Newsletter, v.10 n.1, June 2008.
- [11] Ganter, B., Wille, R. (2004) “Conceptual graphs and formal concept analysis”, Lecture Notes in Computer Science Volume 1257, 1997, pp 290-303 Springer-Verlag
- [12] Mansoul, A., Atmani, B. (2009) “Fouille de données biologiques : vers une représentation booléenne des règles d’association”, CEUR-WS:04-Dec-2009/Vol-547, Conférence Internationale sur l’Informatique et ses Applications CHIA’09, Saida – Algérie
- [13] Mansoul, A., Atmani B. (2010) “Vers un automate cellulaire pour la fouille de données : Partie I : la représentation booléenne des règles d’association”, ASD 5/6 Novembre 2010, Sfax, Tunisie.
- [14] Pasquier, N., Bastide Y., N., Lakhal, L. (2000) “Mining minimal non-redundant association rules using frequent closed itemsets”, Computational logic. International conference No1, London, royaume uni (24/07/2000), vol. 1861, pp. 972-986.
- [15] Vaillant, B., Meyer, P., Prudhomme, E., Lallich, S., Lenca, P., Bigaret, S. (2005) “Mesurer l’intérêt des règles d’association”, Atelier Qualité des Données et des Connaissances (DQK 05, Actes de EGC), pp. 69-78
- [16] Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L. (2001) “Intelligent structuring and reducing of association rules with formal concept analysis”, Proc. KI’2001. conference, LNAI 2174, Springer.

AUTHORS

Abdelhak Mansoul is an Assistant Professor at Skikda University and affiliated researcher in Oran Computer Lab of Oran University. His research interests are in Database Management System, Data Mining, decision support systems, and simulation.

Baghdad Atmani is a professor in computer science at the University of Oran (Algeria). His interest field is Data Mining and Machine Learning Tools. His research is based on Knowledge Representation, Knowledge-based Systems and CBR, Data and Information Integration and Modeling, Data Mining Algorithms, Expert Systems and Decision Support Systems. His research are guided and evaluated through various applications in the field of control systems, scheduling, production, maintenance, information retrieval, simulation, data integration and spatial data mining.

COMBINING DECISION TREES AND K-NN FOR CASE-BASED PLANNING

Sofia Benbelkacem, Baghdad Atmani and Mohamed Benamina

Computer Science Laboratory of Oran (LIO), University of Oran, Algeria
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algeria
{sofia.benbelkacem, atmani.baghdad, benamina.mohamed}@gmail.com

ABSTRACT

In everyday life, we are often faced with similar problems which we resolve with our experience. Case-based reasoning is a paradigm of problem solving based on past experience. Thus, case-based reasoning is considered as a valuable technique for the implementation of various tasks involving solving planning problem. Planning is considered as a decision support process designed to provide resources and required services to achieve specific objectives, allowing the selection of a better solution among several alternatives. However, we propose to exploit decision trees and k-NN combination to choose the most appropriate solutions. In a previous work [1], we have proposed a new planning approach guided by case-based reasoning and decision tree, called DTR, for case retrieval. In this paper, we use a classifier combination for similarity calculation in order to select the best solution to the target case. Thus, the use of the decision trees and k-NN combination allows improving the relevance of results and finding the most relevant cases.

KEYWORDS

Case-Based Reasoning, Classifier Combination, Data Mining, Case Retrieval, Decision Tree, Planning

1. INTRODUCTION

Planning is currently of great interest because it combines two major areas of Artificial Intelligence, exploration and logic. The intersection of these two areas has led to improved performance over the last twenty years [2]. The planning emergence in Artificial Intelligence led to the so-called classical planning [3]. But the classical planning has multiple drawbacks like the unrealistic assumptions that recognize the full knowledge of the environment, it is insensitive to changes in the environment, it does not deal with the possibility of failure or uncertainty in the environment or the presence of other agents or unpredictable situations, etc. To address these problems, a planner must be able to reason in the real world with the notion of time and resources, support more expressive representation of knowledge, evolve using past experience, cooperate with other planners, etc [4]. The rejection of the classical planning paradigm has resulted in new planning techniques aimed at solving the problems which can't be solved by traditional planning systems. Among these techniques, we are interested in case-based planning. Case-based planning is based on the reuse of past successful plans for the development of new plans. A plan for a set of objectives is not built piece by piece but by changing a memory map that partially or fully satisfies the objectives. So, the case-based planning provides significant time savings by avoiding trying to solve problems already treated. Then, to take advantage of past experience and optimize computing time, instead of synthesizing plans from primitive operators, David C. Wyld et al. (Eds) : SAI, CDKP, ICAITA, NeCoM, SEAS, CMCA, ASUC, Signal - 2014 pp. 115–122, 2014. © CS & IT-CSCP 2014

DOI : 10.5121/csit.2014.41112

we adopt the case-based planning principle by combining the case-based reasoning and planning to implement our planning system guided by case-based reasoning.

In a previous work [1], we proposed a new case-based planning technique based on case-based reasoning (CBR) and decision tree, that we called DTR (Decision Tree for Retrieval). It is a decision support system that fits in the medical context. We based our approach on case-based reasoning and decision tree for many reasons. On the one hand, in care planning it is common to encounter patients who need follow the same treatment plan than others. On the other hand, the knowledge of doctors is not based only on rules but also on their theoretical knowledge and experience. For this, we use the case-based reasoning which is a paradigm of problem solving based on past experience [5]. The case-based reasoning will allow us to optimize time, given that in the medical field time is an important factor which must not be neglected. Another factor to consider in the medical field is that the data generated in health organizations are increasing. To manage a large amount of data we used data mining. We introduced an induction decision tree in the retrieval phase of case-based reasoning process. This step requires the use of a similarity measure between cases. We therefore used a retrieval phase guided by decision tree as a measure of similarity [1]. However, we cannot always rely on the solutions proposed by the retrieval by decision tree that could provide impertinent solutions if there is not enough examples (cases) in the case base. To overcome this drawback, we propose to use different classifiers and combine their predictions with the majority vote in order to achieve a more relevant result. The objective of this work is to improve the results quality of the proposed approach by using a classifier combination. Indeed, the main idea behind the combination of classifiers is an increase in the quality of results [6]. To perform the experiment, we evaluated the approach on real cases in the medical field, specifically for tuberculosis treatment.

The paper is organized as follows. In Section 2, we mention some works about the similarity measures used for retrieval. Then, in Section 3 we give a description of the proposed approach. Section 4 presents some experimental results, which include a combination of classifiers. Finally, Section 5 is devoted to conclusions and perspectives of this work.

2. LITERATURE REVIEW

The retrieval step of case-based reasoning process requires the use of a similarity measure. This notion of similarity between cases has been the subject of several works implementing various similarity measures. Nunez et al. [7] propose a new similarity measure for case retrieval. It takes into account the different nature of the quantitative or qualitative values of the continuous attributes depending on its relevance. Thus, different criterions of distance are applied for continuous attributes. Guo & Neagu [8] propose a similarity-based classifier combination system. The classifiers studied include voting-based k-nearest neighbours, weighted k-nearest neighbours, k-nearest neighbours model-based classifier and contextual probability-based classifier. Juarez et al. [9] propose a temporal similarity measure for heterogeneous event sequences, based on the overall uncertainty of a temporal constraint network. The temporal similarity is measured by describing a unique temporal scenario of temporal relations and calculating the uncertainty produced. Petridis et al. [10] present a system built on a similarity metric using a graphical representation of shapes for retrieval. The special feature of this system is that similarity is derived primarily from graph matching algorithms. Zhong et al. [11] propose a two-layer case retrieving method applied to emergency field. This method is based on structural and attribute similarity degrees. First, the structural similarity degrees between the historical cases and the current problem are analyzed. Second, the attribute similarity degrees between them are analyzed. At last, the synthetic similarity degrees between them are calculated. This method can avoid failing to calculate the similarities among the cases with the missing values and the similarity degrees between the historical cases. Hashemi et al. [12] propose a new measuring similarity

method between work pieces using numeric and some symbolic attributes. This method is a similarity measurement system used for fixture design. It is composed of template retrieval and nearest neighbor. Kumar Jha et al. [13] propose a case-based decision support system for patients with diabetes. This system uses similarity by ontology to retrieve similar cases from the case base and generates a basic care plan.

A novel planning method, called DTR, is proposed here to improve the retrieval step of case-based reasoning. Figure 1 illustrates the general architecture of the proposed approach. It consists of several steps going from the project description to the data classification.

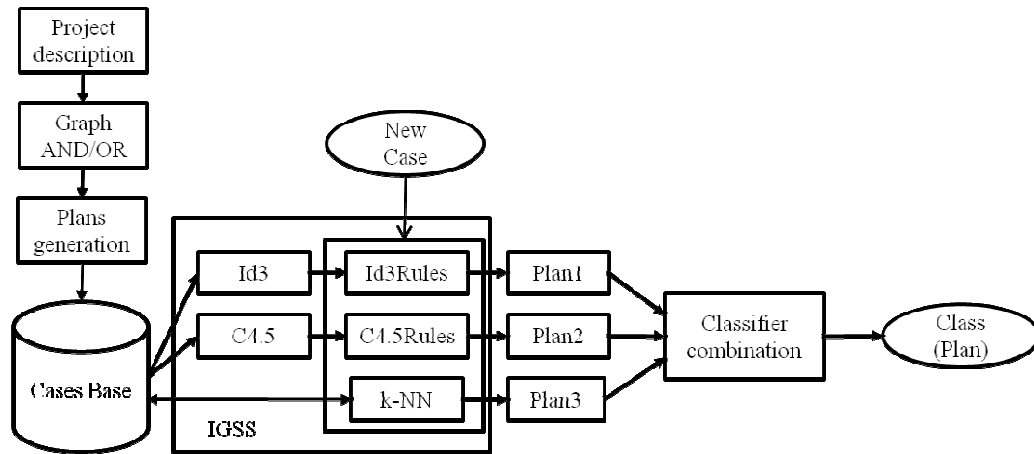


Figure 1. General architecture of the proposed approach

We call project, the set of actions to be undertaken to meet an identified need within a specified time. The organization and the sequence of tasks are generally in the form of tables or graphs. We describe the project representing the sequence of tasks in a table to generate the graph AND/OR [1]. Table 1 shows the project description of tuberculosis treatment.

Table 1. Project description of tuberculosis treatment.

Rubric	Task	Description	Anteriority
	Begin	Start of treatment	-
Treatments	A	Rx thoracic + bacteriologic exam	Begin
	B	2RHZE/4RH (2tablets/day)	A
	C	2RHZE/4RH (3tablets/day)	A
	D	2RHZE/4RH (4tablets/day)	A
	E	2SRHZE/1RHZE/5RHE	A
	F	3OKZE/18OEZ	A
	G	3EthOKZC/18EthOZ	A
Controls	H	Controls (I)	B, C, D
	I	Controls (II)	E
	J	Controls (III)	F, G
	End	End of treatment	H, I, J

A graph AND/OR is a graph whose nodes represent tasks and the edges represent relationships between tasks. A task represents the action performed for a period of time and relationships between tasks are constraints to satisfy [3]. The graph AND/OR given in the Figure 2 is generated from the project of tuberculosis treatment described in the previous step.

We use the data mining tool IGSS (Induction Graph Symbolic System) to build the classification model. IGSS has been developed in our research team SIF (Simulation, Intégration et Fouille de données) to enrich the graphical environment of Weka platform. It uses Boolean modeling to optimize the induction graph and automatic generation of rules [15]. Provided that the training set ΩA is representative of the original population, we can deduce classification rules which are of the form: If Condition Then Conclusion. Condition is a logical expression consisting of disjunction of a conjunction that will be called premise and Conclusion is the majority class in the node described by the condition.

The decision tree can be used in different ways [20]: classification of new data, estimation of an attribute, extraction of classification rules for the target attribute, etc. In our case, it is to classify new data. The new data is a new case which we do not know its solution part. To find the solution part we apply the retrieval step of case-based reasoning. The retrieval involves looking for similar cases to the new data. We treat this step using decision tree. The new data will be incorporated into the classification model which consists of a decision tree and classification rules. The classification model is responsible to classify new data by assigning a plan (class).

4. EXPERIMENT AND DISCUSSION OF RESULTS

During the construction of the classification model, we used only a single classifier to build the decision tree. But with one classifier and few examples in the case base, we can sometimes leads to an impertinent result. For this, we found useful to adopt a combination of classifiers in order to improve the quality of results and we test it on different datasets. First, we evaluate the proposed approach on six public datasets extracted from the UCI machine learning repository [18]. General information about these datasets is listed in Table 3. Then, we perform some experiments on a dataset of real cases which represent patients treated for tuberculosis. An extract of the case base is given in Table 2. In Table 3, the meaning of the title in each column is as follows, NA: Number of attributes, NN: Number of Nominal attributes, NO: Number of Ordinal attributes, NB: Number of Binary attributes, NI: Number of Instances and CD: Class Distribution.

Table 3. General information about UCI datasets.

Dataset	NA	NN	NO	NB	NI	CD
Glass	9	0	9	0	214	70:17:76:0:13:9:29
Hepatitis	19	6	1	12	155	32:123
Ionosphere	34	0	34	0	351	126:225
Iris	4	0	4	0	150	50:50:50
Wine	13	0	13	0	178	59:71:48
Zoo	16	16	0	0	90	37:18:3:12:4:7:9

We used the Percentage split method with a rate of 80% to evaluate the prediction accuracy of three classifiers Id3 [16], C4.5 [17], k-NN [19] and their combination. This method takes 80% of data inside the case base for the training set and 20% of the test set. K-NN is k nearest neighbors, we took k = 5 and the classifiers Id3 and C4.5 are designed for construction of the decision tree. Additionally, we adopted a combination of classifiers by majority vote. It is to count the number of votes for each class offered by different classifiers and choose the class with the highest number of votes (the most proposed class by the classifiers). We consider the C4.5 classifier as the most priority. If all classes have the same number of votes (each classifier gives a different result) then we take the proposed solution by the Id3 classifier. To assess the performance of the proposed approach, we compare our experimental results with four similarity-based classifier combination methods Maximal Similarity-based Combination (MSC), Average Similarity-based Combination (ASC), Weighted Similarity-based Combination (WSC) and MV proposed by Guo

& Neagu [8]. The experimental results are given in Table 4. Table 4 shows that the classification accuracy of the three classifiers Id3, C4.5 and K-NN is slightly better than other methods over five datasets while the other methods have a better performance with the Wine dataset. However, we note that the performance seems better with the combination of the classifiers Id3, C4.5 and k-NN than the ensemble classifiers and other methods in most of the cases.

Table 4. Comparison between classifier combination and other methods with UCI datasets.

Dataset	Id3	C4.5	k-NN (k=5)	Classifier combination	Other methods			
					MV	MSC	ASC	WSC
Glass	100	100	93.92	100	69.52	70.95	70.95	70.95
Hepatitis	87.5	76.12	71.61	87.5	85.33	87.33	86.67	87.33
Ionosphere	94.32	95.44	90.88	95.5	88.57	89.43	88.86	89.43
Iris	96.66	96	94.66	96.7	96.00	96.67	96.67	96.67
Wine	89.88	80.33	76.40	89.9	95.29	96.47	96.47	96.47
Zoo	98.01	99.01	95.04	99.01	95.56	95.56	96.67	96.67

Next, we calculated the metrics Precision, Recall, F-measure and Accuracy for each classifier (Id3, C4.5, k-NN) and for the combination of these classifiers applied to the real case base about tuberculosis. The performance evaluation is given in the Table 5. The experimental results presented in Table 5 show that C4.5 has a better performance than the other classifiers and the classifier combination obtains the highest Precision, Recall, F-measure and Accuracy among other classifiers on tuberculosis dataset. Thus, we can see that the classifier combination with the majority vote can improve the relevance of results.

Table 5. Performance evaluation with the case base of tuberculosis.

Evaluation metrics	Id3	C4.5	k-NN	Classifier combination
Precision	62.5	83.3	71.4	83.3
Recall	83.3	83.3	83.3	83.3
F-measure	71.4	83.3	76.9	83.3
Accuracy	80.9	90.4	85.7	90.4

5. CONCLUSIONS

We have proposed a new approach of case-based planning based on case-based reasoning and decision trees. The objective of our approach is to provide support to practitioners in selecting the appropriate treatment. We implemented our approach on real cases involving patients treated for tuberculosis. We used the IGSS tool for building decision tree and generate classification rules from the training set. To improve the quality of results, we used combination of classifiers. Thus, the use of multiple methods simultaneously can possibly afford to combine the advantages without accumulating disadvantages. For this, we combined decision trees and k-NN in order to get the mostly proposed solution (most relevant). As a method for combining classifiers, we used the majority vote because it is a fairly simple feature fusion and the most used. Faced with new data, the system classifies this data by associating a class that corresponds to a plan. To assess the performance of the proposed approach, we calculated evaluation metrics. The results of experimentation show that the performance becomes higher with the combination of classifiers. Thus, the proposed solution for the new case is more relevant. As future work, we propose to combine with other classifiers which could probably give better results. Moreover, we intend to apply this approach in another important area for further search, it is the paediatric emergency planning.

REFERENCES

- [1] Benbelkacem, S., Atmani, B. & Mansoul, A. (2012) "Planification guidée par raisonnement à base de cas et datamining: Remémoration des cas par arbre de décision", Atelier aIde à la Décision à tous les Etages Aide@EGC2012, pp62-72.
- [2] Bibai, J. (2010) "Segmentation et évolution pour la planification: le système Divide-And-Evolve", Doctoral Dissertation, University of Paris-sud XI Orsay.
- [3] Baki, B. & Bouzid, M. (2006) "Planification et ordonnancement probabilistes sous contraintes temporelles", Actes du 15e congrès francophone de Reconnaissance des Formes et Intelligence Artificielle, pp99-107.
- [4] Vlahavas, I. & Vrakas, D. (Eds.) (2005) *Intelligent Techniques for Planning*, Idea Group.
- [5] Kolodner, J. (1993) *Case based Reasoning*, Morgan Kaufmann.
- [6] Chitroub, S. (2004) "Combinaison de classifieurs: une approche pour l'amélioration de la classification d'images multisources/multidates de télédétection", *Télédétection*, Vol. 4, No. 3, pp289-301.
- [7] Nunez, H., Sanchez-Marre, M., Cortes, U., Comas, J., Martinez, M., Rodriguez-Roda, I. & Poch, M. (2004) "A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations", *Environmental Modelling & Software*, Vol. 19, No. 9, pp809-819.
- [8] Guo, G. & Neagu, D. (2005) "Similarity-based classifier combination for decision making", *Proc. of IEEE International Conference on Systems, Man and Cybernetics*, pp176-181.
- [9] Juarez, J., Guil, F., Palma, J. & Marin, R. (2009) "Temporal similarity by measuring possibilistic uncertainty in CBR", *Fuzzy Sets and Systems*, Vol. 160, No. 2, pp214- 230.
- [10] Petridis, M., Saeed, S. & Knight, B. (2010) "An automatic case based reasoning system using similarity measures between 3D shapes to assist in the design of metal castings", *Expert Update*, Vol. 10, No. 2, pp43-51.
- [11] Zhong, Q., Zhang, X., Guo, S., Ye, X. & Qiu, J. (2010) "The method of case retrieving in the emergency field based on CBR", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.
- [12] Hashemi, H., Shaharoun, A. & Izman, S. (2013) "To improve machining fixture design: A case based reasoning paradigm", *Journal of Basic and Applied Scientific Research*, Vol. 3, No. 5, pp931-937.
- [13] Jha, M.K., Pakhira, D. & Chakraborty, B. (2013) "Diabetes detection and care applying CBR techniques", *International Journal of Soft Computing and Engineering*, Vol. 2, No. 6, pp132-137.
- [14] Benbelkacem, S., Atmani, B. & Benamina, M. (2013) "Planification basée sur la classification par arbre de décision", *Conférence Maghrébine sur les Avancées des Systèmes Décisionnels*.
- [15] Atmani, B. & Beldjilali, B. (2007) "Knowledge discovery in database: Induction graph and cellular automaton", *Computing and Informatics Journal*, Vol. 26, No. 2, pp1001-1027.
- [16] Quinlan, J. (1986) "Induction of decision trees", *Machine Learning*, Vol. 1, pp81-106.
- [17] Quinlan, J. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo.
- [18] Newman, D.J., Blake, C.L. & Merz, C.J. (1998) *UCI repository of machine learning databases*, University California, Irvine.
- [19] Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K. (2003) "Knn model-based approach in classification", *International Conference on Ontologies Databases and Applications of Semantics*, pp986-996.
- [20] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) *Classification And Regression Trees*, Chapman and Hall, New York.

AUTHORS

Sofia BENBELKACEM is a PhD student at the University of Oran and affiliated researcher in Oran Computer Lab. Her research interests include Data mining, Planning, Case-based reasoning and Medical decision support systems.

Baghdad ATMANI received his PhD in Computer Science from the University of Oran (Algeria) in 2007. He is currently a Professor in Computer Science. His interest field is artificial intelligence and machine learning. His research is based on knowledge representation, knowledge-based systems, CBR, data mining, expert systems, decision support systems and fuzzy logic. His research are guided and evaluated through various applications in the field of control systems, scheduling, production, maintenance, information retrieval, simulation, data integration and spatial data mining.

Mohamed BENAMINA is a PhD student at Oran University and affiliated researcher in Oran Computer Lab. His research interests include Data mining with Ontology, Fuzzy Logic, Fuzzy Expert Systems and Fuzzy Reasoning.

COMPETENCY MODEL FOR INFORMATION SYSTEMS' SPECIALIZATION TRACK UTILIZING RIASEC AND VALUES SEARCH MODELS

Risty Moyo-Acerado¹, Lorena W. Rabago², and Bartolome T. Tanguilig³

College of Information Technology Education,
Technological Institute of the Philippines, Quezon City, Philippines

riszty@yahoo.com¹

lwr823@yahoo.com²

bttanguilig_3@yahoo.com³

ABSTRACT

This paper introduces the competency models for Operations Manager, User Interface Designer, and Application Developers. It will serve as a guide for Information Systems students to identify which among the three of the offered tracks would be most suited for them to pursue according to their knowledge, skills, values and interests. The Holland's RIASEC model and the Values Search model of Bronwyn and Holt were utilized to determine the most dominant interest and most dominant values of the industry computing experts. Survey assessment forms were sent to IT Operations Manager, User Interface Designer, and Application Developer. Most dominant values and interests of industry computing experts were determined as well as the knowledge and skills which are mostly required by the industry in their particular area. Based on the result of the survey, it shows that application developer and user interface designer have a closely related values. Thus a second round of a survey would be needed to come up with the most exclusive dominant values for the particular information systems specialization track.

KEYWORDS

e-learning, career assessment, profile matching, competency model

1. INTRODUCTION

Many factors are rarely considered in determining a student's career options which starts from choosing their college program and even in selecting elective courses. Selecting a field of specialization has never been a big deal for many students. Usually, students select those courses which they think are easy and require less paper works or projects. A fourth year irregular student said, "*Honestly, I just follow my friends when taking elective courses. If that would mean they will be my classmates, then I would definitely take up those elective courses they have enrolled provided that the schedule will not give me long vacant hours...*" Another student said, "*When I am choosing free elective courses my friends and I choose only 1 section to have fun in every vacant hour.*" Moreover, students choose a college program that they think is cool, trending, or easy to pass courses to earn a degree. Sometimes, they choose tracks based from the strong influence of their parents and peers. Most often they do not completely realize the impact of their career choice in their career.

One alumna said, *“Other alumnus who didn't like IT/IS/CS but was able to finish the program was due to the fact that they felt they needed to finish the course just to earn a degree. It's just a matter of formality by all means. As long as they earn a degree and get any job. I know someone who is underemployed. She is not confident enough to apply for any programming position because she is not confident of her programming skills.”* Often, students would only realize the importance of choosing the right track once they get into the real job or even from the start of sending applications to the prospective companies and desired positions where there is tight competition. Thus, many graduates end up getting jobs that are far from their field of specialization contributing to the perennial problem of job mismatching. Some may easily find jobs which seemingly are related to their field, but promotions or career growth takes longer compared to others who are competent in their field. As a result, underemployment rate increases.

According to National statistics Office of the Philippines, the underemployment rate last 2nd and 3rd quarters of 2012 was 19.2. Underemployment means a situation in which a worker is employed but not in the desired capacity, whether in terms of compensation, hours, or level of skills and experience. [11] Moreover, [9] “Underemployment and unemployment varies a great deal depending on the major when there's a skills mismatch. With regards to compensation, [9] underemployed and unemployed graduates earn as much as 10 percent less over their careers compared to their fully employed peers. However, despite of the compensation struggle, some graduates would better be underemployed rather than unemployed.

Therefore, helping students determine the track of specialization according to their values, interests, knowledge, and skills would lead them to a more successful career. With the new curriculum of Information Systems program in Technological Institute of the Philippines (TIP), three tracks are now offered aside from elective courses it has been offering. These tracks are Application Development, Operations Management, and User interface designing.

This study focuses on constructing a competency model for Information Systems Program of TIP which would help students to decide on which track of the Information Systems Program they would consider.

2. CONCEPTUAL FRAMEWORK

Career counselling in Technological Institute of the Philippines was not so active yet compared with other universities until the institution implemented the student advising system. Based on the Information Systems, TIP Quezon City Self Survey Report (IS TIPQC SSR) [3] new advising scheme is intended to provide students with knowledge and guidance on academic policies, plan of study progression, career options, instructional support, and job opportunities, among others, and to monitor the attainment of relevant student outcomes.” It is also intended to support students in maintaining and completing their particular plan of study on time and thus improve the institutional retention and completion rates. One motivation in conceptualizing the proposed competency model is the school's advising scheme. Having clear background of students' competencies could help faculty advisers in advising the students. The implementation of the competency model proposed by [5] focusing on a more specific engineering skills resulted to excellent performance of their students. This paper aimed to construct a model that would help to determine not just knowledge and skills of the students but as well as their values and interests. Figure 1 represents the conceptual framework of this study. Skills and knowledge regarding the computing tools and theories from the aligned courses of Information Systems Program of TIP were identified to design a survey. While the RIASEC interest test of John Holland and value search test of Bronwyn and Holt were utilized to determine the values and interests of industry computing experts in the field related to the Information Systems three tracks. The expected

output of this paper is the competency model of Information Systems' Specialization tracks which was based on the results of survey.

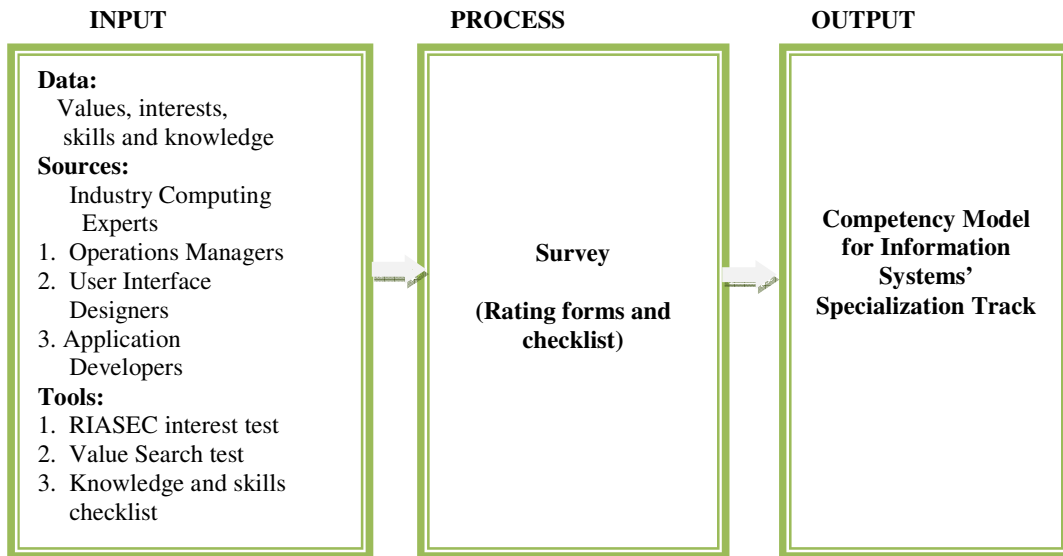


Figure 1. Conceptual Framework of the study

The Information Systems track of the College of Information Technology Education of Technological Institute of the Philippines and the aligned courses of each track are shown on the following tables below.

Table 2 shows Information Systems Track 1 - IT Operations Management, where students concentrate on business operations related to computing processes; Track 2 (table 3): User Interface Designing, where students are trained to create interfaces that organize information for users; and Track 3(table 4): Application Development, where students are trained to simplify the tasks of the end user or resolve recurring problems through process automation.

Table 1. Fundamental And Aligned Courses Of The Information Systems Program Tracks.

TRACK 1 Operation Management	TRACK 2 User Interface Designing	TRACK 3 Application Development
BP(IS100)	OOP (ITE003)	DS(DS201)
MIS(101)	OOP (ITE004)	Deployment & Maintenance (IS400)
Evaluation of Buss. Performance (IS200)		IT Infrastructure (IT001)
PPM(IT203)		
IT QA (IT303)		
SAD(IT202)		
DB(ITE006)		
DB Oracle (ITE007)		

Table 1 shows fundamental courses mapped in each Information Systems track. The knowledge and skills listed on the rating form used in the survey were based from these aligned courses.

Table 2. Information Systems Track 1 - Operations Management

Course	Description
IT401	Data Mining & warehousing
IS403	Business Process Management
IS404	Enterprise Systems
SAP501	Business Analytics Using SAP BW

Table 3. Information Systems Track 2 - User Interface Design

Course	Description
ITE010	Fundamentals of HCI
IS405	Knowledge Management System
IT504	GUI System Development
IT200	Multimedia System Development

Table 4. Information Systems Track 3 - Application Developer

Course	Description
ITE505	Application Development
CS305	Computer Security
IS503	IT audit and control
IS504	IT Security & Risk Mgt.

Figure 2 shows the Framework on the Assessment of Students' Competency for Information Systems Specialization Tracks. The competency model, which is the focus of this paper, is constructed based on result of the evaluations conducted from the industry computing experts. The model could then be used to determine which Information Systems track is the most appropriate for the student to pursue. The same rating tests would be sent to students to determine the student's most dominant and identified values, interest, knowledge, and skills.

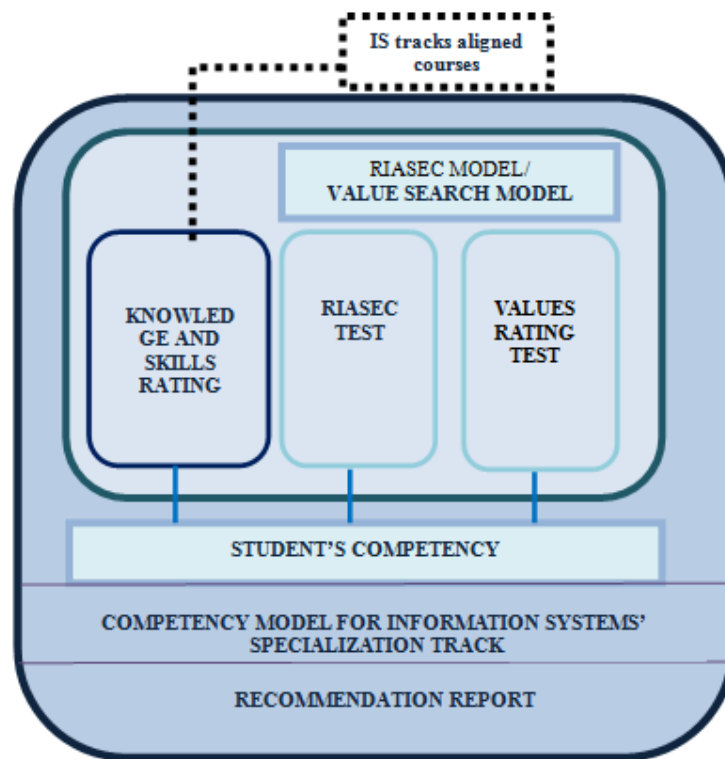


Figure 2. The Framework on the Assessment of Students' Competency for Information Systems Specialization Tracks

3. METHODS, TECHNIQUES, MODE OF INQUIRY

The assessment forms shown on the appendices were used to collect data from the industry computing experts particularly from the IT Operations Manager, User Interface Designer, and Application Developers.

The table shown in appendix 4 is the assessment rating form with regards to the interests expected from the aspiring operations manager, user interface designer, and application developer. This form is based from the RIASEC test designed by the Career Educational Center of University of Hawaii. [4] The RIASEC Holland Code, which stands for: R- ealistic, I- nvestigative, A- rtistic, S- ocial, E- nterprising, C- onventional was used in this study to determine the interest of industry computing experts.

Knowledge and skills listed on the appendix 1, 2, and 3 are based on the aligned courses for each track offered in the Information Systems' curriculum which would be the possible competencies acquired by the students during their training in the institution. Assessment checklist of Knowledge and skills for user interface designer composed mainly of programming languages and tools used in the courses offered in the Information Systems curriculum is shown in appendix 2. Programming languages, database systems, and other application development tools are listed on the knowledge and skills checklist for application developer as shown in appendix 3 while for the expected skills and knowledge from operations manager are shown in appendix 1. With the revision of the Information Systems' curriculum industry experts were already considered as the program formed an advisory board composed of alumni, faculty, institution's officers, and practitioners so the initial list were all based on the updated curriculum.

The Value Search Test found in appendix 6 was used in determining the most dominant values of the industry computing experts. The Value Search Map of Bronwyn and Holt, please refer to figure 3, were utilized to categorize the values. It has eight value categories namely, Understanding, benevolence, tradition, security, power, excitement, achievement, and self-direction. *The mapping of values can help you better understand how your values can influence and motivate your career decisions. The values test is one necessary input to best determine the Information systems specialization track of the student wherein it would help in evaluating on how well the values are integrated in the work.* [6]

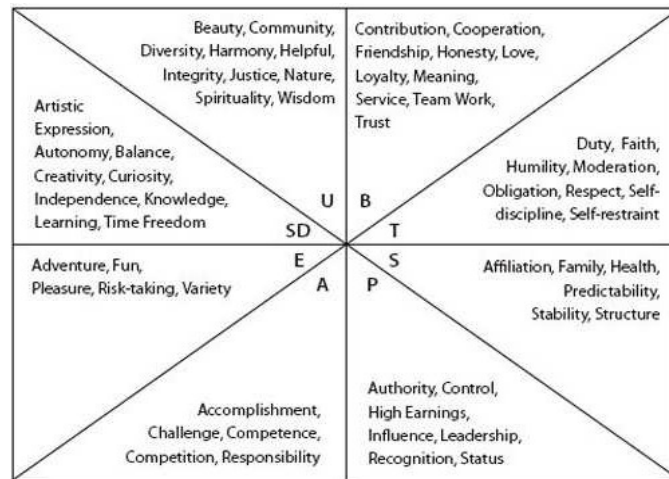


Figure 3. The Value Search Map of Bronwyn and Holt

These value search tests were sent to at least ten of the experienced operations managers, user interface designers, and application developers from the industry computing expert. At least five years of experience in the related field was considered. Determining the minimum years in service, which is five years, is based on the most common requirement of recruitment process in the computing industry.

The same assessment ratings forms will be used when the competency model is utilized in the identification of the most suitable Information Systems' track for the student to pursue.

Table 5. Profile of Respondents

PROFESSION	Number of respondents	Minimum of years in service
(Track1). Operations Manager	10	5
(Track 2). User Interface Designer	10	5
(Track 3). Application Developer	10	5

4. RESULTS AND ANALYSIS

The summation of the most required knowledge and skills of the industry computing experts were computed. The same computation was used in coming up with the top ten most dominant values. Values are then categorized using the Values Search Map [6]. Meanwhile, for the computation of the most dominant interest of the expert, each category in the assessment form was first added and categorized. Finally, results of all forms were summed up and categorized to come up with the top three most dominant interests of computing expert in each track of specialization.

Table6. Result Of Values Survey Based On The Values Search Framework

Types of Values	Application Developer	User interface Designer	Operations Manager
Openness to change:	24	33	16
self direction, excitement			
Self-enhancement	24	29	35
achievement and power			
Self-transcendence	19	26	15
universality and benevolence			
Conformity	21	17	15
Traditions and security			

Table 6 shows the result of the survey that was conducted to determine the most dominant values of experts from application developers, user interface designers, and operations manager. The user interface designers' most dominating value is the openness to change which scored 33 points. While the top values of operations manager falls under the category of self-enhancement with a score of 35 points, which means achievement and power are their highest concerns.

On the other hand, basic knowledge on programming languages and in operating systems is necessary but not required in IT operations manager. Most of the listed knowledge and skills required from application developers and user interface designers were confirmed. Please refer to table 7 and assessment rating form appendix 1, 2, and 3.

Finally, with regard to the top three interests: Operations Manager has Social, Investigative, and Conventional; User Interface Design-Artistic, Social, and Enterprise; Application Developer – Realistic, Investigative, and Conventional. With regard to the dominating values of the computing experts, openness to change and self enhancement are the top values of application developer which both values got the highest score of 24 points as shown in Table 8.

Table 7. Most Required Skills And Knowledge

	Operations Manager	User Interface Designer	Application Developer
1	Business process	HCI	VB.net
2	PPM	CSS	MySQL
3	Accounting management	Photoshop	PHP
4	Economic and accounting principles	Javascript	SQLite
5	Knowledge on Server side technologies	Understanding on diffrent mobile/desktop/web applications	Java
6	Knowledge on usability issues	Java	Oracle
7	CMS	Typography	Javascript
8	Interface design	Layouting	Design and efficiency evaluation and Q.A tools
9	Basic knowledge of Programming languages	Flash	ASP
10	Basic knowledge of operating systems	Correl draw /illustrator	Postgre

Table 8. The Top Three Interests Of Operations Manager, Application Developer, And User Interface Designer

Operations Manager		Applications Developer		User Interface Design	
Interest	Total	Interest	Total	Interest	Total
S	27	I	40	S	29
C	25	C	31	E	28
I	20	R	27	A	25

Top three interests of operations manager as shown on table 8 are Social, Conventional, and Investigative. While for the applications developer top three interests are investigative, conventional, and realistic. User interface designer’s top three interests are, social, enterprise, and artistic.

5. COMPETENCY MODELS FOR INFORMATION SYSTEMS SPECIALIZATION TRACKS

The most dominant values of operations manager is under the category of self-enhancement. Among the three tracks it got the highest score of 35 which is 13% higher than that of user interface designer and 8% higher than the application developer as shown in table

Table 9. Operations Manager Competency Model

Values	Interests	Knowledge and Skills
Self-enhancement Achievement (A)→Accomplishment, Challenge, Competence, Competition, Responsibility Power (P)→Authority, control, High earnings, Influence, leadership, recognition, Status	Social, Conventional, Investigative	BPM, PPM, Economic & Actg. Principles, Client/Server Side technology, Usability Issues, CMS, Interface Designs/HCI, Basic knowledge on operating systems.

Table 10. Applications Developer Competency Model

Values	Interests	Knowledge and Skills
Openness to Change - Self-Direction (SD): Pursues independent thought or action. Enjoys the ability to choose, create, and explore - Excitement (E): Seeks pleasure or sensuous gratification. Enjoys unpredictability and variety in life.	Investigative, Conventional, Realistic	VB.net, query language, PHP, SQLite, Java, Oracle, Javascript, ASP, Design and efficiency evaluation and Q.A tools
Conformity • Tradition (T)→ Respect, commitment, and acceptance of the customs and ideas that one’s culture or religion expects of individuals. • Security (S)→ Desire for safety, harmony, and stability of society, relationships, and self.		

Application developer’s most dominant values are under the categories of openness to change and self-enhancement which got the score of 24 each respectively. However, the conformity category

got the highest score of 21 compared with the scores of operations manager, and user interface designer.

Table 11. User Interface Designer Competency Model

Values	Interests	Knowledge and Skills
Self-transcendence	Social, Enterprise, Artistic	HCI, CSS, Photoshop, JavaScript, Understanding on different mobile/desktop/web applications, Java, Typography, layouting, Flash
Universality (U): Understanding, appreciation, tolerance, and protection for the welfare of people and nature.		
Benevolence (B): Concern for the protection and enhancement of the welfare of people with whom one is in frequent contact.		
Openness to Change		
Self-Direction (SD): Pursues independent thought or action. Enjoys the ability to choose, create, and explore		
Excitement (E): Seeks pleasure or sensuous gratification. Enjoys unpredictability and variety in life.		

The most dominant values of user interface designer falls under the category of openness to change. While the values under self-transcendence got higher scores compared with other tracks.

6. CONCLUSION AND FUTURE WORKS

Based on the result of the survey, it shows that application developer and user interface designer have closely related values. Thus, another round of conducting a survey would be needed to come up with the most exclusive dominant values for the particular information systems specialization track.

For future works, the competency models can be implemented to a computer application that will cater the processing of the identification of Information Systems track to be recommended to the information systems students according to their competency.

REFERENCES

- [1] Philippines National Statistics Office Website. [www. Census.gov.ph/statistics/survey/labor-force](http://www.census.gov.ph/statistics/survey/labor-force). Retrieved on October 18, 2013.
- [2] TIP Quezon City IS ABET SSR. Technological Institute of the Philippines. Quezon City, Philippines
- [3] Singh, Raghav. Generation U. Too Many Underemployed College Grads. <http://www.ere.net/2013/07/19/generation-u-too-many-underemployed-college-grads/>
- [4] Quiang, Li and Shiyang, Wen. A competency Model for Civil Engineering
- [5] Career and Technical Education Center. www.hawaii.edu/cte/publications/RIASEC.pdf
- [6] Bronwyn, Llewellyn and Holt, Robin. Values test. <http://www.netplaces.com/career-tests/values-and-your-career/values-test.htm>. retrieved on october 16, 2013.

APPENDICES**APPENDIX 1**

Assessment Rating Form For Operations Manager – Knowledge And Skills

CHECK THE KNOWLEDGE/SKILLS WHICH ARE REQUIRED FOR USER OPERATIONS MANAGER		
KNOWLEDGE & SKILLS	Required	
	Y	N
Management (people, time, resources)		
Colour skills and design knowledge		
Expert on different operating systems		
Microsoft office applications		
Coding languages: HTML, XHTML,css		
Basic knowledge: PHP, MySQL, JavaScript, JQuery, AJAX		
Content Management Systems(CMS)		
Streaming formats: flash, QuickTime, Windows media,		
Business Process		
In-depth knowledge of usability issues		
Project and planning management		
Economic and accounting principles		
Understanding of server side technologies		
Understanding of Interface design		
Other Application Or Tools, Skills, knowledge (Pls. Indicate: On The Blank Cells Below)		

APPENDIX 2

Assessment Rating Form For User Interface Designer – Knowledge And Skills

CHECK THE KNOWLEDGE/SKILLS WHICH ARE REQUIRED FOR USER INTERFACE DESIGNERS		
KNOWLEDGE & SKILLS	Required	
	Y	N
Cascading Sytle Sheet(Css)/CS5		
Photoshop		
Java		
Javascript		
Flash/illustrator		
Design and Efficiency Evaluation/Q.A Tools		
Layouting		
Color Scheme/Color Theory		
CorelDraw		
HUMAN COMPUTER INTERACTION Principles (HCI)		
Understanding/Knowledge Regarding Differences Of Web/Mobile/Desktop Applications		
Typography		
Multimedia		
Communication Skills		
Leadership Skills		
Other Application Or Tools, Skills, knowledge (pls. indicate: on the blank cells)		

APPENDIX 3

Assessment Rating Form For Application Developer –Knowledge And Skills

CHECK THE KNOWLEDGE/SKILLS WHICH ARE REQUIRED FOR AN APPLICATION DEVELOPER		
KNOWLEDGE & SKILLS	Required	
	Y	N
VB.net		
Sqlite		
Oracle		
MySQL		
Postgre		
PHP		
MSSQL		
Java		
ASP		
JavaScript		
Design and Efficiency		
Evaluation/Q.A Tools		
Other Application Or Tools, Skills, knowledge (Pls. Indicate: On The Blank Cells Below)		

APPENDIX 4

Interests Assessment Rating Form For Operations Manager, User Interface Designer, And Application Developer

1. I like to work on cars	●					
2. I like to do puzzles		●				
3. I am good at working independently			●			
4. I like to work in teams				●		
5. I am an ambitious person, I set goals for myself					●	
6. I like to organize things, (files, desks/effices)						●
7. I like to build things	●					
8. I like to read about art and music			●			
9. I like to have clear instructions to follow						●
10. I like to try to influence or persuade people						●
11. I like to do experiments		●				
12. I like to teach or train people					●	
13. I like trying to help people solve their problems						●
14. I like to take care of animals	●					
15. I wouldn't mind working 8 hours per day in an office						●
16. I like selling things						●
17. I enjoy creative writing			●			
18. I enjoy science		●				
19. I am quick to take on new responsibilities						●
20. I am interested in helping people						●
21. I enjoy trying to figure out how things work	●					
22. I like putting things together or assembling things	●					
23. I am a creative person					●	
24. I pay attention to details						●
25. I like to do filing or typing						●
26. I like to analyze things (problems/ situations)					●	
27. I like to play instruments or sing						●
28. I enjoy learning about other cultures						●
29. I would like to start my own business						●
30. I like to cook					●	
31. I like acting in plays						●
32. I am a practical person						●
33. I like working with numbers or charts					●	
34. I like to get into discussions about issues						●
35. I am good at keeping records of my work						●
36. I like to lead						●
37. I like working outdoors					●	
38. I would like to work in an office						●
39. I'm good at math						●
40. I like helping people						●
41. I like to draw						●
42. I like to give speeches						●

APPENDIX 5

Value Assessment Rating Form For Operations Manager, User Interface Designer, And Application Developer

CHECK THE OTHER TOP 10 VALUES THAT WOULD BE CRITICAL TO JOB SATISFACTION FOR AN APPLICATION DEVELOPER		
accomplishment		Curiosity
adventure		Diversity
affiliation		Duty
Artistic expression		Faith/ spirituality
authority		Family
balance		Friendship
beauty		Fun
challenge		Harmony
community		Health
control		Helpfulness
contribution		High earnings
cooperation		Honesty
Humility		Independence
influence		Integrity
justice		Knowledge
leadership		Learning
love		Loyalty
meaning		Moderation
pleasure		Obligation
predictability		Recognition
respect		Responsibility
Risk taking		Self discipline
Self restraint		Service
wisdom		Stability
structure		Team work
status		Time freedom
trust		Variety
Pls. write other values you think is critical or required:		

EFFECTS OF GOP ON MULTIVIEW VIDEO CODING OVER ERROR PRONE CHANNELS

A.B Ibrahim¹ and A.H Sadka²

¹Department of Electronic & Computer Engineering,
Brunel University, London, United Kingdom
Abdulkareem.Ibrahim@brunel.ac.uk
Abdul.Sadka@brunel.ac.uk

ABSTRACT

In this paper, an investigation of the effects of group of pictures on H.264 multiview video coding content over an error prone environment with varying packet loss rates is presented. We analyse the bitrate performance for different GOP and error rates to see the effects on the quality of the reconstructed multiview video. However, by analysing the multiview video content it is possible to identify an optimum GOP size depending on the type of application used. A comparison is demonstrated for the performances between widely known H.264 data partitioning error resilience technique and multi-layer data partitioning technique with different error rates and GOP in terms of their perceived quality. Our simulation results turned out that Multi-layer data partitioning technique shows a better performance at higher error rates with different GOP. Further experiments in this work have shown the effects of GOP in terms of visual quality and bitrate for different multiview video sequences.

KEYWORDS

Multiview Video Coding, Group of Pictures, Error rates, Bitrate, and Video quality.

1. INTRODUCTION

Three-dimensional technology widely known as 3D technology has transformed many fields of discipline such as entertainment, communications, medicine and many more. 3D can be perceived in a number of different ways. We shall restrict our understanding to just multiview video coding in this paper. Generally, the main concept of video coding is to exploit the statistical correlation between consecutive frames. The MVC extension of the H.264/AVC exploits these similarities between frames, simplifies the decoding process, and advances new features that are specific to multiview video coding [1]. Multiview video coding has emerged as advancement in video coding technology. The multiview video coding system enables efficient encoding of sequences that are captured from different cameras at different locations at the same time. The H.264 MVC codec takes as an input several synchronized bitstream that are captured from several different cameras and generate a single bitstream as an output for storage or transmission [2]. The work in [3] gives a detailed overview of the MVC standard. The structure of MVC is defined by a concept known as matrix of pictures (MOP). In this technique, each row consists of a group of pictures (GOP) normally captured by the base view and each column represents the time domain of the video.

2. BACKGROUND

The H.264/AVC international standard [4] has specifies a coding standard of video data. H264 defines three picture types namely I-frame, P-frame, and B-frame. In a standard reference multiview video encoder all the pictures in a multiview video are encoded with a fixed GOP length depending on the settings and applications. The arrangement of these three picture types in a sequence is distributed statistically within the group-of pictures. The special type of I-frame at the beginning of a sequence also known as an IDR frame serves as an entry point to facilitate random seeking or switching between channels. This can further be used in providing coding robustness to transmission errors [5] which are only coded with moderate compression to reduce the spatial redundancies in the multiview video sequence. I frames are generally larger than P and B frames which means the less you have the longer the GOP size and the more compression you can get. But in multiview video content transmission especially in error prone channels, very long GOP can have an adverse effect of propagating error spatially, temporally and in interview direction. P frames are coded in an efficient way through the concept of motion compensation from either a past I or P frame which are mostly used as a reference to predict further. B frames have a very high compression ratio which requires the presence of both a past and future reference pictures for motion compensation.

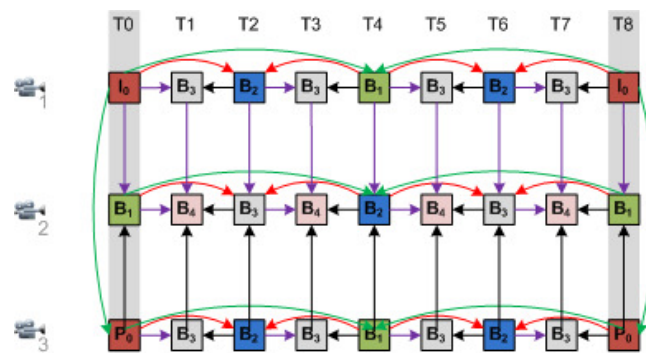


Figure 1. MVC prediction structure with GOP size of 8 [6].

Fig. 1 depicts a multiview video coding prediction structure with GOP size of 8 where I, P, and B represents the encoding of pictures in intra mode, predicted mode and bi-predicted mode respectively. The compressed multiview video data is highly sensitive to noise and information is loss due to the removal of statistical and subjective redundancy in the video by the compression scheme [7]. H.264/AVC employs variable length coding (VLC) in order to achieve higher compression gain. This type of predictive coding technique makes the video data highly sensitive to bit errors, and the effects of errors on the perceptual video quality can be quite severe. Thus, it is necessary to provide an effect technique and configuration settings that can make the MVV bitstream more robust to transmission error and to improve the visual quality of the reconstructed multiview video [8]. The effectiveness of H.264/AVC coding depends on many coding parameters one of which is GOP size and its internal organization [9]. Most standard reference H.264 codecs use a fixed size for the GOP to encode video sequences. The GOP size can have different values as specified by the standard, however, once a given size is chosen, it becomes applicable to the entire coding process and the corresponding standard decoder can be able to sort out the positioning of these frames during decoding process.

2.1. Concept of Data Partitioning in H.264/AVC

The H.264/MPEG-4 AVC standard is established to represent complete video information in a much lower level called the *slice*. A H.264 video slice consists of an arbitrary integer number of successive macroblocks that represent different types of video data [10]. Slice header conveys information that is common to all the MBs in the slice such as the slice types which determine which MBs types are allowed, and frame number that the slice corresponds to, reference picture settings and default quantization parameter. The slice data section consists of a series of MBs that make a slice.

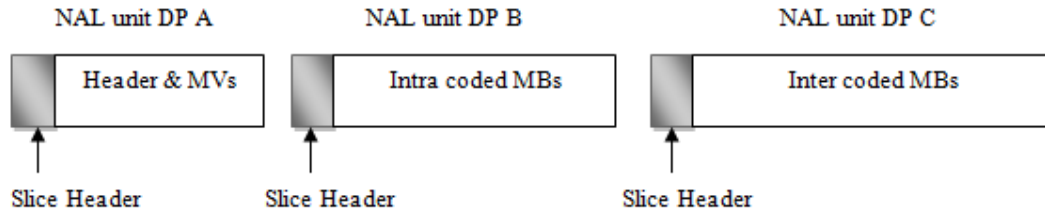


Figure 2. H.264/AVC Slice layout with data partitioning

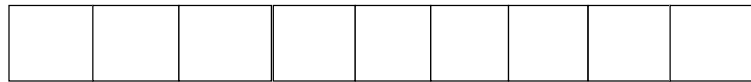


Figure 3. Multi-Layer data partitioning technique

In DP technique, MV and the residual information are separated by a boundary marker which is a uniquely decodable codeword. The codeword indicates the end of header information in a slice and the beginning of residual information [11]. Recent study on the concept of DP can be found in [12]. Data partition, nonetheless, creates more than one bit string (partitions) in every slice, and rearrange all symbols of a slice into a separate partition that have a close semantic relationship with each other Fig. 3. In H.264/AVC, when data partition is enabled, each slice of the coded bitstream is divided into three separate partitions with each of the partitions being from either type A, type B or type C partitions. Type A partition consists of header information, Quantization parameter (QP), Macroblock type, reference indices and motion vectors. The intra partition also called type B consists of the Discrete Cosine Transform (DCT) intra coded coefficients and the inter partitions also known as type C partitions contain DCT coefficients of motion compensated Inter-frame coded MBs. Type C partition in many cases is the biggest partition of a coded slice and yet the least sensitive to error because its information does not synchronise the encoder and the decoder [13]. Each partition is placed in a separate Network Abstraction Layer (NAL) unit and may be transmitted separately over a network. The use of both Types B and C will require a type A partition and not vice-versa.

2.2. Previous Work

The implementation of data partitioning technique for MVC is presented in [14]. A video slice without any ER mechanism may be affected by transmission errors that can lead to the loss of the entire information within the slice. Implementation of error resilience techniques such as data partitioning in the JMVC reference software is necessary because there is no provision for any ER technique in the MVC in the reference software. Therefore, in order to analyse the performance of MVC in error-prone networks, implementation of a valid error resilience technique such as data partitioning as shown in Fig. 2 is employed and implemented in the JMVC 8.5 reference software.

From the H.264 data partitioning technique, a video slice can be recovered when either partitions B or C, or both, are affected by transmission errors as long as the partition A is not affected or lost as a result of losing the header and motion information contained therein. It has been observed that the performance of H.264/AVC data partitioning technique in MVC is not too encouraging and further error performance improvements can be made through the introduction of the proposed multi-layer data partitioning technique depicted in Fig.3.

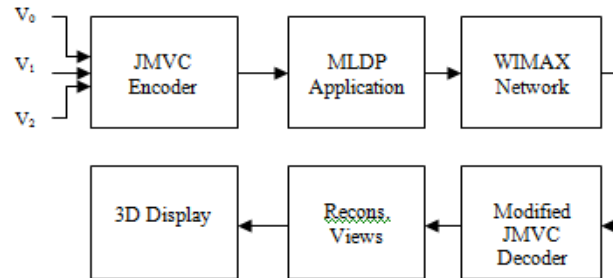


Figure 4. Flow diagram of the Multi-Layer DP technique

2.3. Multi-Layer Data Partitioning Technique

In an attempt to make the MVV bitstream more error resilient to the transmission errors in an error prone network, we propose a technique that can create two-layer of partitioning for each slice in the multiview video bitstream. The general concept of the technique is illustrated in Fig.4. The assembled multiview video bitstream is parsed in the developed Multi-Layer DP application for increased robustness against the transmission errors before transmitting over the wireless network. The partitioned bitstream is received by the modified JMVC reference decoder in order to decode and reconstruct the multiview video bitstream for viewing at the display. Multi-Layer DP adopts a mechanism that restructures a video slice as shown in Fig. 3. A_0 partition consists of the header information of frame 0 from view 0, and A_1 partition consists of the header and motion information of frame 1 from view 1 and A_2 partition consists of the header and motion information of frame 2 from view 2. B_0 consists of the residual information of intra coded MBs of frame 0, B_1 consists of the residual of intra coded MBs in frame 1 and B_2 consists of the residual of intra coded MBs of frame 2 and C_0 is an empty partition, C_1 consists of residuals of inter coded MBs and C_2 consists of the residual of inter coded MBs of frame 2 and in that sequence it continues till nth view and nth last slice of the multiview bitstream.

Note that, partition C_0 is empty because there is no residual information of inter-coded MB's in frame 0 which is an intra-coded frame. I-frames are self-referential and do not require any sort of information from other frames to be predicted, so it consists of only intra coded MBs. The H.264 compliant encoder needs not to send empty partitions to the decoder because a standard H.264 decoder will assume missing partitions are empty partitions and are designed to handle the multiview bitstream accordingly [15]. During the decoding process of the MLDP bitstream, the decoder is modified to cope with the lost video data due to errors in the wireless channel.

The effects of displaying a frame reconstructed from a corrupted data can adversely degrade visual perception by introducing artefacts. In order to support the MLDP technique more effectively and to minimise the effects of channel errors in the multiview video bitstream, a simple and commonly known error concealment technique is employed and developed in the JMVC 8.5 reference decoder. Lost data in the bitstream can be concealed by copying the information from previously received error free slices. Frames that are generated by copying related video data in order to replace lost information are not always perceptually noticeable by a viewer which is an advantage of this technique especially in low-activity scenes [16]. In our

approach, we are able to support multi-layer data partitioning technique with improved quality by employing frame copy error concealment which works fairly well with MVC and is simple to implement; however, there are more complex techniques that use an elaborate approach to exploit the redundancy within the video frame in order to come up with a more efficient estimate of the lost data [3].

Time first coding Fig. 5 is a MVV bitstream format representation that allows all views to be encoded and then assembled in a time domain for suitable transmission. The decoder on the other side can receive and reorder the bitstream in the right decoding order, which can allow it to decode all the pictures in different views in the same time domain and display the videos in the correct order.

Time first coding supports the implementation of frame copy error concealment in MVC. That is because of the display nature of all the frames across the views in the same time domain, which makes it easier to conceal missing pictures from previously received pictures in the reference list.

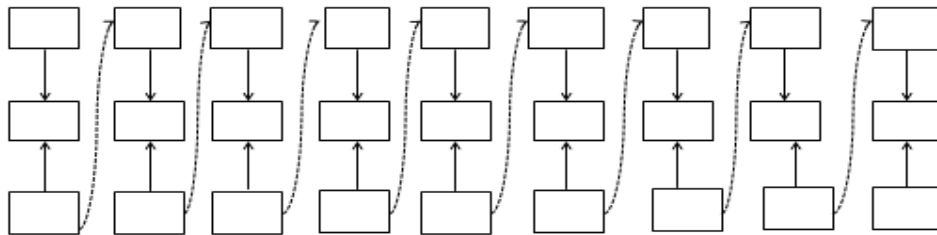


Figure 5. Time first coding [1].

Currently, the MVC reference decoder only accepts H.264 compliant bitstream and does not support the decoding of erroneous coded video sequence. In order to be able to decode the corrupted multiview video bitstream, the H.264/AVC frame copy error concealment technique is implemented in the JMVC reference decoder to adapt and cope with the losses within the bitstream. Frame copy error concealment technique is simple and usually quite effective in a video content where the motion is not large [17]. In Addition, the JMVC 8.5 reference codec has two types of reference frame lists that is also part of the standard and can be used to support frame copy error concealment in MVC. The first list is a reference list 0 which can be used for both P and B frames while reference list 1 is only applicable for B-frames. The main difference between the two reference lists is that list 0 utilizes the temporally earlier key frames (I or P) within the GOP in a sequence while in the case of the reference picture list 1; it utilizes temporally closer reference frames which can be a B frame [18]. Conceptually, reference list 1 can ensure smoother pictures because the frame to be copied is nearer to the picture to be reconstructed.

2.4. Proposed decoding scheme

H.264/AVC Frame copy error concealment technique is implemented in the JMVC reference decoder and further modified to decode the Multi-layer DP bitstream with losses as earlier discussed in the previous section. The technique is optimized to reconstruct all the views successfully from the multiview coded bitstream with a higher level of quality in conformance with the standard [19].

Part of the reasons and motivation to adopt frame copy error concealment technique in our work is its convenience to replace missing pictures especially in the case of packet loss network.

The flowchart in Fig. 6 illustrates the implementation of frame copy error concealment technique. The technique can conceal lost information in the MVV bitstream with an improved perceptual quality based on some experimental results presented later in the paper.

When the ML data partitioned bitstream is transmitted over the network and is received, it is first buffered and rescheduled back to the standard H.264 DP format for processing. Note that, the multi-layer data partitioning technique employed during source coding is only to make the multiview video bitstream more resilient to channel errors during transmission or streaming over the simulated wireless network. After successfully delivering the bitstream across the network, then the received bitstream is rescheduled back to the standard H.264 data partitioned format for decoding.

The decoder checks if the buffer is full then all the frames are sent directly for decoding. Also, note that all the slices are partitioned into three different partitions encapsulated into VCL NAL units of DP A, DP B and DP C respectively. The decoding of these types of slices is such that the loss of one partition might make another partition useless. In order to correctly decode partitions B and C, it is important for the H.264 standard compliant decoder to know how each and every macroblock is predicted within a slice. This information is stored in partition A as part of header information. Therefore, loss of partition A can render partitions B and C useless even when correctly received and decoded. Partition A does not necessarily require the information from partition B and C to be correctly decoded.

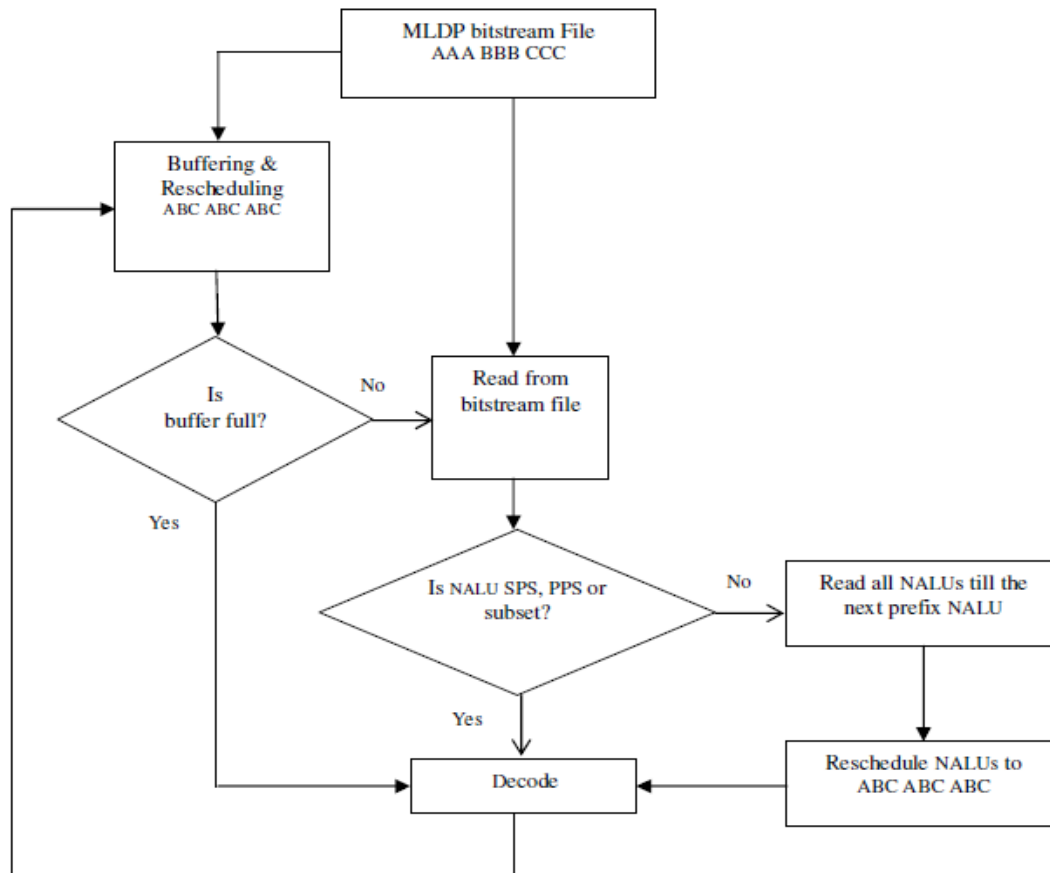


Figure 6. Decoding scheme for erroneous MVV bitstream

So, if only partition A is correctly received then the error concealment algorithm can utilize useful information such as motion vectors to reconstruct the slice. However, if partition A is lost regardless of whether partition B or/and C is/are received. The frame copy error concealment is invoked by the decoder to replace the missing picture information by a previously received picture in the reference list. If the buffer is empty, then the NAL units are read from the MVV bitstream and the decoder determines whether it is a non-VCL NAL unit or VCL NAL unit? All non-VCL NAL units are sent directly for decoding while the VCL NAL units are all read until the next prefix NAL unit is detected and are rescheduled to the H.264 format before decoding. The whole process is restarted again through a looping system.

3. SIMULATION

To show the performance of 3D MVV bitstream over a wireless error-prone network, a number of coding and transmission experiments and simulations are performed in both JMVC 8.5 reference software and OPNET 16.1 network simulator [20]. This section describes the conditions used in the experimental setup.

3.1. Video Encoder Settings

Different MVV test sequences were used in the experiment and simulations such as Ballroom, Exit, and Vassar. Frame size of 640X480, Frame rate = 25 f/s, Number of Frames per view = 250, and Quantization parameter (QP) was carefully selected and set to 31 and an intra-coded frame was inserted every 13th frame in order to limit the temporal error propagation. The JMVC 8.5 reference software and simulations were configured as in [18]. Symbol mode is set on Content Adaptive Variable Length Coding (CAVLC) to support the DP in the extended profile, also one slice per NAL unit is considered as part of the H.264/AVC network friendly design [21]. Three views of each of the MVV test sequences were considered to generate the MVV bitstream used for transmission over the simulated network.

3.2 Transmission Simulation Setup

This section describes some of the necessary conditions and parameters used in the network simulation setup. The robustness of MLDP against channel errors is demonstrated by examining the performance of the MVV bitstream transmitted over a WIMAX simulated channel under varying channel conditions and error rates. Multiple Subscriber Stations (SS) that represents source and destination nodes are configured to share a common Base Station that is connected to the core network through an IP backbone. The WIMAX model in OPNET does not have a direct approach to upload and introduce errors in an MVV bitstream file. Trace file of the MVV bitstream need to be generated first and simulated across the network. Transmission error distribution formats have been developed for different error rates in the model. The network simulation methodology is similar to the work in [22]. The MVV bitstream is decoded and evaluated after being received by the application client for different error rates in the network. The objective quality of the reconstructed videos is measured and analysed in terms of Peak Signal to Noise Ratio (PSNR) which is a widely known objective metric used to measure the reconstructed video quality [23].

3.3. Experimental Results and Analysis

This section describes the performance evaluation and results of the effects of GOP size on multiview video bitstream over the wireless network. The values of GOP sizes used in the experiments are 4, 8, 12, and 16 respectively. Also, the error rates used are 0%, 1%, 5%, 10%,

15%, and 20% respectively. For every GOP size and error rate considered, ten different simulations are conducted, and the average results are generated. The perceptual quality of each reconstructed view is measured in terms of peak signal to noise ratio (PSNR) for all the different simulations and error rates used in the experiment. Ballroom sequence experimental values for perceptual quality are generated in table 1 for different loss rates and GOP sizes

In Fig. 6, we evaluate both the H.264 DP and the multi-layer DP method for different error rates and GOP size. We have observed for different runs in our simulations that high error rate (20%) multi-layer DP has a better and improved quality performance than the H.264 DP technique in many instances. Note that, video coding works either as fixed quality and variable bitrate and vice-versa.

In this experiment for various quality levels with GOP of 4, 8, 12 and 16, corresponding constant bitrates of 1909.69kb/s, 1619.76kb/s, 1527.94kb/s, and 1374.75kb/s are respectively reported for ballroom test sequence Fig. 7.

In Fig. 8 and 9, the results of the experiment have revealed that a small number of GOP size means more I frames. This can have a tendency to consume more of bits because of the frequent occurrence of intra frames within the GOP. However, having more I-frames increases the multiview bitstream size. It can have a tendency of reducing the efficiency of the multiview video coding. Different applications can have different GOP requirements such as real time and offline applications each having a different latency or delay requirement [24].

In Fig. 10, the results obtained illustrate that lower GOP size can give a better perceptual quality in the multiview video. This is because low GOP means more intra frames within the GOP with less prediction error which can result in a higher video quality. In video communications over-error prone environment, trade-off between perceptual quality and bitrate consumption is important and necessary [25]. In most cases, applications requiring a high level of quality in an error-prone network can have a higher bitrate in order to make the MVV bitstream more resilient to channel noise and that result in visual quality improvement. [26]

3.3.1. Objective and Subjective analysis

Table 1. Numerical simulation results

Ballroom GOP4			Ballroom GOP8	
PLR (%)	H264 DP (dB)	H264 ML (dB)	H264 DP (dB)	H264 ML(dB)
0	35.45	35.45	35.16	35.16
1	34.53	34.93	34.67	34.72
5	28.54	28.90	30.28	27.97
10	24.73	24.37	26.82	24.96
15	21.04	22.93	21.35	21.90
20	18.65	20.04	18.09	19.04

Ballroom GOP 12			Ballroom GOP 16	
PLR (%)	H264 DP (dB)	H264 ML (dB)	H264 DP (dB)	H264 ML(dB)
0	34.99	34.99	34.83	34.83
1	34.74	32.83	34.38	33.41
5	30.42	30.10	30.42	31.82
10	24.24	24.22	24.61	25.59

15	20.94	21.63	19.23	22.52
20	18.23	20.09	16.01	19.17

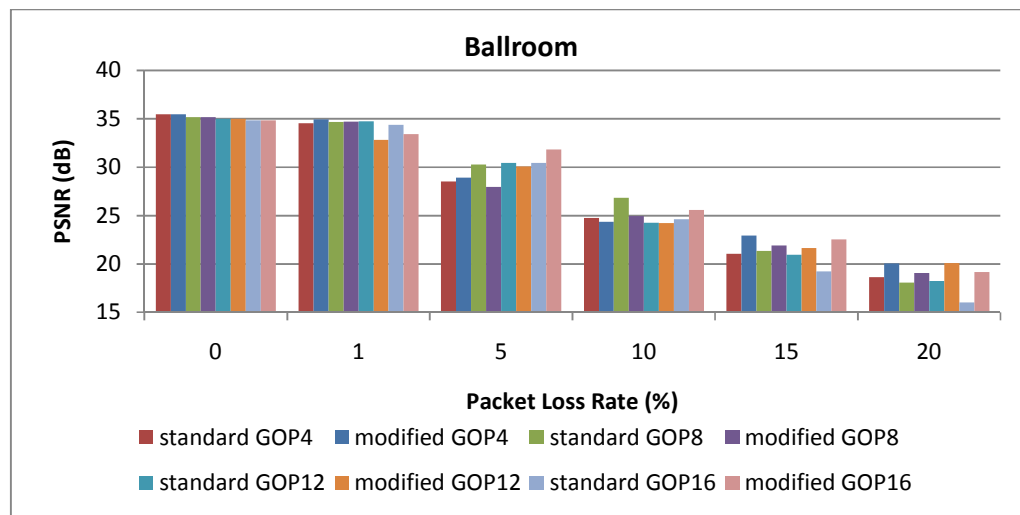


Figure 6. Quality evaluation for different error rates and GOP sizes

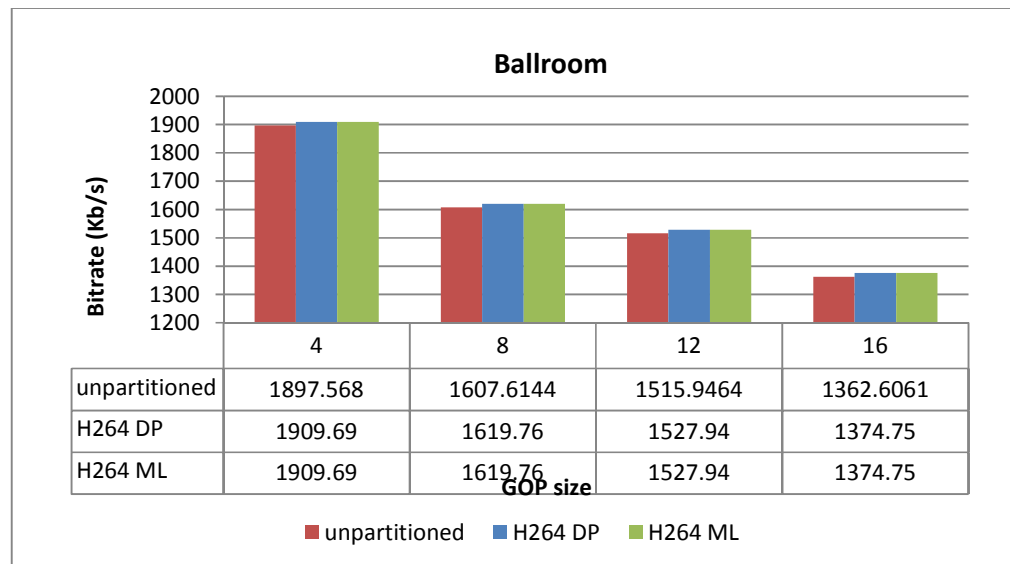


Figure 7. Bitrate performance for different GOP sizes

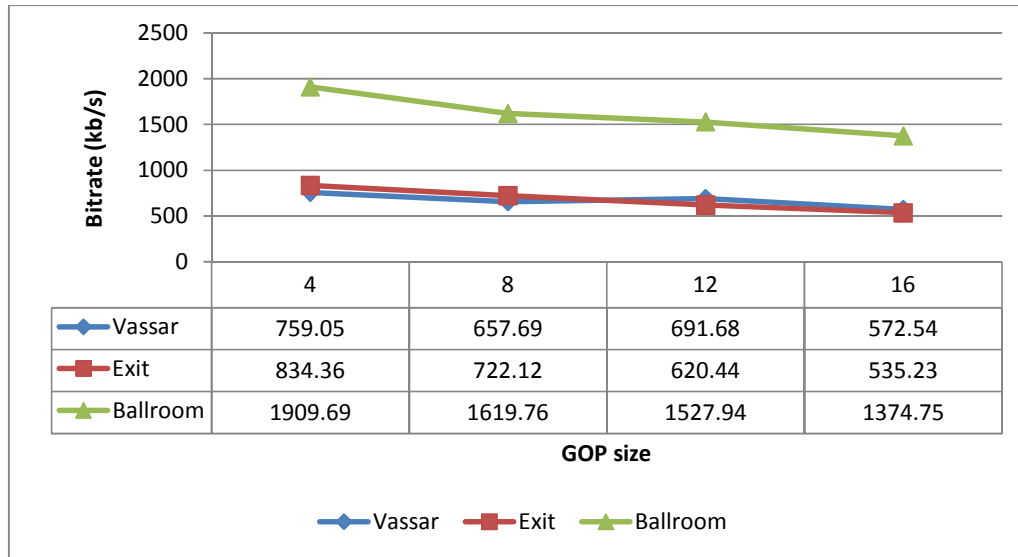


Figure 8. Bitrate performance for different GOP and test sequences

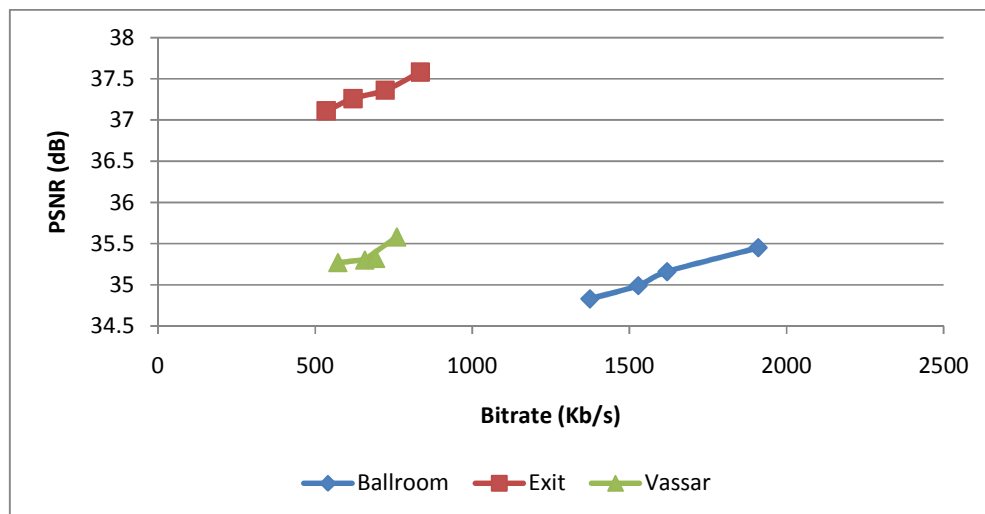


Figure 9. Quality and bitrate evaluation for different test sequences

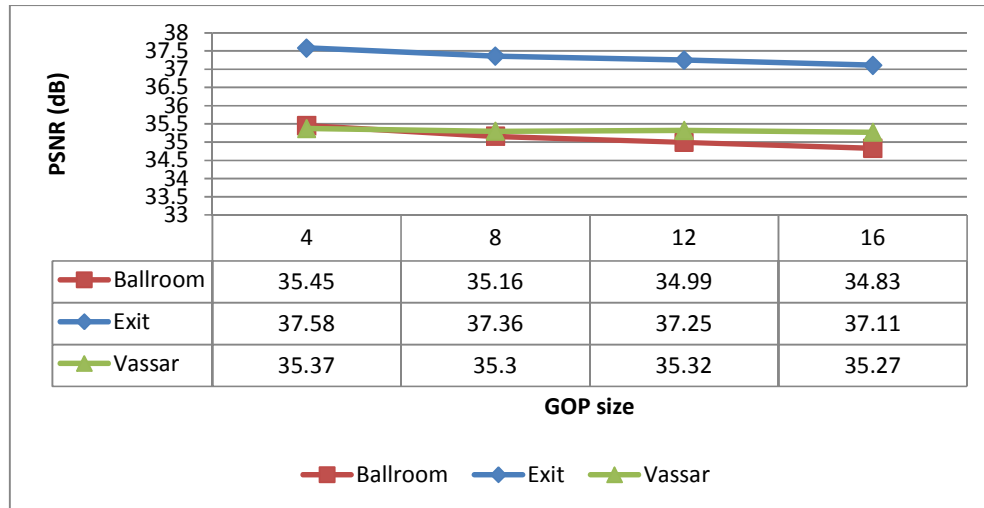


Figure 10. Quality evaluation for different GOP sizes and test sequences

The subjective results are presented for ballroom sequence for different views in Figures 11-13. In the demonstration, frame 121 is chosen from each view at 20% loss rate and a GOP of 16. It can be observed that Multi-Layer DP technique can improve the perceptual quality performance than H.264 DP technique. The greyscale effect in Multi-layer DP technique is completely removed. We can observe closely in the Multi-layer DP that these frames are not reconstructed with the best quality when compared with the original frames. This is because of the high error rate used in the network simulations and the inability of the frame copy error concealment to recover high losses. At such error rate of 20% and GOP of 16, the multi-layer DP technique could recover most of the lost video information with improved quality compared to H264 DP technique at the same error rate and GOP size. It is important to analyse the effects of error propagation within a GOP of the multi-layer data partitioned bitstream. In hierarchical GOP like the one in multiview video coding, the reference decoder uses the I-frame in the base view and the anchor frames in the non-base view either directly or indirectly as reference frames for all other frames with the GOP. If an error occurs in the I-frame of view 1, it can result to artefacts that can continue to propagate throughout the GOP structure. The effect can be experienced in both temporal and interview manner until the next random access point. At this point, the decoder refreshes with the next intra coded frame in view 1 or the anchor frames in either view 2 and 3. It has been noticed that losses within the I-frame that does not affect the header information such as intra coded MBs coefficient can also propagate errors throughout the GOP. P-frames are coded using motion compensation prediction from previous reference frames. From Fig. 1, anchor frame such as the one in view 3 is forward predicted from the I-frame in view 1, subsequent prediction of other non-anchor frames in both view 3 and view 2 takes reference from their preceding P-frame. Any form of loss in this frame can further propagate error through the remainder of the GOP until the next refresh frame is received within the multi-layer partitioned bitstream. It can be highlighted that the impact of P-frame or anchor frame of view 3 can be almost as significant as losing an I-frame due many of interdependencies from other frames. Due to the hierarchical nature of MVC bitstream, anchor frame in view 2 that is interview predicted from view 1 and view 3 is used to predict other non-anchor frames temporally within the GOP. So the effect of error is limited to view 2 only and less severe than I and P-frames in the multiview video bitstream.



Original



H264 DP



ML DP

Figure 11. Subjective quality comparison of frame 121 of view 0 at 20% error rate and GOP=16 for Ballroom sequence.



Original



H264 DP



ML DP

Figure 12. Subjective quality comparison of frame 121 of view 1 at 20% error rate and GOP=16 for Ballroom sequence.



Original



H264 DP



ML DP

Figure 13. Subjective quality comparison of frame 121 of view 2 at 20% error rate and GOP=16 for Ballroom sequence.

4. CONCLUSIONS

The GOP within a video sequence is one of the key coding parameters that determine the video quality perception of the viewer, more importantly, the GOP size and the motion within the sequence. Large GOP size improves the compression efficiency, which can allow more or higher video content to be transmitted for a given bitrate. However, the effects of error propagation or artefacts due to transmission error in an IP network might be longer. It is necessary to wisely decide what GOP structure and size to support any application such as streaming or transmitting videos. The work in this paper examines the effect of GOP size on erroneous multi-layer data partition bitstream when transmitted over error-prone networks. However, the study in this paper focuses and illustrates the performance of the two algorithms for worst case scenario. Two different techniques namely H264 DP and multi-layer DP are used to demonstrate this effect. Our experimental results illustrate that the Multi-Layer DP technique can improve the visual perception of reconstructed videos for higher error rates within allowable compression efficiency and bitrate. From the results obtained, we can assume and suggest that multi-layer DP technique can suitably be utilized for delivering multiview video content over bandwidth constraint and high error rate channel at a GOP size of 16. Please note that the work in this paper is not claiming to achieve a remarkable visual quality. We are proposing based on simulated results a different approach that can apparently improve the visual quality of multiview video in a very high error rate channel. Part of our future work is to optimize the multi-layer data partitioning technique by implementing error protection technique. The idea is to protect the multiview data from the high error rate in the channel. The decoder error concealment algorithm is going to be extended to employ the hybrid method that can fully exploit the redundancies between macroblocks in both spatial/temporal and interview direction. We anticipate that from our current findings and results, more and better visual quality can be achieved when these techniques are implemented while considering the cost of bit rate and coding efficiency.

ACKNOWLEDGEMENTS

The authors would like to thank the Petroleum Technology Trust Fund (PTDF) for the research sponsorship.

REFERENCES

- [1] Y. Chen, Y. Wang, K. Ugur, M. M. Hannuksela, J. Lainema and M. Gabbouj, "The emerging MVC standard for 3D video services," *EURASIP Journal on Applied Signal Processing*, vol. 2009, pp. 8, 2009.
- [2] P. A. Akiki and H. W. Maalouf, "A two-stage encoding scheme for holographic data transmission," in *Multimedia and Ubiquitous Engineering (MUE)*, 2011 5th FTRA International Conference on, 2011, pp. 138-142.
- [3] M. Ebian, M. El-Sharkawy and S. El-Ramly, "Enhanced dynamic error concealment algorithm for multiview coding based on lost MBs sizes and adaptively selected candidates MBs," in *Proceedings of the Fourth International Conference on Signal and Image Processing 2012 (ICSIP 2012)*, 2013, pp. 435-443.
- [4] A. Hermans, "H. 264/MPEG-4 Advanced Video Coding," 2012.
- [5] T. Fang and L. Chau, "An error-resilient GOP structure for robust video transmission," *Multimedia, IEEE Transactions on*, vol. 7, pp. 1131-1138, 2005.
- [6] H. Mohib, "End-to-end 3D video communication over heterogeneous networks," 2014.
- [7] A. Vetro, J. Xin and H. Sun, "Error resilience video transcoding for wireless communications," *Wireless Communications, IEEE*, vol. 12, pp. 14-21, 2005.
- [8] S. Khan, Y. Peng, E. Steinbach, M. Sgroi and W. Kellerer, "Application-driven cross-layer optimization for video streaming over wireless networks," *Communications Magazine, IEEE*, vol. 44, pp. 122-130, 2006.

- [9] B. Zatt, M. Porto, J. Scharcanski and S. Bampi, "Gop structure adaptive to the video content for efficient H. 264/AVC encoding," in Image Processing (ICIP), 2010 17th IEEE International Conference on, 2010, pp. 3053-3056.
- [10] I. E. Richardson, The H. 264 Advanced Video Compression Standard. John Wiley & Sons, 2011.
- [11] M. Sun, Compressed Video Over Networks. CRC Press, 2000.
- [12] L. Al-Jobouri, M. Fleury and M. Ghanbari, "Protecting H. 264/AVC data-partitioned video streams over broadband WiMAX," Advances in Multimedia, vol. 2012, pp. 10, 2012.
- [13] S. Wenger, "H. 264/avc over ip," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 13, pp. 645-656, 2003.
- [14] A. B. Ibrahim and A. H. Sadka, "Implementation of error resilience technique in multiview video coding," in IEEE Southwest Symposium on Image Analysis and Interpretation, San Diego, California, 2014, pp. 1-4.
- [15] Y. Dhondt, S. Mys, K. Vermeirsch and R. Van de Walle, "Constrained inter prediction: Removing dependencies between different data partitions," in Advanced Concepts for Intelligent Vision Systems, 2007, pp. 720-731.
- [16] O. Hohlfeld, "Stochastic packet loss model to evaluate QoE impairments," PIK-Praxis Der Informationsverarbeitung Und Kommunikation, vol. 32, pp. 53-56, 2009.
- [17] Y. Wang and Q. Zhu, "Error control and concealment for video communication: A review," Proc IEEE, vol. 86, pp. 974-997, 1998.
- [18] G. J. Sullivan, P. N. Topiwala and A. Luthra, "The H. 264/AVC advanced video coding standard: Overview and introduction to the fidelity range extensions," in Optical Science and Technology, the SPIE 49th Annual Meeting, 2004, pp. 454-474.
- [19] I. Rec, "H. 264 & ISO/IEC 14496-10 AVC," Advanced Video Coding for Generic Audiovisual Services.ITU-T, 2003.
- [20] Z. Lu and H. Yang, Unlocking the Power of OPNET Modeler. Cambridge University Press, 2012.
- [21] T. Stockhammer and M. Bystrom, "H. 264/AVC data partitioning for mobile video communication," in Image Processing, 2004. ICIP'04. 2004 International Conference on, 2004, pp. 545-548.
- [22] J. Yi, Protocole De Routage À Chemins Multiples Pour Des Réseaux Ad Hoc, 2010.
- [23] S. Yasakethu, C. T. Hewage, W. A. C. Fernando and A. M. Konoz, "Quality analysis for 3D video using 2D video quality models," Consumer Electronics, IEEE Transactions on, vol. 54, pp. 1969-1976, 2008.
- [24] D. Wu, Y. T. Hou and Y. Zhang, "Transporting real-time video over the Internet: Challenges and approaches," Proc IEEE, vol. 88, pp. 1855-1877, 2000.
- [25] M. Flierl and B. Girod, "Generalized B pictures and the draft H. 264/AVC video-compression standard," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 13, pp. 587-597, 2003.
- [26] A. Aggoun, P. Amon, I. Arbel, A. Chernilov, J. Cosmas, G. Garcia, A. Jari, S. Keller, M. Mattavelli and C. Kontopoulos, "Multimedia delivery in the future internet," 2008.

Abdulkareem Bebeji Ibrahim received the B.ENG. degree in electrical engineering from Bayero University Kano, Nigeria, in 2005, and the MSc. degree in satellite communication and space systems from the University of Sussex, Brighton, United Kingdom, in 2011. He is currently pursuing his PhD. degree in electronic and computer engineering at Brunel University London. His current research interests include error resilience and concealment for 3D multiview video coding and perceptual 3D multiview video quality.



Professor Sadka received the Ph.D. degree in electrical and electronic engineering from Surrey University, Surrey, UK, in 1997. He has nearly 20 years' worth of academic experience and a long track record of scientific leadership in the area of Video Processing and Communications. He is the former Head of the Department of Electronic and Computer Engineering at Brunel University and the Founding Director for the Centre for Media Communications Research.



He has over 200 publications in refereed journals and conferences 3 patents and a specialised book entitled "Compressed Video Communications" published by Wiley in 2002. To date, he has managed to attract circa £4M worth of research grants and contracts and has graduated 20 PhD students. He is widely supported by industry and runs his consultancy company VIDCOM. He is a fellow of the IET, a fellow of the HEA and a senior member of the IEEE.

MOLECULAR DYNAMICS SIMULATION MODEL OF AFM-BASED NANOMACHINING

Rapeepan Promyoo^{1,3}, Hazim El-Mounayri^{1,3,*} and Kody Varahramyan^{2,3}

¹Department of Mechanical Engineering, Indiana University Purdue University
Indianapolis, Indianapolis, IN, USA

²Department of Electrical and Computer Engineering, Indiana University
Purdue University Indianapolis, Indianapolis, IN, USA

³Integrated Nanosystems Development Institute (INDI),
IUPUI, Indianapolis, IN, USA

*Corresponding author (helmouna@iupui.edu)

ABSTRACT

In this paper, a developed three-dimensional Molecular Dynamics (MD) model for AFM-based nanomachining is applied to study mechanical indentation and scratching at the nanoscale. The correlation between the machining conditions, including applied force, depth, tip speed, and defect mechanism in substrate/workpiece is investigated. The simulations of nanoscratching process are performed on different crystal orientations of single-crystal gold substrate, Au(100), Au(110), and Au(111). The material deformation and deformed geometry are extracted from the final locations of atoms, which are displaced by the rigid indenter. The simulation also allows for the prediction of forces at the interface between the indenter and substrate. Material properties including modulus of elasticity and hardness are estimated. It is found that properties vary significantly at the nanoscale. In addition to the modeling, an AFM is used to conduct actual indentation and scratching at the nanoscale, and provide measurements to which the MD simulation predictions are compared. Due to computational time limitation, the predicted forces obtained from MD simulation only compares well qualitatively with the experimental results.

KEYWORDS

AFM-based Nanomachining, Molecular Dynamics (MD), Nanoindentation, Nanoscratching, Simulation

1. INTRODUCTION

Atomic force microscope (AFM) has been considered a potential manufacturing tool for operations including machining, patterning, and assembling with in situ metrology and visualization [1]. AFM-based nanomachining generally involves nanoindentation and nanoscratching, which have been commonly used in the characterization of surfaces or small-scale materials [2]. It also has the ability to perform in situ repair/re-manufacturing of the position, size, shape, and orientation of single nanostructures. Some applications of AFM-based nanomachining include fabrication of micro-/nano-devices, individualized biomedicine and drug delivery, molecular reading and sorting, ultrahigh density memory, nanoscale circuitry, and fabrication of metal nanowires [3-13].

Recently, AFM tips have been used as cutting tools for surface modification. Nanochannels, nanoslots, and complex nanopatterns can be fabricated by directly scratching the substrate [9]. These AFM-based mechanical indentation and scratching techniques have been successfully applied to produce complex geometries and high aspect-ratio 3D nano-objects on both flat and curved surfaces [10]. Nanoindentation and nanoscratching are capable of fabricating complex structures, and advances in materials, pattern transfer processes, and cost reductions of AFM equipment have allowed these methods to become a viable but not yet scalable method for many nanoscale devices [14]. Process throughput is low due to limited removal speed, tip-surface approach, contact detection, desired force profile, and tool wear. Parallel fabrication using multiple AFM tip arrays has been reported [5]. However, parallel fabrication currently does not allow precise control over size, shape, position, or orientation of individual structures. A fundamental understanding of substrate deformations/separations and the tip is needed to achieve controllable nanomanufacturing [1]. Attempts have been made to study the correlation between machining parameters, machined geometry, and surface properties for better control of AFM-based nanomachining processes both experimentally [15-21] and computationally [22-55]. This includes experiments on few types of materials to investigate the effects of parameters such as applied load, scratching speed, feed rate, scratching direction, tip geometry, tip angle, tip radius, and number of scratching cycles. These parameters which also depend on material properties and crystal orientation of the substrate, affect the depth, width, chip formation, and surface roughness of the machined surface. Due to experimental limitations, computational models are therefore essential to achieve a more comprehensive/complete understanding of the roles of the parameters affecting the final nano-geometry in AFM-based nanomachining. On the other hand, a more extensive experimental study is necessary to inform the development of accurate and realistic predictive models. The experimental data is also needed to validate the computational models.

To address the need for computational models of AFM-based nanomachining, some efforts have been made to model nanoindentation and nanoscratching using MD simulations [22-55]. MD simulation presents itself as a viable alternative to the expensive traditional experimental approach. Such a simulation was initiated in the late 1950s by Alder and Wainwright [56-57] in the field of statistical mechanics and has been successfully applied to investigate various phenomena at nanoscale. The advantage of MD simulation over continuum model simulation (FE) is that it allows for a better, more detailed understanding of the ways defects are created, the transition from elastic to plastic behavior, and crystal structure effects in materials [22]. Numerous studies have been reported on MD simulations of nanoindentation and nanoscratching. The effects of several parameters such as crystal orientation [41, 45, 46], indenter shape and orientation [33, 39, 40, 44], penetration or scratching depth [37, 42, 47, 48], scratching speed [47, 48], feed (on nanoscratching) [34, 35], and temperature [25, 45, 49] have been investigated on different types of bulk and thin film materials. In addition, mechanical properties including Young's modulus, friction coefficient and hardness of materials have also been reported [26, 42]. MD simulation quality depends on the accuracy of the potential energy function used. Also, the complexity of the potential energy function directly affects computational time. The selection of the potential function depends on material type. Various types were investigated in MD simulations: silicon [22, 30, 31], gold [32], copper [25, 33-35], aluminum [36-38], silver [39, 40], iron [41, 42] and nickel [43, 44]. However, MD simulation involves the interaction of a large number of atoms as deformation occurs on an atomic scale. One major concern in MD simulation is the high computational time required. Existing MD models are limited in the size of simulated volume as well as time scale, inhibiting the ability to capture all important attributes for deformation. To keep the processing time under control, most existing models of nanoindentation use less than 100,000 atoms. The largest models of nanoindentation found in the literature contain approximately 10 million atoms [58], which are enabled by parallel computing.

In this paper, three-dimensional MD simulations of AFM nanoindentation and nanoscratching are performed to investigate the effects of tip speed and crystal orientation for the case of gold material. The simulation allows for the prediction of forces at the interface between an indenter

and a substrate. The material deformation and deformed geometry are extracted based on the final locations of the atoms, which have been displaced by the rigid tool. Mechanical properties including Young's modulus and hardness of materials are also reported. In addition, an AFM is used to conduct actual indentation and scratching at the nanoscale, and provide data with which to validate MD simulation. The results of the simulation as well as the AFM data are presented and compared.

2. METHODOLOGY

MD simulation is used to simulate the time dependent behavior of a molecular system. MD simulations of AFM-based nanomachining in this study are implemented using LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) [59, 60]. The LAMMPS code run in parallel uses distributed memory message passing techniques and spatial decomposition of simulation domain. The inputs required in MD simulation are initial positions and velocities of atoms in the system along with other information such as boundary conditions, potential energy function, time steps, etc. The outputs of the simulation include trajectories of atoms in the system, forces, energy of the system, and other physical quantities of interest. The MD simulation model and the potential functions used in this study are explained in the following sections.

2.1. Simulation Model

The schematic model used in the MD simulation of AFM nanoscratching is shown in Figure 1. The simulation model consists of a single crystal gold workpiece and a three-sided pyramidal indenter. Diamond is selected as indenter tip. The indenter tip is modeled as a rigid body. The initial positions of atoms in the model are calculated from the default lattice position. For example, face center cubic (fcc) structure is applied in the modeling of gold workpieces. On the other hand, diamond structure is used for modeling of diamond indenter. The workpiece in the MD simulation is divided into three different zones: boundary, thermostat, and the Newtonian zones. A few layers of boundary and thermostat atoms are placed on the bottom side of the workpiece. Fixed boundary conditions are applied to the boundary atoms. The atoms are fixed in the position to reduce the edge effects and maintain the symmetry of the lattice. Periodic boundary conditions are maintained along the x- and y-direction. The periodic boundary conditions are usually employed when a simulation seek to investigate the behavior of an isolated system, to avoid spurious edge effects and thereby simulate the behavior of a much larger crystal system. The thermostat zone is applied to the MD simulation model to ensure that the heat generated during the indentation process can be conducted out of the indentation region properly. The temperature in the thermostat zone is maintained by scaling the velocities of the thermostat atoms for each computational time step. In the Newtonian zone, atoms move according to Newton's equation of motion.

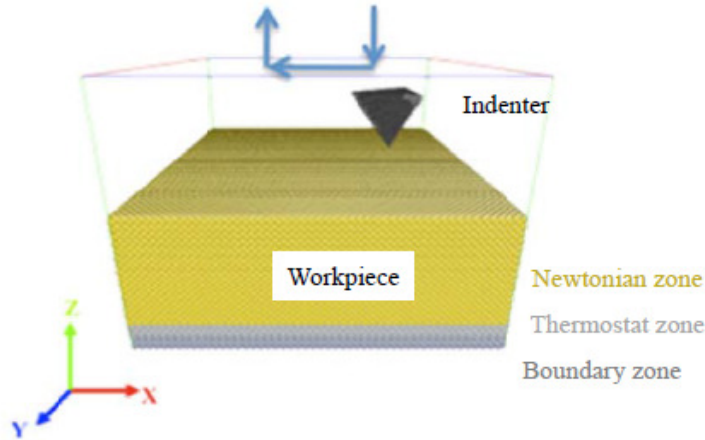


Figure 1. Schematic MD simulation model of AFM nanoscratching

2.2. Potential Energy Function

The motion of the atoms in the Newtonian zone is determined by the forces derived from potential energy function and Newton's equation of motion. The interaction of each atom can be approximated by a potential energy function in accordance with Newtonian mechanics. The quality of the MD simulation results depends on the accuracy of the potential energy function used. On the other hand, the complexity of the potential energy function directly affects the computational time [61]. The selection of the potential function depends on the type of material used in the model. The potential energy function used for the interaction between atoms in the gold (Au) workpiece materials is the Embedded Atom Method (EAM) potential [62]. The Morse potential [63] is employed for the interaction between the gold workpiece and diamond indenter tip in the MD simulations.

The Morse potential [63] is a commonly used empirical potential energy function for bonded interactions. The Morse potential energy function U can be expressed as a function of interatomic distance r as the following formula:

$$U(r) = D \left\{ \exp[-2\alpha(r - r_e)] - 2 \exp[-\alpha(r - r_e)] \right\} \quad (1)$$

where r is the distance between the atoms, r_e is the equilibrium bond distance, D is the cohesive energy, and α is a parameter controlling the width of the potential. The single independent variable in the equation is r . The constant parameters, r_e , α , and D , can be determined on the basis of the physical properties of the material. The parameters used in the Morse potential for gold are listed in Table 1. The parameters between gold and carbon (Au-C) are calculated from the following equations.

$$D_{Au-C} = \sqrt{D_{Au} \cdot D_C} \quad (2)$$

$$\alpha_{Au-C} = \sqrt{\alpha_{Au} \cdot \alpha_C} \quad (3)$$

$$r_{e\ Au-C} = \sqrt{r_{e\ Au} \cdot r_{e\ C}} \quad (4)$$

The EAM potential [62] is an extension of the two-body potential that has been developed for metals. The basic approach of the EAM, which evolved from the density-function theory, is based upon the recognition that the cohesive energy of a metal is governed not only by the pair-wise

potential of the nearest neighbor atoms, but also by embedding energy related to the electron gas that surrounds each atom.

Table 1. Parameters used in the Morse potential energy function

Parameter	Au-Au [64]	C-C [25]	Au-C
D (eV)	0.475	2.423	1.073
α (\AA^{-1})	1.583	2.555	2.011
r_e (\AA)	3.024	2.522	2.762

The interatomic force between any two atoms can be obtained from the potential energy function (U) such that

$$F_{ij} = -\frac{\partial U}{\partial r_{ij}} \quad (5)$$

where F_{ij} is the interatomic force between atom i and j at a distance r_{ij} from atom i . The total force exerted on a particular atom is then calculated as the following equation.

$$F_i = \sum_{j=1, i \neq j}^N F_{ij}(r_{ij}) \quad (6)$$

where F_i is the resultant force on atom i and N is the total number of atoms. After calculating force on each atom, velocities and positions are calculated from Newton's second law of motion.

In this study, material properties, Young's modulus and hardness, are calculated using the formulations developed by Oliver and Pharr [65]. They used data directly drawn from the load-displacement curve and correlated the projected contact area, A_c , to the contact depth, h_c , where h_c may be expressed as

$$h_c = h_{max} - 0.72 \frac{P_{max}}{S_{max}} \quad (7)$$

where h_{max} is the maximum depth of indentation, P_{max} is the maximum applied load and S_{max} is the slope of the unloading curve at the maximum applied load. The contact area, A_c , is thus found from the geometry of the indenter as a function of the contact depth, h_c . Once the contact area is known, the hardness, H , is estimated from the maximum indentation load P_{max} divided by the projected contact area, i.e.

$$H = \frac{P_{max}}{A_c} \quad (8)$$

The Young's modulus is calculated by the reduced elastic modulus, E_r , which takes into account the combined elastic effects of indenter tip and sample, as follows:

$$E_r = \frac{1}{2} \sqrt{\frac{\pi}{A_c}} \frac{dP}{dh} \quad (9)$$

where dP/dh is the slope of tangent line at the beginning of the unloading curve and A_c is the projected area at the maximum depth of indentation. The Young's modulus of the sample, E_s , is then calculated from the following equation.

$$\frac{1}{E_r} = \frac{1-\nu_s^2}{E_s} + \frac{1-\nu_i^2}{E_i} \quad (10)$$

where E_i is the Young's modulus of the indenter, and ν_s and ν_i are the Poisson's ratios of the sample and indenter, respectively.

2.3. Ensembles of Statistical Thermodynamics

Statistical ensembles are usually characterized by fixed values of thermodynamic variables such as energy, E , temperature, T , pressure, P , volume, V , particle number, N , or chemical potential, μ . One fundamental ensemble is called the microcanonical ensemble and is characterized by constant particle number, N , constant volume, V and constant total energy, E , and is denoted as the NVE ensemble. Other examples include the canonical, or NVT ensemble, the isothermal-isobaric or NPT ensemble, and the grand canonical or μ VT ensemble. In the current study, microcanonical or NVE ensemble is applied in the Newtonian zone. The system is isolated from changes in number of atoms (N), volume (V) and energy (E). It corresponds to an adiabatic process with no heat exchange. A microcanonical molecular dynamics trajectory may be seen as an exchange of potential and kinetic energy, with total energy being conserved.

2.4. Parallel MD Simulation

The parallel MD simulations of AFM-based nanomachining are implemented using LAMMPS [59, 60]. The LAMMPS code run in parallel uses distributed memory message passing techniques and spatial decomposition of simulation domain. In spatial decomposition, the simulation domain is divided into a set of equal smaller sized domains. Each sub-domain is distributed to different processor for calculation. Since nearby atoms are placed on same processor, only neighboring atoms on different processor need to be communicated by Message Passing Interface (MPI). Communication is minimized to optimal level by replicating force computations of boundary atoms. Non-uniformity of data distribution can occur for spatial decomposition as interaction between tool and workpiece arise. The parallel MD simulation is run on the Big Red II supercomputer [66]. Big Red II is Indiana University's main system for high-performance parallel computing. Big Red II combines the longstanding leadership of Cray supercomputers with IU-developed technology. The Cray XE6/XK7 supercomputer is capable of one thousand trillion floating-point operations per second, or one petaFLOPS, making it the fastest university-owned supercomputer in the world.

2.5. MD Simulation Conditions

MD simulations of AFM-based nanomachining were conducted on single crystal gold with the use of parallel computing. Table 2 gives the conditions used in the MD simulations of AFM-based nanomachining. The dimensions of the workpiece and indenter, the depth of indentation and the tip speeds are given. The dimensions of the workpiece are expressed in terms of the lattice constants. The lattice constant of gold (aAu) is 4.080 Angstroms (\AA).

Table 2. MD simulation conditions used in the MD simulations of AFM-based nanomachining

Workpiece material	Gold (Au)
Workpiece dimension	Indent: $120a_{Au} \times 120a_{Au} \times 120a_{Au}$ Scratch: $160a_{Au} \times 320a_{Au} \times 40a_{Au}$
Crystal orientation	Au: (100), (110), (111)

Number of atoms in the workpiece	Indent: 6,912,000 atoms Scratch: 8,192,000 atoms
Indenter tip material	Diamond
Indenter type	Three sided pyramid
Indentation depth	1 - 7 nm
Nanoindentation tip speed	1, 10 m/s
Bulk temperature	293 K
Time steps	1 fs (10^{-15} s)

3. EXPERIMENTAL SETUP

A Veeco Bioscope AFM was used to conduct actual indent and scratch at the nanoscale, and provides data for evaluation of the MD simulation predictions. The AFM provides resolution on the nanometer (lateral) and angstrom (vertical) scales. A diamond probe (Bruker DNISP indentation probe) with a spring constant of 250 N/m was used in the experiments. The indenter tips have three-sided pyramid shapes. Nanoindentation is made by forcing the tip into the workpiece until the required cantilever deflection is reached. The tip is then lifted to its initial position above the workpiece. Nanoindentation can be made at various forces and rates, using the deflection of the cantilever as a measure of the indentation force. The indentation force, F , is directly proportional to the deflection of the cantilever can be calculated from the well-known Hooke's law:

$$F = kx \quad (11)$$

where k is the cantilever stiffness or spring constant in N/m and x is the deflection of the cantilever. Nanoscratching is performed by forcing the tip into the workpiece until the required cantilever deflection is reached. The tip is then moved horizontally for a specified length and then lifted to its initial position above the workpiece. The nanoindentation and nanoscratching experiments were conducted at various applied forces and tip velocities.

4. RESULTS AND DISCUSSION

MD simulation results of AFM nanoindentation and nanoscratching are presented in this section. All MD simulation snapshots are visualized by Atomeye [67]. The different colors shown in the following figures represent coordination number, which is a measure of how many nearest neighbors exist for a particular atom. For example, atoms in perfect fcc crystals have 12 nearest neighbors and their atomic coordination number is accordingly 12. Atoms with coordination numbers that are not 12 usually represent the location of defects and vacancies. The purpose of using this coordination number coloring is to clearly see the defects and dislocations of atoms.

In nanoindentation process, the indenter tip moves vertically into the surface of substrate. The atoms in the substrate are compressed beneath the tip and the deformation can be seen in the vicinity of the tip. The material apart from the tip seems to effect very little by the motion of the tip. MD simulation snapshots of nanoindentation are shown in Figure 2. The figure shows the initial stage of indenter tip and workpiece material in nanoindentation followed by the movement of the tip into the workpiece material at various time intervals. At the surrounding of contact surface between the indenter tip and the workpiece, a material pile-up is observed. Figure 3 shows top and cross-sectional views of MD simulation snapshots of nanoindentation. The tip is located at the maximum indentation depth at the time of 55 ps (Figure 3 (a)), while the tip is moved to its

initial point at the time of 115 ps (Figure 3(b)). The elastic deformation on the top surface of the gold workpiece undergoes elastic recovery after the tool tip was moved upward from the workpiece. It can be seen from Figure 3 that some deformation on the surface disappeared after the tool tip was moved up. Moreover, the depth of indentation mark and subsurface deformation decrease after the tip was removed from the workpiece.

Figure 4 shows the MD simulation snapshots of nano-scratching with the scratching depth of 5 nm. The crystal orientation of workpiece material is Au(100) and the direction of scratching is [100]. The scratching length is 30 nm and the tip speed is 10 m/s. The atoms in the workpiece are compressed beneath and in front of the tip and assembled to form a small chip. The material pile-up can be seen along the resulting groove. Several types of defects, including vacancies and Shockley partial dislocation loops, can be observed during the simulation. The dislocation loops are highly mobile and participate in various interactions among themselves and with other defects. The dislocation loops on the top surface are emitted in front of the tip and generally move out of the computation domain at a side boundary and come inside from the opposite side of boundary, due to the periodic boundary conditions applied to all four side boundaries.

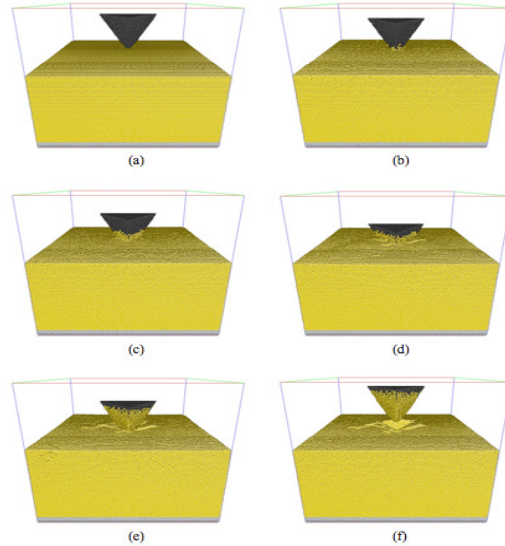


Figure 2. MD simulation snapshots of nano-indentation at various times: (a) 0 ps; (b) 10 ps; (c) 30 ps; (d) 55 ps; (e) 80 ps; (f) 115 ps

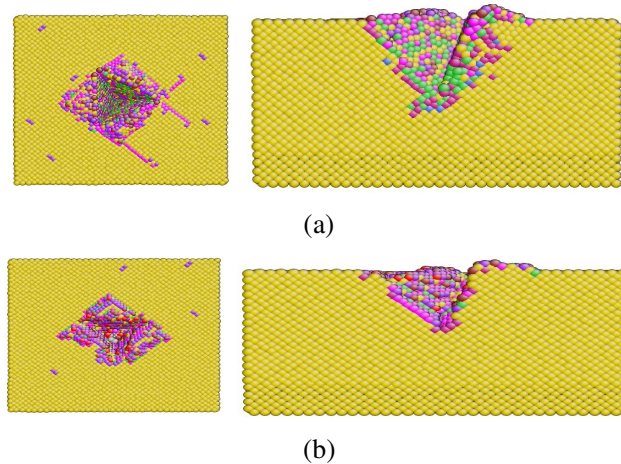


Figure 3. Top (left) and cross-sectional (right) views of MD simulation snapshots of nano-indentation: (a) time = 55 ps; (b) time = 115 ps

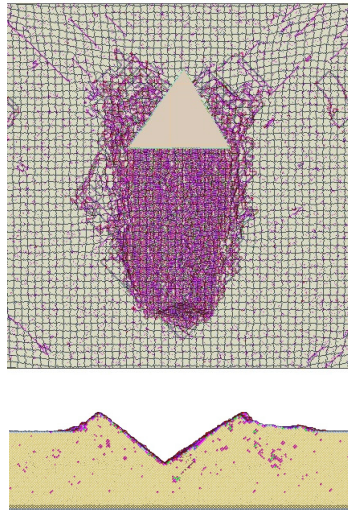


Figure 4. MD simulation snapshots of nano-scratching with a scratching depth of 5 nm.

The effect of scratching depth on material deformation was investigated. MD simulations of nanoscratching were conducted with scratch depths varying from 1 to 7 nm. Top and cross-sectional views of MD simulation snapshots of nanoscratching at various scratching depths are shown in Figs. 5 - 6, respectively. As the scratching depth increases, the deformation is found to penetrate much deeper from the surface and the height of material pile-up also increases. In addition, more dislocation loops on the top surface can be observed. With increasing depth, the dislocations reach side boundaries sooner and re-enter from the opposite side. Some of these partial dislocations interact with other defects to form more defects on the top surface. This indicates that a larger computation domain is needed.

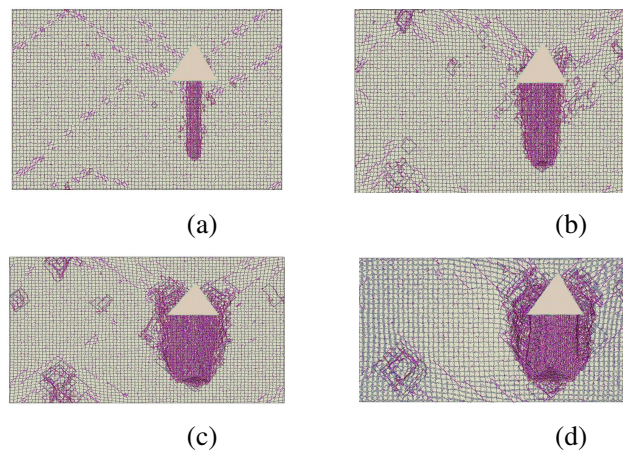


Figure 5. Top view of MD simulation snapshots of nanoscratching with different depths of scratch: (a) 1 nm; (b) 3 nm; (c) 5 nm; (d) 7 nm.

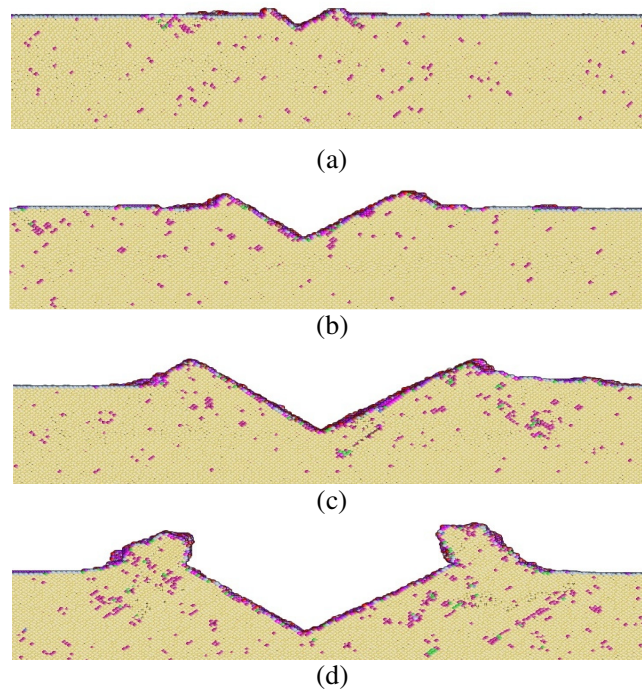


Figure 6. Cross-sectional views of MD simulation snapshots of nanoscratching with different depths of scratch: (a) 1 nm; (b) 3 nm; (c) 5 nm; (d) 7 nm.

The effects of crystal orientation are presented. Here, the MD simulations of nanoscratching were conducted on three different crystal orientations: Au(100), Au(110), and Au(111). The scratching depths are 5 nm for all the three cases. Figs 7 – 9 show cross-sectional (a, b) and top (c, d) views of different crystal orientations. Different pattern of surface and subsurface deformation can be observed for different crystal orientations.

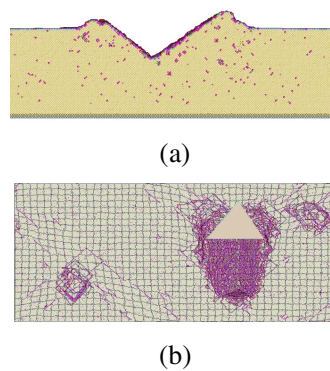


Figure 7. Cross-sectional (a) and top (b) views of MD simulation snapshots of nanoscratching of Au(100)

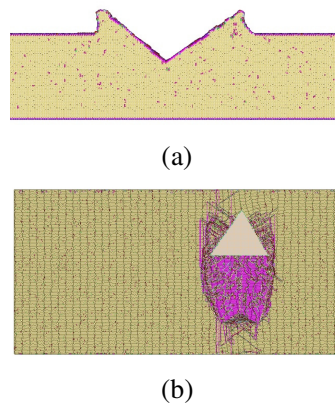


Figure 8. Cross-sectional (a) and top (b) views of MD simulation snapshots of nanoscratching of Au(110)

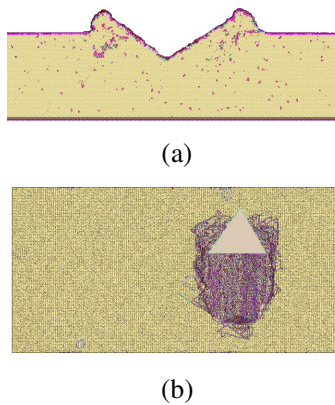


Figure 9. Cross-sectional (a) and top (b) views of MD simulation snapshots of nanoscratching of Au(111)

Fig. 10 shows AFM image and cross-sectional profile of nanoindentation for different applied forces, increasing from right to left: 65, 70, 75, 80, 85 μN . The AFM experiments were repeated for five times (five rows shown in Fig. 10). The indentation depths increase as the applied forces increase. The variation of indentation forces with depths of indentation at different tool tip speeds is shown in Figure 11. It can be observed that the indentation force increases as the depth of indentation increases. The simulation results were compared with the experimental results. Due to the limitation on computational time, it should be noted that the tool tip speeds used in this MD simulation are a lot higher than those used in the experiment. The typical speed used in the experiment is approximately 5-10 $\mu\text{m/s}$. Therefore, the effect of tool tip speed on the indentation force is also investigated in this paper. It can be observed from Figure 11 that the indentation force increases as the tool tip speed decreases. Since the tip speed plays an important role on the indentation force, the quantitative values of the indentation force obtained from MD simulation are not comparable to the experimental results. However, the increasing trends of indentation force are the same for both simulation and experimental results. AFM experiments of nanoindentation were also carried out to investigate the effect of the tip speed. Fig. 12 shows the experimental results of AFM nanoindentation for different applied forces and different tip speeds. The tip speeds were increased from 1 to 10 $\mu\text{m/s}$ from top to bottom rows. Fig. 12 (b) shows the cross-sectional profile of AFM nanoindentation. The blue line represents the cross-sectional profile for the tip speed of 1 $\mu\text{m/s}$. The red line represents the cross-sectional profile for the tip speed of 10 $\mu\text{m/s}$. It can be seen from Fig. 12 (b) that the indentation depth increases as the tip speed increases.

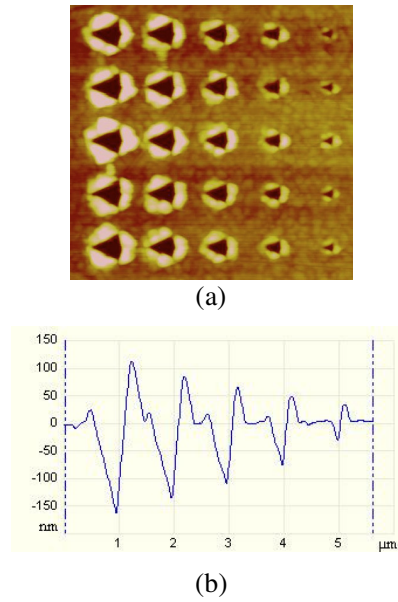


Figure 10. (a) AFM image of nanoindentation for different applied forces (increasing from right to left); (b) cross-sectional profile of AFM image

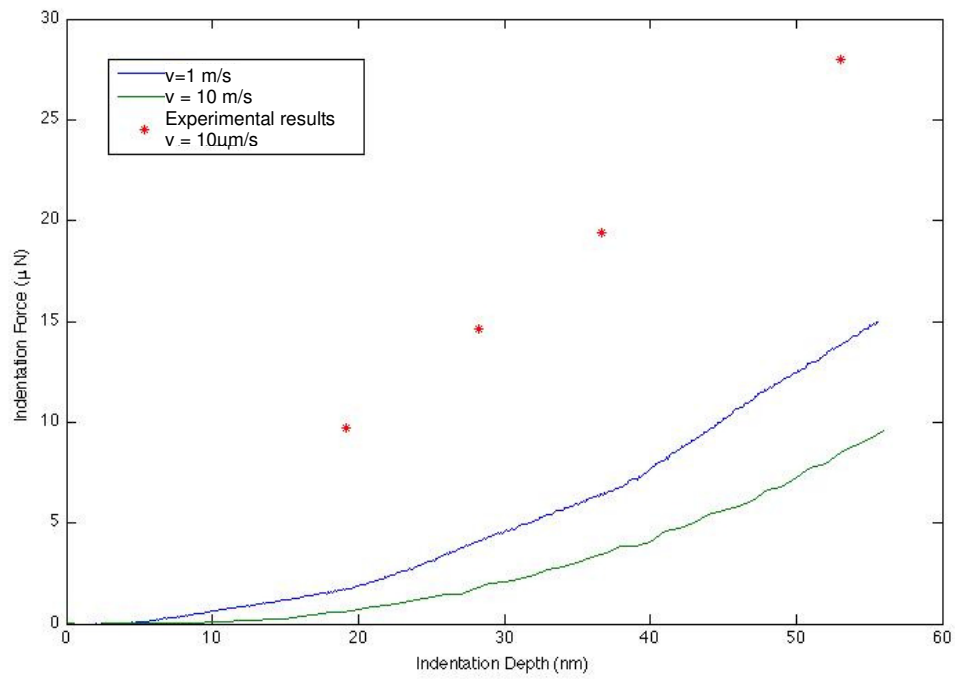


Figure 11. Variation in indentation forces with depths of indentation.

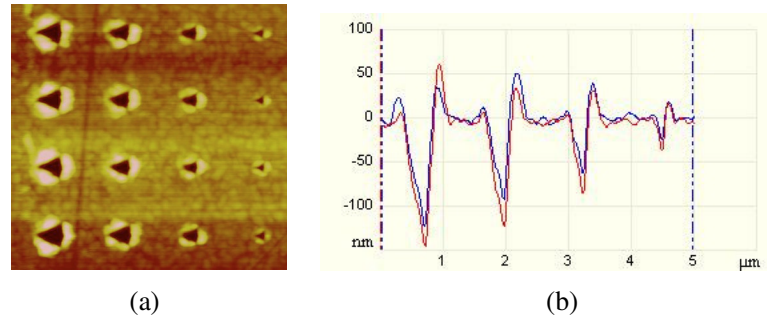


Figure 12. (a) AFM image of nanoindentation for different applied forces (increasing from right to left) and tip speeds (increasing from top to bottom); (b) cross-sectional profile of AFM image

Figure 13 shows the AFM experimental results of nanoscratching with five different depths, namely 20, 30, 40, 50, 60 nm (increasing from right to left). As can be seen from figure 13(b), the surface roughness varies between 0 to 20 nm across the gold substrate. In order to obtain a meaningful result, the depths of scratch used in the experiments must be higher than 20 nm which are ten times the depths used in the MD simulation. For this reason, the quantitative values, i.e. forces, obtained from MD simulation are not comparable to the experimental results; only the qualitative values are discussed here. The height of the material pile-up along the scratch groove is found to increase as the depth of cut increases in both MD simulation and AFM experiments. However, the material pile-up on the left side is observed to be higher than the right side which is different from the MD simulation results. One possible explanation for the discrepancy may be that the x-rotation of the AFM probe used in the experiment was set to 12 degree as recommended by manufacturer. During the nanoscratching process, the pile-up material in front of the tip increases, but not enough to form chip. Thus, no chip formation is detected in both MD simulation and experiments for the observed depths of scratch.

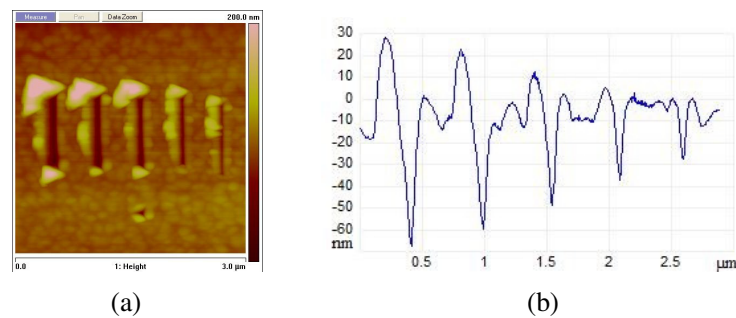


Figure 13. AFM experimental results of nano-scratching with five scratch depths of 20, 30, 40, 50, 60 nm increasing from right to left: (a) AFM image of nanoscratching (b) cross-sectional profile

Figure 14 shows load-displacement curve for the case of gold material and diamond indenter tip. As the indentation depth of the diamond tip continues to increase, the load curve continues to go up and until it reaches a maximum depth. After reaching the specified maximum depth, the tip begins to unload and return to its original position. The slope of the unloading curve at the maximum load is determined and used in the calculation of hardness and Young's modulus in Eq. (8) and (10), respectively. Two different methods were used to calculate the contact area. In the first method, the contact area was calculated from the geometry of the indenter as a function of the contact depth, h_c . In the second method, the contact area was estimated from the location of displaced atoms at the interface between tip and sample. The material properties of diamond indenter used in the calculation are $E_i = 1140$ GPa and $\nu_i = 0.07$. Table 3 shows the values of Young's modulus and hardness of gold obtained from the calculations.

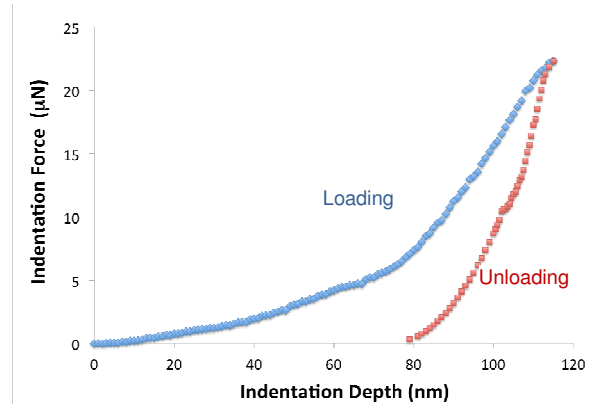


Figure 14. Load-displacement curves for the case of gold material and diamond indenter tip.

Table 3. Young's modulus and hardness of materials at nanoscale

Young's modulus (GPa)	Hardness (GPa) Method 1	Hardness (GPa) Method 2
235	113.56	98.13

Comparing to the macroscale properties, the Young's modulus of gold is approximately 57-120 GPa; while it was found to be 235 GPa from our calculation. The values of hardness obtained from method 1 in our calculation is slightly larger than those obtained from method 2. Both Young's modulus and hardness in our analysis were about two to three times larger than those at the macroscale. This discrepancy is a result of the scale differences. Bulk material typically has constant material properties regardless of its size, but size-dependent properties are often observed at the nanoscale. Nanoscale material has a high surface area and a large fraction of the atoms are on its surface. This can give rise to size effects in material properties at the nanoscale. Moreover, the defect of the material such as grain boundaries and dislocations was different at different scales. In addition, an assumption of perfect defect-free single-crystal material was applied in the MD simulation; while, in general, materials at macroscale were poly-crystalline and contained several types of defects.

3. CONCLUSIONS

MD simulations of AFM-based nanoindentation and nanoscratching were conducted to investigate the effect of indentation and scratching depth, crystal orientation and tip speed. Material properties at the nanoscale were also extracted and compared with macroscopic properties. Several types of defects, including vacancies and Shockley partial dislocation loops, could be observed during the simulation. With increasing depth of scratch, the dislocations reach side boundaries sooner and re-enter from the opposite side. Some of these partial dislocations interact with other defects to form more defects on the top surface. Due to the periodic boundary condition applied to all the four side boundaries, the simulation domain should be large enough to avoid the re-entering of dislocations. For different crystal orientations, different pattern of surface

and subsurface deformation can be observed. The effect of indentation depths and tip speeds was investigated and found that indentation force increases as depth of indentation and tip speed increase. Material properties, e.g. Young's modulus and hardness, of the materials at the nanoscale are different from those at the macroscale. Hence, due to different material properties between nano- and macro-scale, materials at nanoscale are typically considered new types of material. As can be seen from the presented results, these machining parameters affected the final nano-geometry in AFM-based nanomachining. The findings from this work can be applied to the fabrication of nanochannels/nano-fluidic devices. However, a more extensive experimental study is necessary to better validate the computational models. This will be reported in our future work.

ACKNOWLEDGEMENTS

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute, and in part by the Indiana METACyt Initiative. The Indiana METACyt Initiative at IU is also supported in part by Lilly Endowment, Inc.

REFERENCES

- [1] A.P. Malshe, K.P. Rajurkar, K.R. Virwani, C.R. Taylor, D.L. Bourell, G. Levy, M.M. Sundaram, J.A. McGeough, V. Kalyanasundaram and A.N. Samant, "Tip-based nanomanufacturing by electrical, chemical, mechanical and thermal processes," *CIRP Annals - Manufacturing Technology*, Vol. 59, pp. 628-651, 2010.
- [2] A. C. Fischer-Cripps, *Nanoindentation*, Springer, New York, 2002.
- [3] A. N. Shipway, E. Katz and I Willner, "Nanoparticle Arrays on Surfaces for Electronic, Optical, and sensor applications," *ChemPhysChem*, Vol. 1(1), pp. 18-52, 2000.
- [4] M. Liu, N. A. Amro, C. S. Chow and G-Y Liu, "Production of Nanostructures of DNA on Surfaces," *Nano Letters*, Vol. 2(8), pp. 863-867, 2002.
- [5] P. Vettiger, M. Despont, U. Drechsler, U. Dürig, W. Häberle, M. I. Lutwyche, H. Rothuizen, R. Stutz, R. Widmer and G. K. Binnig, "The 'millipede' – more than one thousand tips for future AFM data storage," *IBM Journal of Research and Development*, Vol. 44(3), pp. 323-340, 2000.
- [6] C. R. Taylor, E. A. Stach, G. Salamo and A. P. Malshe, "Nanoscale dislocation Patterning by Ultralow Load Indentation," *Applied Physics Letters*, Vol. 87(7), 073108, 2005.
- [7] G. F. Zheng, F. Patolsky, Y. Cui, W. U. Wang and C. M. Lieber, "Multiplexed electrical detection of cancer markers with nanowire sensor arrays," *Nature Biotechnology*, Vol. 23(10), pp. 1294-1301, 2005.
- [8] X. Li, H. Gao, C. J. Murphy and K. K. Caswell, "Nanoindentation of silver nanowires," *Nano Letters*, Vol. 3, pp.1495-1498, 2003.
- [9] X. Li, P. Nardi, C. W. Baek, J. M. Kim and Y. K. Kim, "Direct nanomechanical machining of gold nanowires using a nanoindenter and an atomic force microscope," *Journal of Micromechanics and Microengineering*, Vol. 15, pp. 551-556, 2005.
- [10] Y. D. Yan, T. Sun, X. S. Zhao, Z. J. Hu and S. Dong, "Fabrication of microstructures on the surface of a micro/hollow target ball by AFM," *Journal of Micromechanics and Microengineering*, Vol. 18, 035002, 2008.
- [11] Y. J. Chen, J. H. Hsu and H. N. Lin, "Fabrication of metal nanowires by atomic force microscopy nanoscratching and lift-off process," *Nanotechnology*, Vol. 16, pp. 1112-1115, 2005.
- [12] Y. T. Mao, K. C. Kuo, C. E. Tseng, J. Y. Huang, Y. C. Lai, J. Y. Yen, C. K. Lee and W. L. Chuang, "Research on three dimensional machining effects using atomic force microscope," *Review of Scientific Instruments*, Vol. 80, 065105, 2009.
- [13] T. Fang, C. Weng and J. Chang, "Machining characterization of nano-lithography process using atomic force microscopy," *Nanotechnology*, Vol. 11, pp. 181-187, 2000.
- [14] S. Diegoli, C. A. E. Hamlett, S. J. Leigh, P. M. Mendes and J. A. Preece, "Engineering nanostructures at surfaces using nanolithography," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, Vol. 221(4), pp. 589-629, 2007.
- [15] T. Sun, Y. D. Yan, J. F. Xia, S. Dong, Y. C. Liang and K. Cheng, "Research on micro machining using AFM diamond tip," *Key Engineering Materials*, Vols. 259-260, pp. 577-581, 2004.

- [16] J. C. Huang, C. L. Li, and J. W. Lee, "The Study of Nanoscratch and Nanomachining on Hard Multilayer Thin Films Using Atomic Force Microscope," *Scanning*, Vol. 34, pp. 51-59, 2012.
- [17] Y. Yan, T. Sun, Y. Liang and S. Dong, "Investigation on AFM-based micro/nano-CNC machining system," *International Journal of Machine Tools and Manufacture*, Vol. 47, pp. 1651-1659, 2007.
- [18] F. Zhang, H. Zhang, Y. Yan and J. Wang, "Research on nano-scale material removal process using atomic force microscopy," *Key Engineering Materials*, Vols. 359-360, pp. 269-273, 2008.
- [19] H. Zhang, J. Kuai and F. Zhang, "Minimum thickness of cut in nanomachining using atomic force microscopy," *2010 International Conference on E-Product E-Service and E-Entertainment (ICEEE)*, Henan, China, November 7-9, 2010.
- [20] A. A. Tseng, J. Shirakashi, S. Nishimura, K. Miyashita and Z. Li, "Nanomachining of permalloy for fabricating nanoscale ferromagnetic structures using atomic force microscopy," *Journal of Nanoscience and Nanotechnology*, Vol. 10, pp. 456-466, 2010.
- [21] Z. Q. Wang, N. D. Jiao, S. Tung and Z. L. Dong, "Atomic force microscopy-based repeated machining theory for nanochannels on silicon oxide surfaces," *Applied Surface Science*, Vol. 257, pp. 3627-3631, 2011.
- [22] W. C. D. Cheong and L. C. Zhang, "Molecular dynamics simulation of phase transformations in silicon monocrystals due to nano-indentation," *Nanotechnology*, Vol. 11, pp. 173-180, 2000.
- [23] R. Komanduri, N. Chandrasekaran and L.M. Raff, "MD simulation of indentation and scratching of single crystal aluminum," *Wear*, Vol. 240, pp. 113-143, 2000.
- [24] D. Christopher, R. Smith and A. Richter, "Atomistic modelling of nanoindentation in iron and silver," *Nanotechnology*, Vol. 12, pp. 372-383, 2001.
- [25] T. Fang, C. Weng and J. Chang, "Molecular dynamics analysis of temperature effects on nanoindentation measurement," *Material Science and Engineering*, Vol. A357, pp. 7-12, 2003.
- [26] X. M. Liu, Z. L. Liu and Y G Wei, "Nanoscale Friction Behavior of the Ni-film/substrate system under scratching using MD simulation," *Tribology Letters* Vol. 46, pp. 167-178, 2012.
- [27] A. Gannepalli and S. K. Mallapragada, "Molecular dynamics studies of plastic deformation during silicon nanoindentation," *Nanotechnology*, Vol. 12, pp. 250-257, 2001.
- [28] I. Salehinia, S.K. Lawrence and D.F. Bahr, "The effect of crystal orientation on the stochastic behavior of dislocation nucleation and multiplication during nanoindentation," *Acta Materialia*, Vol. 61, pp. 1421-1431, 2013.
- [29] C. F. Sanz-Navarro, S. D. Kenny and R. Smith, "Atomistic simulations of structural transformations of silicon surfaces under nanoindentation," *Nanotechnology*, Vol. 15, pp. 692-697, 2004.
- [30] T. Akabane, Y. Sasajima and J. Onuki, "Computer simulation of silicon nanoscratch test," *Materials Transactions*, Vol. 47, pp. 1090-1097, 2006.
- [31] H. Okabe, T. Tsumura, J. Shimizu, L. Zhou and H. Eda, "Experimental and Simulation Research on Influence of Temperature on Nano-Scratching Process of Silicon Wafer," *Key Engineering Materials*, Vol. 329, pp. 379-384, 2007.
- [32] T. Fang, W. Chang and C. Weng, "Nanoindentation and nanomachining characteristics of gold and platinum thin films," *Materials Science and Engineering A*, Vol. 430, pp. 332-340, 2006.
- [33] T. Fang and C. Weng, "Three-dimensional molecular dynamics analysis of processing using a pin tool on the atomic scale," *Nanotechnology*, Vol. 11, pp. 148-153, 2000.
- [34] T. Fang, C. Weng, and J. Chang, "Molecular dynamics simulation of nano-lithography process using atomic force microscopy," *Surface Science*, Vol. 501, pp. 138-147, 2002.
- [35] Y. Yan, T. Sun, S. Dong and Y. Liang, "Study on effects of the feed on AFM-based nano-scratching process using MD simulation," *Computational Materials Science*, Vol. 40, pp. 1-5, 2007.
- [36] H. Yu, J. B. Adams and L. G. Hector Jr, "Molecular dynamics simulation of high-speed nanoindentation," *Modeling and Simulation in Materials Science and Engineering*, Vol. 10, pp. 319-329, 2002.
- [37] S. Jun, Y. Lee, S. Kim and S. Im, "Large-scale molecular dynamics simulations of Al(111) nanoscratching," *Nanotechnology*, Vol. 15, pp. 1169-1174, 2004.
- [38] Y. Lee, J. Park, S. Kim, S. Jun and S. Im, "Atomistic simulations of incipient plasticity under Al(111) nanoindentation," *Mechanics of Materials*, Vol. 37, pp. 1035-1048, 2005.
- [39] D. Mulliah, D. Christopher, S. D. Kenny and R. Smith, "Nanoscratching of silver (100) with a diamond tip," *Nuclear Instruments and Methods in Physics Research B*, Vol. 202, pp. 294-299, 2003.
- [40] D. Mulliah, S. D. Kenny, R. Smith and C. F. Sanz-Navarro, "Molecular dynamics simulations of nanoscratching of silver (100)," *Nanotechnology*, Vol. 15, pp. 243-249, 2004.
- [41] R. Smith, D. Cristopher and S. D. Kenny, "Defect generation and pileup of atoms during nanoindentation of Fe single crystals," *Physical Review B*, Vol. 67, 245405, 2003.

- [42] C. Lu, Y. Gao, G. Y. Deng, G. Michal, N. N. Huynh, X. H. Liu and A. K. Tieu, "Atomic-scale anisotropy of nanoscratch behavior of single crystal iron," *Wear*, Vol. 267, pp. 1961-1966, 2009.
- [43] Z. Lin, J. Huang and Y. Jeng, "3D nano-scale cutting model for nickel material," *Journal of Materials Processing Technology*, Vol. 192-193, pp. 27-36, 2007.
- [44] Y. Gao, C. Lu, N. N. Huynh, G. Michal, H. T. Zhu and A. K. Tieu, "Molecular dynamics simulation of effect of indenter shape on nanoscratch of Ni," *Wear*, Vol. 267, pp. 1998-2002, 2009.
- [45] I. Gheewala, R. Smith and S. D. Kenny, "Nanoindentation and nanoscratching of rutile and anatase TiO₂ studied using molecular dynamics simulations," *Journal of Physics: Condensed Matter*, Vol. 20, 354010, 2008.
- [46] Y. Liang, J. Chen, M. Chen, D. Song and Q. Bai, "Three-dimensional molecular dynamics simulation of nanostructure for reciprocating nanomachining process," *Journal of Vacuum Science and Technology B*, Vol. 27, pp. 1536-1542, 2009.
- [47] J. Chen, Y. Liang, Q. Bai, Y. Tang and M. Chen, "Mechanism of material removal and the generation of defects by MD analysis in three-dimensional simulation in abrasive processes," *Key Engineering Materials*, Vol. 359-360, pp. 6-10, 2008.
- [48] Z. Lin and J. Huang, "A study of the estimation method of the cutting force for a conical tool under nanoscale depth of cut by molecular dynamics," *Nanotechnology*, Vol. 19, 11570, 2008.
- [49] D. Mulliah, S. D. Kenny, E. McGee, R. Smith, A. Richter and B. Wolf, "Atomistic modeling of ploughing friction in silver, iron and silicon," *Nanotechnology*, Vol. 17, pp. 1807-1818, 2006.
- [50] R. Promyoo, H. El-Mounayri, and A. Martini, "AFM-Based Nanomachining for Nano-fabrication Processes: MD Simulation and AFM Experimental Verification," *ASME International Manufacturing Science & Engineering Conference*, Erie, PA, October 2010.
- [51] R. Promyoo, H. El-Mounayri, K. Varahramyan, and A. Martini, "Molecular Dynamics Simulation of AFM-Based Nanomachining Processes," *ASME International Manufacturing Science & Engineering Conference*, Corvallis, OR, June 13 – 17, 2011.
- [52] R. Promyoo, H. El-Mounayri, and K. Varahramyan, "AFM-Based Manufacturing for Nano-fabrication Processes," *TSME International Conference on Mechanical Engineering*, Krabi, Thailand, October 2011.
- [53] R. Promyoo, H. El-Mounayri, and K. Varahramyan, "AFM-Based Nanoindentation Process: A Comparative Study," *ASME International Manufacturing Science & Engineering Conference*, Norte Dame, IN, June 4-8, 2012.
- [54] R. Promyoo, H. El-Mounayri, K. Varahramyan, and V. Kumar, "AFM-Based Nanofabrication: Modeling, Simulation, and Experimental Verification," *ASME International Manufacturing Science & Engineering Conference*, Madison, WI, June 10 – 14, 2013.
- [55] R. Promyoo, H. El-Mounayri, and K. Varahramyan, "AFM-based nanoindentation using a 3D molecular dynamics simulation model," *Journal of Materials Science and Engineering A*, Vol. 3(6), pp. 369-381, 2013.
- [56] B. J. Alder and T. E. Wainwright, 1959, "Studies in molecular dynamics. I. General method," *Journal of Chemical Physics*, Vol. 31, pp. 459-466, 1959.
- [57] B. J. Alder and T. E. Wainwright, "Studies in molecular dynamics. II. Behavior of a small number of elastic spheres," *Journal of Chemical Physics*, Vol. 33, pp. 1439-1451, 1960.
- [58] R. Komanduri and L. M. Raff, "A review on the molecular dynamics simulation of machining at the atomic scale," *Proceedings Institution of Mechanical Engineers*, Vol. 215 (B), pp. 1639-1672, 2001.
- [59] S. J. Plimpton, S. J., "Fast parallel algorithms for short-range molecular dynamics," *Journal of Computational Physics*, Vol. 117, pp. 1-19, 1995.
- [60] S. J. Plimpton, R. Pollock and M. Stevens, "Particle-mesh Ewald and rRESPA for parallel molecular dynamics simulations," *Proc of the Eighth SIAM Conference on Parallel Processing for Scientific Computing*, Minneapolis, MN.
- [61] P. Walsh, R. K. Kalia, A. Nakano and P. Vashishta, "Amorphization and anisotropic fracture dynamics during nanoindentation of silicon nitride: A multimillion atom molecular dynamics study," *Applied Physics Letters*, Vol.77, pp.4332-4334, 2000.
- [62] M. S. Daw and M. I. Baskes, "Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals," *Physical Review B*, Vol. 29, pp. 6443-6453, 1984.
- [63] P. M. Morse, "Diatomic molecules according to the wave mechanics II vibrational levels," *Physical Review*, Vol. 34, pp. 57-64, 1929.
- [64] I. M. Torrens, *Interatomic potentials*, Academic, New York, 1972.

- [65] W.C. Oliver, G.M. Pharr, "Measurement of hardness and elastic modulus by instrumented indentation: Advances in understanding and refinements to methodology," *Journal of Material Research*, Vol. 19, pp. 3-20, 2004.
- [66] <http://rt.uits.iu.edu/bigred2/index.php>
- [67] J. Li, "AtomEye: an efficient atomistic configuration viewer," *Modeling and Simulation in Materials Science and Engineering*, Vol. 11, pp. 173-177, 2003.

FACTORS INFLUENCING ACTUAL USE OF MOBILE LEARNING CONNECTED WITH E-LEARNING

Young Ju Joo¹ Sunyoung Joung² Eui Kyoung Shin³ Eugene Lim⁴ and Miran Choi⁵

^{1,3,4,5} Department of Education, Ewha Womans University, Seoul, South Korea

youngju@ewha.ac.kr

luvsiiin@naver.com

lim_u@naver.com

bohemiran@naver.com

² Department of Education, Kookmin University, Seoul, South Korea

sjoung@kookmin.ac.kr

ABSTRACT

The purpose of the current study is to propose discriminated management strategies for mobile learning environments after observing the effects of mobile self-efficacy on performance expectancy and effort expectancy, the social influence on intention of use, and the effects of facilitating conditions and intention of use on learners' actual use of mobile learning by adding mobile self-efficacy to the UTAUT model proposed by Venkatesh et al. (2003). We established hypotheses to determine whether mobile self-efficacy, performance expectancy, effort expectancy, social influence, and facilitating conditions affect intention of use and whether intention of use affects actual use. Results showed that when mobile self-efficacy and performance expectancy is higher, so is the intention of using mobile learning services. It was confirmed that the factors had significant indirect effects on the actual use by mediating the intention of use and that the intention of use directly affected actual use. However, the current research reported that effort expectancy, social influence, and facilitation conditions did not have significant effects on the intention of using mobile learning services. These results will contribute substantially to the design of effective mobile learning environments.

KEYWORDS

Structural relationship, Technology Acceptance Model, Mobile learning

1. INTRODUCTION

Despite the tendency of cyber-universities to offer many services, e-learners are likely to limit their mobile learning to services within the administrative context, such as announcements. Actual use of mobile services connected with e-learning is insufficient (Lee, 2010). Learners in traditional as well as cyber-universities identify the complexity of log-in and authentication processes, speed problems due to simultaneous access, and limitations from unstable functions as factors inhibiting the use of mobile learning services (Min, Sin, Ryu, & Gwak, 2014).

It was discovered that the current level of mobile learning practice consists simply of converting e-learning contents to mobile learning contents. Learners satisfied with using e-learning do not

dare to switch to mobile learning (Choi & Rho, 2014). Accordingly, it is necessary to devise strategies to increase mobile learning service adoption by analyzing the factors that increase practical use of mobile learning to expand the usage of mobile learning.

Although several theories have been devised to predict the adoption and diffusion of new technology, such as the Theory of Reasoned Action (TRA), the Technology Acceptance Model (TAM), it is necessary to develop an integrative theory and model for the adoption and diffusion of new technologies because of individual differences in technology adoption and the heterogeneity of research environments (Venkatesh, Morris, Davis, & Davis, 2003). Venkatesh and his colleagues (2003) proposed a Unified Theory of Acceptance and Use of Technology (UTAUT) Model that integrates the eight current models after reexamining the validity. The UTAUT is used as the latest model for analyzing intention of using new technology and actual use. The eight models are the TRA, TAM, Theory of Planned Behavior (TPB), Technology Acceptance Model-Theory of Planned Behavior (TAM-TPB), Integrated Model of Technology Acceptance and Use, Motivation Model, PC Utilization Model, Diffusion of Innovation Theory, and Social Cognition Theory (Venkatesh et al., 2003).

The UTAUT model is well known to affect performance expectancy, effort expectancy, social influence, and facilitating conditions. Among them, performance expectancy and effort expectancy directly affect social influences, and facilitating conditions and intention of use affect actual use (Venkatesh et al., 2003). Meanwhile, Venkatesh and his colleagues (2003) referred to self-efficacy as a variable directly affecting actual use in conjunction with facilitating conditions and intention of use. They excluded self-efficacy from the final model because of the post-research results, which proved the non-significant effects of self-efficacy on actual use (Venkatesh & Zhang, 2010).

However, recent empirical studies (Chiu & Wang, 2008; El-Gayar & Moran, 2006; Luarn & Lin, 2005) have reported self-efficacy as directly affecting actual use of new technologies and intention of using information systems. There are contradictory research results between self-efficacy and actual intention in the mobile learning environments. Therefore, we will examine the effects of mobile self-efficacy by adding it as an individual variable in the mobile learning environments connected with e-learning.

The purpose of current study, then, is to propose discriminated management strategies for mobile learning environments after observing the effects of mobile self-efficacy on performance expectancy and effort expectancy, the social influence on intention of use, and the effects of facilitating conditions and intention of use on learners' actual use of mobile learning by adding mobile self-efficacy to the UTAUT model proposed by Venkatesh et al. (2003)

We established hypotheses to determine whether mobile self-efficacy, performance expectancy, effort expectancy, social influence, and facilitating conditions affect intention of use and whether intention of use affects actual use. The research hypotheses are as follows, and the hypothetical research model is displayed in Figure 1.

[Hypothesis 1] Mobile self-efficacy, performance expectancy, effort expectancy, and social influence will affect intention of mobile learning.

[Hypothesis 2] Intention of mobile learning will affect the actual uses of mobile learning.

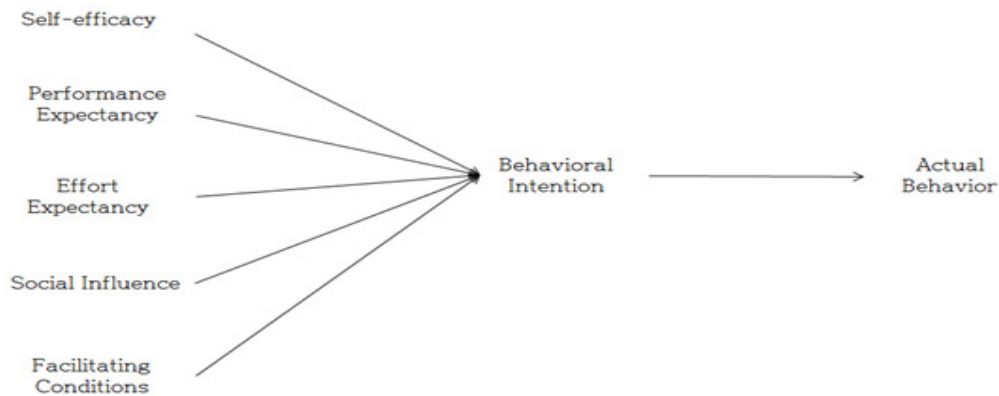


Figure 1. Hypothetical Research Model

* Self-efficacy = Mobile self-efficacy

2. RESEARCH METHOD

2.1. Subjects and Research Procedure

Research subjects were undergraduate students registered for the required course majoring in health wellness at a W cyber university. The W Cyber University provides Wellness Health Programs specialized in Korean culture and practical welfare through mobiles services by both IOS and Android platforms. Subjects were a total of 238 people. Male subjects were 80(33.6%) and female ones were 158(66.4%). Subjects in their 20s were eight (3.4%), 30s were 32(13.4%), 40s were 105(44.1%), 50s were 84(35.7%), and over 60s were 8(3.4%). Mostly were in their 40s and 50s.

2.2. Measurement Instrument

The instrument used in this study contains 28 self-report items including mobile self-efficacy, performance expectancy, effort expectancy, social influence, facilitating conditions, intention of use, and actual use of mobile services. Each item was revised appropriately to meet the current research purpose after content validity tests were conducted by two experts. All variables except actual use were measured on a five-point Likert scale; actual use was measured using actual access time. Each variable was assessed using the measures described below:

To measure mobile self-efficacy, we used the instrument developed by Wang and Wang (2008), which consists of ten questions asking about self-ability or self-belief related to performing mobile operations. For reliability, the measurement has a Cronbach's α of .92. To measure performance expectancy, effort expectancy, social influence, facilitating conditions, and intention of use, we used the instrument by Venkatesh et al. (2003), which consists of three items asking about expecting usefulness, productivity improvement, expenditure reduction, and performance improvement. For reliability, the measurement has a Cronbach's α of .83. Effort expectancy relates to e-learners' perceived ease of use in mobile learning. It was measured through four items with Cronbach's α of .87. Social influence was measured through four items with Cronbach's α of .81. Facilitating conditions is the perceived possibility of receiving help while using mobile learning services. It is measured through four items with Cronbach's α of .80. Intention of use is the intention of continuously using the mobile learning services. It is measured through three items with Cronbach's α of .96. Finally, actual use is the degree of actual usage of mobile services. It was measured using actual access time to mobile learning services during a semester.

2.3. Research Analysis Methods

We conducted reliability tests, factor analysis, descriptive statistics, and correlation analysis with the collected data using SPSS. We used structural equation modeling (SEM) to analyze the structural relationships between relevant variables and actual use of mobile learning. To analyze the SEM, we first evaluated the validity of the measurement model by using the AMOS program. After that, we used two approaches in conducting SEM. We used maximum likelihood estimation (MLE) to verify the data normality and confirmed model fit using a χ^2 test, TLI, CFI, and RMSEA, which are widely used measures of model fit.

3. RESEARCH RESULTS

3.1. Examination of Measurement Model

This study examined the validity of the structural model prior to analyzing it. The results are displayed in Table 1. Observing Table 1, although the χ^2 results of the measurement model were significant (39, $n = 238$, $CMIN = 71.980$) at p level of .001, it is desirable to assess model fit by using other indices since χ^2 is sensitive to sample size and data normality. TLI (.978) and CFI (.987) were both over .90, satisfying the acceptable criteria, and RMSEA (.060) was also acceptable. Therefore, it was confirmed that the measurement model is suitable.

Table 1. Fit of the Measurement Model ($n = 238$)

	CMIN	P	df	TLI	CFI	RMSEA (90% Confidence Interval)
Measurement Model	71.980	.001	39	.978	.987	.060(.037~.081)
Criteria Value				**>.900*	>.900	<.080

The results of confirmatory factor analysis provided robust evidence of construct and distinctive validity. Standard factor loadings in each path of the measurement variable ranged from .78 to .98, both over .50. This means that the selected index variables, which are selected to measure each theoretical variable in each research model, indicate adequate construct validity (Bagozzi & Yi, 1988). The results of examining the correlations between latent variables ranged from .47 to .78, showing low cross-correlations (less than .80). This shows that there is sufficient distinctive validity between the latent variables (Bagozzi et al., 1988).

3.2. Structural Model Examination

First, as shown in the Table 2, the examination result of χ^2 ($[50, n = 238] = 103.581, p = .000$) was significant, and we can consider it an appropriate model with evidence of TLI = .961, CFI = .979, and RMSEA = .067. This means that there is a causal relationship between mobile self-efficacy, performance expectancy, performance expectancy, social influence, facilitating conditions, intention of use, and actual time for use.

Table 2. Fit of the Initial Structural Model ($n = 238$)

	CMIN	P	df	TLI	CFI	RMSEA (90% Confidence Interval)
Initial Structural Model	103.581	.000	50	.961	.979	.067 (.049~.086)
Criteria Value				>.900	>.900	<.080

We examined the direct effects between variables included in the structural model and estimated the path coefficients under the verification that the structural model has a good fit. As shown in Table 3, three paths among six direct paths are statistically significant.

First, the effect of mobile self-efficacy on intention of use was significant ($\beta = .371, t = 4.258, p < .05$). Second, the effect of performance expectancy on intention of use was significant ($\beta = .734, t = 5.852, p < .05$). Third, the effect of effort expectancy on intention of use was not significant ($\beta = -.122, t = -1.039, p > .05$). Fourth, the effect of social influence on intention of use was not significant ($\beta = -.057, t = -.668, p > .05$). Fifth, the effect of facilitating conditions on actual use was not significant ($\beta = -.104, t = -1.216, p > .05$). Finally, the effect of intention of use on actual use was significant ($\beta = .374, t = 4.177, p < .05$).

After examining statistical significance in the initial structural model, the direct effects of performance expectancy, effort expectancy, social influence, and facilitation conditions were found to be not statistically significant. We established a succinct revised structural model under the supposition that there is no significant difference in model fit after removing insignificant paths from the initial structural model. We conducted a χ^2 test to confirm if there is a statistical difference between the initial structural model and a revised succinct model since there is a hierarchical relationship between the initial structural and revised succinct models. From the analysis results, we selected the revised succinct research model as a final research model since there was no significant difference between the initial structural model and the revised succinct model ($\Delta\chi^2 = 5.933, p = 115$).

Table 3. Examination of Fit of Revised Structural Model ($n = 238$)

	CMIN	P	df	TLI	CFI	RMSEA (90% Confidence Interval)
Revised Structural Model	109.515	.000	53	.961	.978	.067 (.049~.086)
Initial Structural Model	103.581	.000	50	.961	.979	.067 (.049~.086)
Criteria Value				>.900	>.900	<.080

The results of MLE estimation to measure the fit of the revised structural model are shown in Table 3. The revised structural model had good fit, with TLI = .961, CFI = .978, and RMSEA = .067. Accordingly, the results of investigating the effects of mobile self-efficacy, performance expectancy, facilitating conditions, intention of use, and actual use are shown in Table 4.

The effect of mobile self-efficacy on intention of use was $\beta = .269$ ($t = 3.844$, $p < .05$), that of performance expectancy on intention of use was $\beta = .596$ ($t = 7.697$, $p < .05$), and that of intention of use on actual use was $\beta = .373$ ($t = 4.162$, $p < .05$).

The research results reported that mobile self-efficacy and performance expectancy affects intention of use, and intention of use affects actual use. The final model including standardized path coefficients is shown in Figure 2.

Table 4. Path Coefficients of Revised Structural Model ($n = 238$)

			Unstandardized Coefficients (B)	Standardized Coefficients (β)	S.E	t	p
Self-Efficacy	→	Intention of Use	00.308	.269*	00.080	3.844	*
Performance Expectancy	→	Intention of Use	000.725	.596*	00.094	7.697	*
Intention of Use	→	Actual Use	271.627	.373*	65.267	4.162	*

* $p < .05$

*Self-Efficacy = Mobile Self-Efficacy



Figure 2. Standardized Path Coefficients of Revised Model

The results show that mobile self-efficacy and performance expectancy affect intention of use, and intention of use affects actual use. Accordingly, we examined the significance of indirect effects between the variables using a Sobel test (Kline, 2011). Mobile self-efficacy ($z = 2.826$, $p = 0.005$) and performance expectancy ($z = 3.662$, $p = 0.000$) were found to have indirect effects on mobile learning service by mediating intention of use. These direct and indirect effects are analyzed in Table 5.

Table 5. Direct and Indirect Effects Analysis of Revised Structural Model ($n = 238$)

Relevant Variables		Unstandardized Estimate			Standardized Estimate		
Relevant Variables		Total	Direct	Indirect	Total	Direct	Indirect
Self-Efficacy	→ Intention of Use	.308	.308	-	.269	.269	-
Performance Expectancy	→ Intention of Use	.725	.725	-	.596	.596	-
Self-Efficacy	→ Actual Use	83.650	-	83.650	.222	-	.222
Performance Expectancy	→ Actual Use	196.840	-	196.840	.100	-	.100
Intention of Use	*→ Actual Use	271.627	271.627	-	.373	.373	-

* Self-Efficacy = Mobile Self-Efficacy

4. CONCLUSIONS AND SUGGESTIONS

According to the current research results, when mobile self-efficacy and performance expectancy are high, so is the intention of using mobile learning services. It was confirmed that the factors had significant indirect effects on actual use by mediating intention of use. It was also confirmed that the intention of use directly affected actual use. However, the current research reported that effort expectancy, social influence, and facilitation conditions did not have significant effects on the intention of using mobile learning services.

Firstly, mobile self-efficacy in the use of mobile learning service connected with e-learning. That is, cyber-learners' feelings of self-confidence and self-ability when using mobile machines significantly affected intention of using mobile learning services.

Secondly, it appeared that performance expectancy increased the intention of mobile learning, and mobile learning services connected with e-learning improved learning outcomes, reduced time and expenditure, and increased the efficiency and effectiveness of learning. The current results will contribute substantially to the design of effective mobile learning environments.

ACKNOWLEDGEMENTS

This work was supported by National Research Foundation of Korea Grant funded by the Korean Government (2012-045331)

REFERENCES

- [1] Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the academy of marketing science*, 16(1), 74-94.
- [2] Chiu, C. M., & Wang, E. T. (2008). Understanding Web-based learning continuance intention: The role of subjective task value. *Information & Management*, 45(3), 194-201.

- [3] Choi, M. N., Roh, H .L. (2014). A Study of the influence of motivations on the intention of taking mobile-learning courses in universities. *The Journal of Educational Information and Media*, 20(1), 77-95.
- [4] El-Gayar, O. F., & Moran, M. (2006). College students' acceptance of tablet PCs: an application of the UTAUT model. *Dakota State University*.
- [5] Lee, J. (2010). M-Learning as a Challenge for Cyber Universities: The Present and the Future. *Journal of Cyber society & Culture*, 1(2), 91-119.
- [6] Luarn, P., & Lin, H. H. (2005). Toward an understanding of the behavioral intention to use mobile banking. *Computers in Human Behavior*, 21(6), 873-891.
- [7] Min, K. B., Shin, M. H., Yu, T. H., Hwak, S. H. (2014). Strategies for Revitalizing E-Learning Through Investigating the Characteristics of E-Learning and the Needs of Distance Learners in the Domestic Universities in Korea. *The Korea Contents Society*, 14(1), 30-39.
- [8] Min, K. B., Shin, M., Shin, Yu, T., Kwak, S. (2014). Strategies for Revitalizing E-Learning through Investigating the Characteristics of E-Learning and the Needs of Distance Learners in the Domestic Universities in Korea. *The Korea Contents Society*, 14(1), 30-39.
- [9] Venkatesh, V., & Zhang, X. (2010). Unified Theory of Acceptance and Use of Technology: US Vs. China. *Journal of Global Information Technology Management*, 13(1).
- [10] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 27(3).
- [11] Wang, Y. S. & Wang, H. Y. (2008). Developing and Validating an Instrument for Measuring Mobile Computing Self-Efficacy. *Cyber Psychology & Behavior*, 11(4), 405-413.

DISTANCE'S QUANTIFICATION ALGORITHM IN AODV PROTOCOL

Meryem Saadoun¹, Abdelmajid Hajami² and Hakim Allali³

Department of Mathematics and Computer Sciences,
Hassan 1st University, Settat, Morocco

¹saadoun.meryem@gmail.com

²abdelmajidhajami@gmail.com

³hakim-allali@hotmail.fr

ABSTRACT

Mobility is one of the basic features that define an ad hoc network, an asset that leaves the field free for the nodes to move. The most important aspect of this kind of network turns into a great disadvantage when it comes to commercial applications, take as an example: the automotive networks that allow communication between a groups of vehicles. The ad hoc on-demand distance vector (AODV) routing protocol, designed for mobile ad hoc networks, has two main functions. First, it enables route establishment between a source and a destination node by initiating a route discovery process. Second, it maintains the active routes, which means finding alternative routes in a case of a link failure and deleting routes when they are no longer desired. In a highly mobile network those are demanding tasks to be performed efficiently and accurately. In this paper, we focused in the first point to enhance the local decision of each node in the network by the quantification of the mobility of their neighbours. Quantification is made around RSSI algorithm a well known distance estimation method.

KEYWORDS

Ad hoc, Mobility, RSSI, AODV, Localization, Distance, GPS-free.

1. INTRODUCTION

Mobile ad hoc network (MANET) is an appealing technology that has attracted lots of research efforts. Ad hoc networks are temporary networks with a dynamic topology which doesn't have any established infrastructure or centralized administration or standard support devices regularly available as conventional networks [1]. Mobile Ad Hoc Networks (MANETs) are a set of wireless mobile nodes that cooperatively form a network without infrastructure, those nodes can be computers or devices such as laptops, PDAs, mobile phones, pocket PC with wireless connectivity. The idea of forming a network without any existing infrastructure originates already from DARPA (Defense Advanced Research Projects Agency) packet radio network's days [2][3]. In general, an Ad hoc network is a network in which every node is potentially a router and every node is potentially mobile. The presence of wireless communication and mobility make an Ad hoc network unlike a traditional wired network and requires that the routing protocols used in an Ad hoc network be based on new and different principles. Routing protocols for traditional wired networks are designed to support tremendous numbers of nodes, but they assume that the relative position of the nodes will generally remain unchanged. In ad hoc, since the nodes are mobile, the network topology may change rapidly and unpredictably and the connectivity among the terminals may vary with time. However, since there is no fixed infrastructure in this network, each mobile node operates not only as a node but also as a router forwarding packets from one node to other mobile nodes in the network that are outside the range

of the sender. Routing, as an act of transporting information from a source to a destination through intermediate nodes, is a fundamental issue for networks. [4]

The problem that arises in the context of ad hoc networks is an adaptation of the method of transport used with the large number of existing units in an environment characterized by modest computing capabilities and backup and fast topology changes.

According to the way of the creation and maintenance of roads in the routing of data, routing protocols can be separated into three categories, proactive, reactive and hybrid protocols. The pro-active protocols establish routes in advance based on the periodic exchange of the routing tables, while the reactive protocols seek routes to the request. A third approach, which combines the strengths of proactive and reactive schemes, is also presented. This is called a hybrid protocol.

Ad-hoc On-Demand Distance Vector routing protocol (AODV) [5] is a reactive routing protocol, who was standardized by the working group MANET [6] with IETF (Internet Engineering Task force), by the (RFC 3561).

The protocol's algorithm creates routes between nodes only when the routes are requested by the source nodes, giving the network the flexibility to allow nodes to enter and leave the network at will. Routes remain active only as long as data packets are traveling along the paths from the source to the destination .When the source stops sending packets, the path will time out and close.

In this paper we propose a solution that enables each node in the network to determine the location of its neighbors in order to create a more stable and less mobile road. For that purpose, we locally quantify the neighbor's distances of a node as the metric of mobility using AODV protocol.

The remainder of this paper is organized as follows. Section 2, describes briefly the AODV protocol. In Section 3, a summary of related work is presented. we present in Section 4 how to quantify, evaluate, estimate mobility in ad hoc network. Section 5 shows the algorithm used the quantification of the distance in AODV protocol. Section 6 presents some simulations and results. Finally Section 7 concludes this paper.

2. AD HOC ON-DEMAND DISTANCE VECTOR

AODV is an on-demand protocol which is capable of providing unicast, multicast [7], broadcast communication and Quality of Service aspects (QoS) [8], [9]. It combines mechanisms of discovery and maintenance roads of DSR (RFC 4728) [10] involving the sequence number (for maintains the consistency of routing information) and the periodic updates of DSDV [11].

At the discovery of routes, AODV maintains on each node transit information on the route discovery, the AODV routing tables contain:

- The destination address
- The next node
- The distance in number of nodes to traverse
- The sequence number of destination
- The expiry date of the entry of the table time.

When a node receives a packet route discovery (RREQ), it also notes in its routing table information from the source node and the node that just sent him the package, so it will be able to retransmit the response packet (RREP). This means that the links are necessarily symmetrical.

The destination sequence number field of a route discovery request is null if the source has never been linked to the destination, else it uses the last known sequence number. It also indicates in this query its own sequence number. When an application sends a route discovery, the source

waits for a moment before rebroadcast its search query (RREQ) road, after a number of trials, it defines that the source is unreachable.

Maintained roads is done by periodically sends short message application called "HELLO" , if three consecutive messages are not received from a neighbor, the link in question is deemed to have failed . When a link between two nodes of a routing path becomes faulty, the nodes broadcast packets to indicate that the link is no longer valid. Once the source is prevented, it can restart a process of route discovery.

AODV maintains its routing tables according to their use, a neighbor is considered active as long as the node delivers packets for a given destination, beyond a certain time without transmission destination, the neighbor is considered inactive. An entered routing table is considered active if at least one of the active neighbors using the path between source and destination through active routing table entries is called the active path. If a link failure is detected, all entries of the routing tables participating in the active path are removed.

3. RELATED WORK

In [12], a geometric mobility metric has been proposed to quantify the relative motion of nodes. The mobility measure between any pair of nodes is defined as their absolute relative speed taken as an average over time. This metric has certain deficiencies: First, it assumes a GPS like scheme for calculation of relative speeds while in a MANET, we cannot assume the existence of GPS, so we have to resort to other techniques for measuring relative mobility. Secondly, it is an "aggregate" mobility metric and does not characterize the local movement of the neighboring nodes to another particular node.

The Reference Point Group Mobility Model (RPGM) proposed in [13] was useful for predictive group mobility management. In RPGM, each group has a logical "center" and the center's motion defines the entire group's motion behavior including location, velocity, acceleration etc.

In [14], They proposed a measure of the network mobility which is relative and depending on neighboring and link state changes. Each node estimates its relative mobility, based on changes of the links in its neighboring. This measure of mobility has no unit, it is independent of any existing mobility models and it is calculated at regular time intervals.

The degree mobility used in [15] was calculated from the change of its neighboring to each node in time. The node mobility degree, represents at a given time t for each node in the ad hoc network, the change variations undergone in its neighboring compared to the previous time $t - 1$. Thus, nodes that join or/and leave the neighboring of a given node will have surely an influence on the evaluation of its mobility.

However, the last two measures are not representative's values of a change node's motion with respect to another node.

We can see that none of the metrics described above are suitable for characterizing the relative mobility of nodes in a particular node's neighborhood in a MANET. Hence, we feel that there is a need to develop such a metric which can be used by any routing protocol.

4. LOCAL QUANTIFICATION OF NEIGHBOURING MOBILITY

In this section, we define how we estimate nodes mobility in ad hoc network. Mobility is quantified locally and independently of localization of a given node. We represent this local quantification node mobility, by calculating of the distance between a node and its neighbors.

The quantification of the distance can be done using 3 methods:

Calculate the exact distance: this is done by two ways:

1st way: The distance calculation using the GPS: This operation is done by using a terminal capable of being localized through a positioning system by satellites: GPS. The principle of localization by the GPS system is based on the use of satellite coordinates and the estimation of distances between the receiver satellites. Distances are obtained from the estimation of the TOA (Time Of Arrival) of the signals transmitted by the satellites [16].

2nd way: Distance calculation function in a simulation environment: Like NS2, OPNET, tor other simulator.

Calculate the distance using the RSSI (Received Signal Strength Indication): in case that the absolute positioning is not accessible, dedicated equipment not available or not possible, in theory, to determine the distance between a transmitter and a receiver we can use the RSSI. RSSI is a generic radio receiver technology metric, which is usually invisible to the user of the device containing the receiver, but is directly known to users of wireless networking of IEEE 802.11 protocol family.

The distance using RSSI can be calculated using the Friis transmission formula:

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2 L}$$

P_r : Receiving power.

P_t : Transmitting Power.

G_t : Gain of a transmitting antenna = ability to radiate in a particular direction in space.

G_r : Gain of a receiving antenna = ability to couple the energy radiated from a direction in space.

λ : is the wavelength.

L : is system loss factor which has nothing to do with the transmission

d : is the distance between the antennas.

Then, to calculate the distance between two nodes that are equipped by transmitting antennas, the formula is:

$$d = \sqrt{\frac{P_t G_t G_r \lambda^2}{(4\pi)^2 L P_r}}$$

Calculate distance using GPS-free [17]: In case that the GPS is not accessible, we can use a GPS-free to localize the neighbor of each node. This method uses a mobile reference to calculate the coordinates of all the nodes in the network. However we can conclude the distance between any nodes. In this part, we use the distance reception power to determinate the distance between the reference and the others nodes.

To implement this method in AODV protocol, we have to choose the reference:

Choice of the reference

a and **b** are two nodes that want to communicate in a MANET network.

We put **i** the center of the coordination system.

Let **N**: The set of nodes in the network.

P_i: The set of one hop neighbor of the node **i**.

d_{ij}: the distance between nodes **i** and **j**.

D_i: All distances **d_{ij}**

We choose **i** / $i \in Pa$, $i \neq b$ and $I \neq a$

We choose **p** and **q** / $p, q \in Pa$, $dpq \neq 0$,

$$p\hat{i}q \neq 180^\circ , p\hat{i}q \neq 0^\circ$$

Node **i** defines its system of local coordination.

We set the x-axis as the line (**ip**).

We set the y-axis as the line (**iq**).

Thus the system is defined:

The node **i** is the center of the system.

$i_x = 0$, $i_y = 0$
p is the node situated on the axis of abscissas

$$p_x = d_{ip} \quad , \quad p_y = 0$$

The node **q** is located on the ordinate axis

$$q_x = d_{iq} \cos \alpha \quad , \quad p_y = d_{iq} \sin \alpha$$

$$\text{with } \alpha = p\hat{i}q$$

Using the theorem of Al-Kashi:

$$d_{pq}^2 = d_{ip}^2 + d_{iq}^2 - 2 * d_{ip} * d_{iq} * \cos \alpha$$

α is calculated using this formula:

$$\alpha = \arccos \frac{dip^2 + diq^2 - dpq^2}{2 * dip * diq}$$

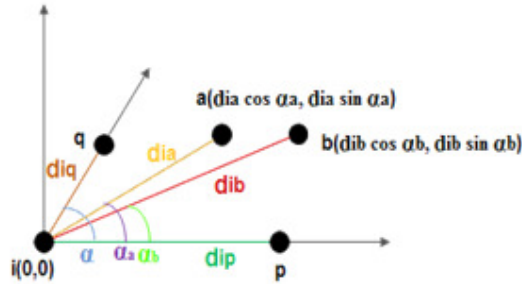


Figure 1 : The reference system

Once the reference is selected, the calculation of the coordinates of the nodes that belong to **Pi** is easy.

$$a \in Pi : \begin{cases} ax = dia \cos \alpha_a \\ ay = dia \sin \alpha_a \end{cases}$$

$$\alpha_a = a\hat{i}p = \arccos \frac{dip^2 + dia^2 - dpa^2}{2 * dip * dia}$$

For **b**, the coordinates are:

$$b \in Pa \begin{cases} bx = dib \cos \alpha_b \\ dib = dia - dab \\ by = dib \sin \alpha_b \end{cases}$$

$$\alpha_b = b\hat{i}p = \arccos \frac{dip^2 + dib^2 - dpb^2}{2 * dip * dib}$$

Then the final system is as follow:

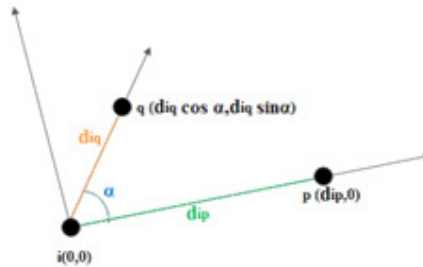


Figure 2: The system of localization local

Changing the Benchmark:

We propose in this part to replace the reference by another composed by the **a**'s neighbors having the smallest distances.

However, we choose $i, p, q / i, p, q \in Pa, i \neq b, i \neq a,$

$$p \hat{=} q \neq 180^\circ, p \hat{=} q \neq 0^\circ$$

And $\forall k \in Pi$ and $\forall k \neq a, b, i, p, q$

$$dak > dai, dak > dap \text{ and } dak > daq.$$

After the quantification of the distance between all nodes, we can describe the behavior of the node in the network by calculate the average of all the distances $Avg(d_{ij})$.

If the average is very high we say that the network nodes are very agitated else the network is supposedly more stable.

5. ALGORITHM OF QUANTIFICATION DISTANCE IN THE AODV ROUTING PROTOCOL

In this part, we propose to use one of those methods in the first function of a AODV protocol (rout establishment between a source and a destination).

A node x wants to communicate with a node y .

x diffuse **RREQ**.

Each node receiving **RREQ**, calculates the distance between itself and the neighbor who sent him **RREQ** (in this part we use the exact distance or the distance using the Pr) and broadcasts its table [**neighbors-distance**] to its neighbors.

To use the third method for the quantification of the distance, the algorithm has to change.

A node x wants to communicate with a node y .

x diffuse **RREQ**.

Each node receiving **RREQ**, calculates the distance between itself and the neighbor who sent him RREQ (in this part we use the exact distance or the distance using the Pr), broadcasts its table [**neighbors-distance**] to its neighbors and choose the reference who has the smallest distance and recalculate the newest distances using the third method.

N.B.: the node who receive the **RREQ** is the node a in the previous part.

6. SIMULATIONS AND RESULTS

In the following simulations, we applied our proposition to the AODV protocol .For this, we have been used the simulator NS-2 [18], with its implementation of AODV protocol of the version NS-2.35 and Ying-3D [19] to represent some results in 3D.

6.1. Environment

The network size considered for our simulations is (1000m×1000m) . The nodes have the same configuration, in particular TCP protocol for the transport layer and Telnet for the application layer. Time for each simulation is of 60s. For each simulation the mobility of the nodes is represented by the choice of an uniform speed between $V_{min}= 0$ and $V_{max} = 100$ m/s. The nodes are moved after a random choice of the new destination without leaving the network (1000m×1000m).

6.2. Discussions of results

The results present the local quantification of neighbor's distances during the simulation. After the application of our proposition on a AODV protocol we obtain:

Table 1: Quantification's results

RREQ's source	RREQ's destination	Time	Distance	Distance/Pr
7	0	0,00128055	165,397	317,654
8	0	0,00128093	277,895	897,138
3	0	0,001281	299,47	1041,43
17	0	0,00128117	351,078	1430,93
15	0	0,00128119	356,321	1474,36
19	0	0,00128119	358,238	1490,13
11	8	0,00301544	152,84	271,179
9	8	0,00301572	238,948	662,809
1	8	0,00301579	258,118	773,428
7	8	0,00301585	276,525	887,669
0	8	0,00301585	277,731	896,079
15	8	0,00301586	280,446	913,022
10	8	0,00301598	315,33	1154,29
3	8	0,00301599	318,399	1176,86
5	8	0,00301617	371,712	1603,97
19	8	0,0030162	380,623	1681,8
1	15	0,00584584	308,585	1105,44
13	15	0,00584587	316,332	1161,63
5	10	0,00903948	103,368	124,039
4	10	0,00903976	187,032	406,083
9	10	0,00903978	194,023	437,009
11	10	0,00903982	206,903	496,956
1	10	0,00903988	224,54	585,286
18	10	0,00903999	256,821	765,673
16	10	0,00904017	309,608	1112,77
12	10	0,00904037	371,486	1602,02
13	10	0,0090404	378,926	1666,83
19	3	0,0112602	66,4831	66,4831
17	3	0,0112604	136,741	217,059
0	3	0,011261	299,018	1038,29

In the following figures, we observe the change of the distance between the node 1 and their neighbours in the first 3s of the simulation.

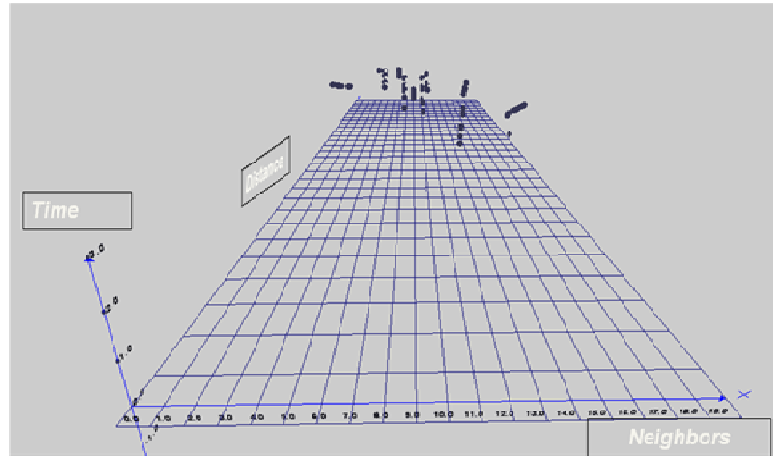


Figure 3: Quantification of the Distance between node 1 and its neighbors during the first 3s of the simulation

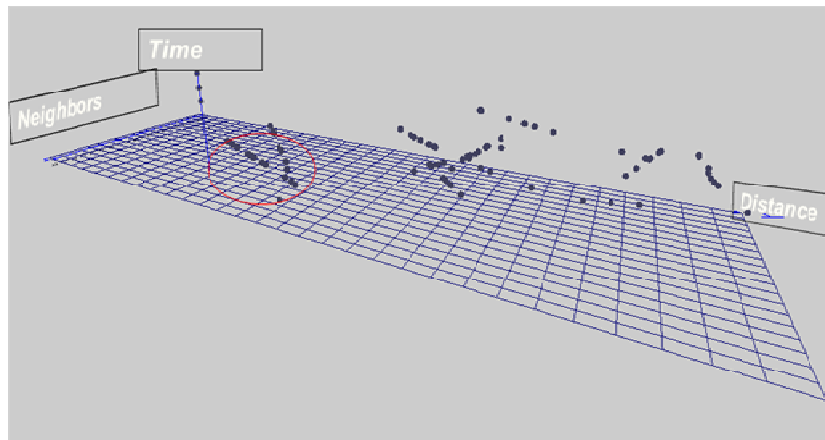


Figure 4: Quantification of the Distance between node 1 and its neighbors during the first 3s of the simulation “Another observation angle”

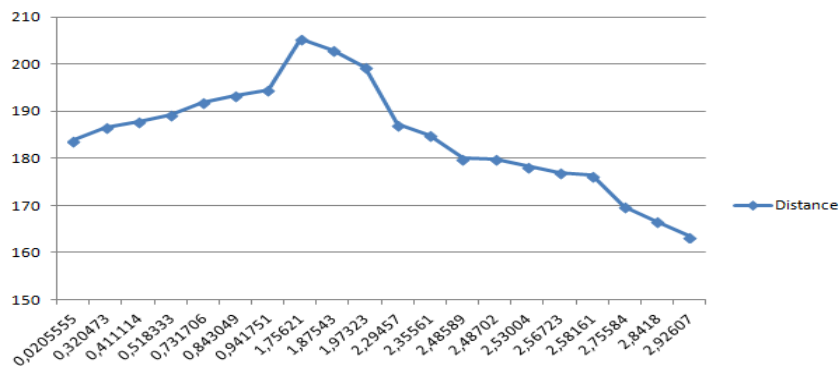


Figure 5: The distance between node 1 and its neighbor node 8 during the first 3s of the simulation

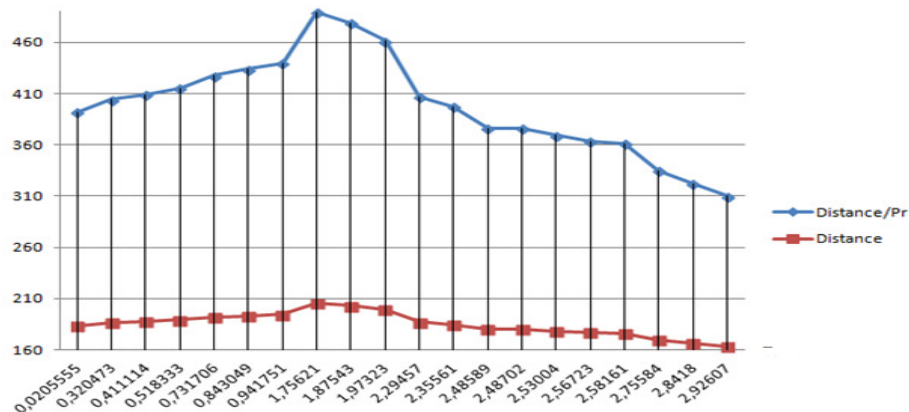


Figure 6: The distance and the distance using RSSI between node 1 and its neighbor node 8 during the first 3s of the simulation

7. CONCLUSION

In this paper, we tried to calculate a local distance between a node and its neighbors in a AODV routing protocol for the ad hoc networks. This metric of mobility that can be used to choose a stable rout to transmit data thus ameliorate the Quality of Service in this kind of networks.

To allow this proposition more really feasible, we present the three methods to calculate the distance between two nodes. First, we use the exact distance with a GPS or using RSSI. In case that the absolute positioning is not accessible, we propose our improved GPS-free implementing in AODV protocol.

REFERENCES

- [1] "Mobile Ad hoc Networking(MANET): Routing Protocol Performance Issues and Evaluation Considerations "Request for Comments 2501, IETF, January, 1999
- [2] J. Jubin and J. D. Tornow, "The DARPA Packet Radio Network Protocols," Proceedings of the IEEE, Vol. 75, No. 1, pp. 21-32, Jan. 1987.
- [3] B. M. Leiner, D. L. Neilson, F. A. Tobagui "Issues in Packet Radio Network esign," Proceedings of the IEEE, Vol. 75, No. 1, pp. 6-20, Jan. 1987.
- [4] Sabina Baraković, Suad Kasapović, and Jasmina Baraković, "Comparison of MANET Routing Protocols in Different Traffic and Mobility Models", Telfor Journal, Vol. 2, No. 1, 2010.
- [5] C. Perkins, B.-R. E. and D. S. "Ad hoc On-demand Distance Vector routing" Request For Comments (Proposed Standard) 3561, Internet Engineering Task Force, July, 2003.
- [6] <http://datatracker.ietf.org/wg/manet/charter/>
- [7] C. Cordeiro, H. Gossain and D. Agrawal "Multicast over Wireless Mobile Ad Hoc Networks: Present and Future Directions" vol. 17, no. 1, pp. 52-59, January/February, 2003
- [8] Sung-Ju Lee, Elizabeth M. Royer and Charles E. Perkins "Scalability Study of the Ad Hoc On-Demand Distance Vector Routing Protocol" In ACM/Wiley International Journal of Network Management, vol. 13, no. 2, pp. 97-114, March, 2003
- [9] Ian Chakeres Elizabeth and M. Belding-Royer "AODV Implementation Design and Performance Evaluation" to appear in a special issue on Wireless Ad Hoc Networkirng' of the InternationalJournal of Wireless and Mobile Computing (IJWMC), 2005
- [10] David B. Johnson, David A. Maltz and Josh Broch "DSR: The Dynamic Source Routing Protocol for Multi-Hop Wireless Ad Hoc Networks" Proceedings of INMC, 2004 - cseweb.ucsd.edu
- [11] Guoyou He. "Destination-sequenced distance vector (DSDV) protocol." Technical report, Helsinki University of Technology, Finland. 2 Dec 2003
- [12] P. Johansson, T. Larsson, N. Hedman, B. Mielczarek, and M. Degermark, "Scenario-based performance Analysis of Routing Protocols for Mobile Ad Hoc Networks," F'roc. ACM Mobicom 1999, Seattle WA, August 1999

- [13] X. Hong, M. Gerla, G. Pei, and C.-C. Chiang, "A Group Mobility Model for Ad Hoc Wireless Networks," Proc. ACM/IEEE MSWiM '99, Seattle WA, August 1999
- [14] N. Enneya, K. Ouidi and M. Elkoutbi "Network Mobility in Ad hoc Networks", Computer and Communication Engineering, 2008. ICCCE 2008. International Conference on, 13-15 May 2008, Kuala Lumpur, Malaysia.
- [15] N. Enneya, M. El Koutbi and A. Berqia "Enhancing AODV Performance based on Statistical Mobility Quantification", Information and Communication Technologies, 2006. ICTTA '06, 2nd (Volume:2), Pages 2455 – 2460.
- [16] Ahmad Norhisyam Idris, Azman Mohd Suldi & Juazer Rizal Abdul Hamid ,Effect of Radio Frequency Interference (RFI) on the Global Positioning System (GPS) Signals,2013 IEEE 9th International Colloquium on Signal Processing and its Applications, 8 - 10 Mac. 2013, Kuala Lumpur, Malaysia.
- [17] S. Capkun, M. Hamdi and J.P. Hubaux, "GPS-free positioning in mobile Ad-Hoc networks," Hawaii International Conference On System Sciences, HICSS-34 January 3-6, 2001 Outrigger Wailea Resort
- [18] "The network simulator - ns-2" January, 2006
- [19] <http://revue.sesamath.net/spip.php?article362>

AUTHORS

Meryem SAADOUNE

Received the master degree in Computer Engineering in 2010 from Faculty of Sciences -HASSAN 2 University- Casablanca, Morocco.

Currently, a PhD Student in Computer Science.

Ongoing research interest :

- QoS in Mobile Ad hoc Networks (MANETs) / Wireless Sensor Network (WSN)
- Next Generation Networks



Abdelmajid HAJAMI

PhD in informatics and telecommunications,

Mohamed V-Souissi University Rabat-Morocco. 2011

Ex Trainer in Regional Centre in teaching and Training

Assistant professor at the Faculty of Science and Technology of Settat in Morocco.

Member of LAVETE Lab at Faculty of Science and Technology of Settat

Research interests:

- Security and QoS in wireless networks
- Radio Access Networks
- Next Generation Networks
- ILE : Informatics Learning Environments



Hakim ALLALI

Was born in Morocco on 1966. He received the

Ph.D degree from Claude Bernard Lyon I University (France) in 1993 and the

"Docteur d'Etat" degree from Hassan II-Mohamedia University, Casablanca (Morocco) in 1997.

He is currently Professor at Faculty of Sciences and Technologies of Hassan 1st University of Settat (Morocco) and director of LAVETE Laboratory.

He is executive manager and founder of IT Learning Campus.

His research interests include technology enhanced learning, modeling, image processing, computer networking and GIS.



INTENTIONAL BLANK

DUAL BAND SEMI CIRCULAR DISK PATCH ANTENNA LOADED WITH L-SHAPED SLOT

Amel Boufrioua

Electronics Department, Technological Sciences Faculty,
University Constantine 1, Constantine, Algeria
boufrioua_amel@yahoo.fr

ABSTRACT

In this paper, a dual frequency resonance antenna is analysed by introducing L-shaped slot in a semi circular patch, different parametric studies have allows and the results in terms of return loss and radiation pattern are given. It is observed that various antenna parameters are obtained as a function of frequency for different value of slot length and width; it is easy to adjust the upper and the lower band by varying these different antenna parameters. The coaxial feed is used to excite the patch antenna. Theoretical results using Matlab are compared with the simulated results obtained from Ansoft HFSS and shown to be in good agreement.

KEYWORDS

L-shaped slot, semicircular patch antenna, dual band

1. INTRODUCTION

With the rapid development of wireless communications, it is desirable to design small size, low profile and wideband multi-frequency planar antennas. Over the past few years, single-patch antennas are extensively used in various communication systems due to their compactness, economical efficiency, light weight, low profile and conformability to any structure. The main drawback to implementing these antennas in many applications is their limited bandwidth. However, the most important challenge in microstrip antenna design is to increase the bandwidth and gain [1-10]. Several techniques that can be used to achieve multi-band performances such as multilayer stacked patch, multi resonator and insertion of slots and slits [2] of various shapes and sizes in the patch antennas have been proposed recently [1-10]. When a microstrip patch antenna is loaded with reactive elements such as slots, stubs or shorting pin, it gives tunable or dual frequency antenna characteristics [6]. The most popular technique for obtaining dual-frequency behavior is to introduce the slots on a single patch [1- 4].

In this paper, we present a semicircular microstrip patch antenna with L-shaped slot. The proposed antenna can completely cover two bands and provides a significant size reduction. Dual frequency is tuned by changing the dimensions of the slot. In this paper the simulation resultants and the performance analyses using Matlab and Ansoft HFSS software of the proposed semicircular microstrip patch antenna with L-shaped slot are presented, a comprehensive parametric study is carried out to investigate the effect of antenna design parameters on the return loss, the bandwidth and radiation of the proposed antenna.

2. L-SHAPED SLOT LOADED SEMICIRCULAR PATCH ANTENNA

The configuration of the proposed antenna is shown in Fig. 1. The semi circular microstrip patch of dimensions $W \times L$ printed on the grounded substrate, which has a uniform thickness of h and having a relative permittivity ϵ_r and the dielectric material is assumed to be nonmagnetic with permeability μ_0 . The analysis of the half disk patch antenna is similar to that of circular disk patch, but the effective radius changes to 50% reduction in size [6]. The L-shaped slot with dimension (L_s , W_s) is embedded in a semicircular patch (see Figure 2), the L-shaped patch semicircular antenna features dual-band behavior.

The patch is fed by a probe coaxial (50Ω) which is easy to fabricate and has spurious radiation [11]. In this feeding technique, the inner conductor of the coaxial connector extends from ground through the substrate and is soldered to the radiating patch, while the outer conductor extends from ground up to substrate.

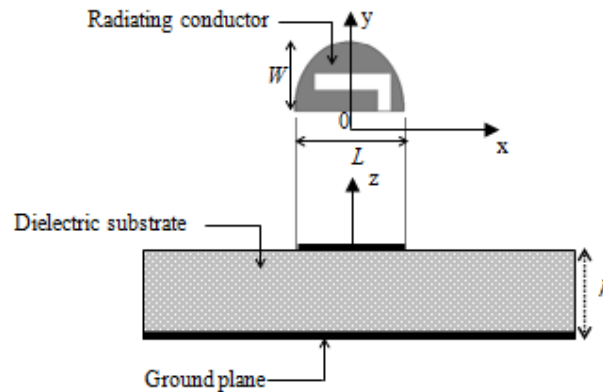


Figure 1. Geometry of L-shaped slot loaded semicircular disk patch antenna

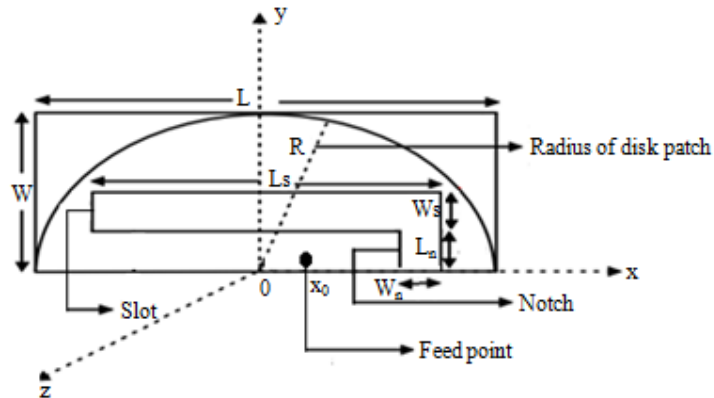


Figure 2. Dimensions of L-shaped slot loaded semicircular disk patch antenna

3. RESULTS AND DISCUSSION

Table 1 shows the different parameters of this proposed semicircular patch antenna loaded with an L-shaped slot with $\epsilon_r=1$. It is worth noting that the feed can be placed at any desired location inside the patch. In this study the feeding is accomplished with a probe coaxial located on the axial of symmetry of the antenna in the point of coordinates x_0 and y_0 .

The simulation is done through programs in Matlab and Ansoft HFSS; we compare our results with those available in the reference [6].

Table1. Design parameters of the proposed antenna.

Parameters	R	h	Ls	Ws	Wn	Ln	(x_0, y_0)
Value (mm)	40	15	46.2	6	2.5	6	(12.6, 0)

The return loss is studied in function of frequency; the effect of different physical parameters on the characteristics of the patch antenna is shown. From (figures 3a and b) given by Matlab code and HFSS software respectively it is clear that the proposed antenna resonate at two frequencies with two band widths

That is seen from the (Figure3. A), given by Matlab2013 code that the lower resonant frequency $fr1=1.48\text{GHz}$ and the upper resonant frequency $fr2=2.28\text{GHz}$. The -10 dB band width of lower and upper resonance frequencies are respectively $BP1=14.86\%$, $BP2=10.96\%$

From (Figure 3. b) obtained from HFSS14.0 software the two resonant frequencies are $fr1=1.42\text{GHz}$; $fr2=2.23\text{GHz}$. The -10 dB band width of the previous lower and upper resonance frequencies are respectively $BP1=8.42\%$, $BP2=19.74\%$.

The theoretical results using Matlab are found to be approximately in good agreement with the simulated results obtained with Ansoft HFSS software and with [6].

The variation of return loss S_{11} according to slot length "Ls" is shown by Figure 4 obtained from Matlab code, it is observed that the increase of the Ls, decreases both the lower and the upper resonance frequencies.

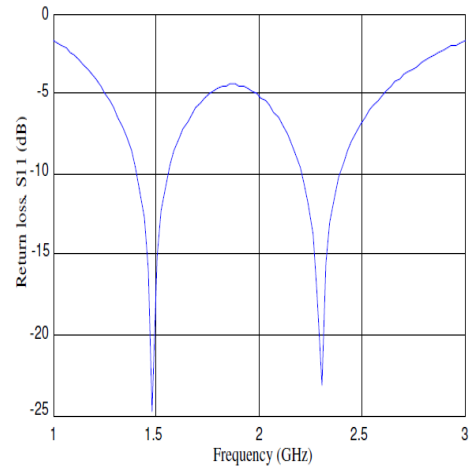
For the variation of S_{11} according to slot width Ws shown by Figure 5 (obtained from Matlab code), it is observed that the upper gap of the resonance frequencies decrease with increasing value of the slot width. It can be seen clearly that the slot length Ls and slot width Ws have a stronger effect on the lower resonance frequencies than the upper resonant frequencies.

Figures 6 and 7 show the variation of the return loss with frequency for different value of notch length and width respectively, it is observed that the notch length Ln and notch width Wn have a stronger effect on the upper resonance frequencies which increase with the increasing value of the notch length and width length than the lower ones which slightly increases with the increase of these parameters (Ln and Wn).

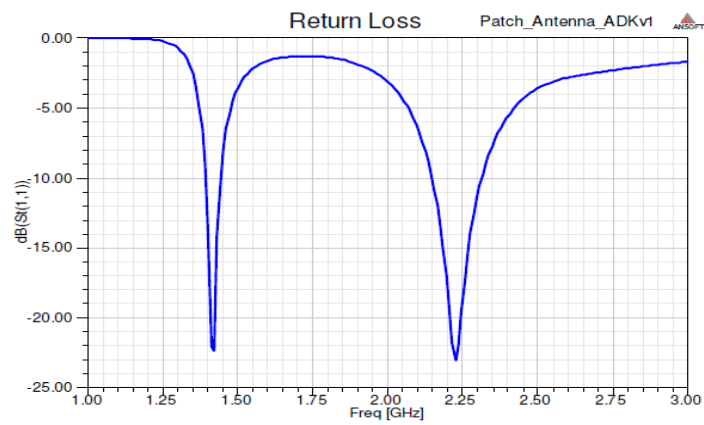
In Figure 8, it is shown that the feed locations have a stronger effect on both the lower and the upper resonance frequencies; furthermore the obtained results show that the lower resonance

frequencies vary more significantly when the x_0 change compared to the upper resonance frequencies as well as the other physical parameters of the proposed antenna.

Radiation pattern of the antenna is shown in Figure 9 and 10 for both upper and lower resonance in both principal planes E and H.



(a)



(b)

Figure 3. Comparative plot of return loss with frequency along with theoretical results obtained from Matlab code (a) and simulated results obtained from Ansoft HFSS software (b).

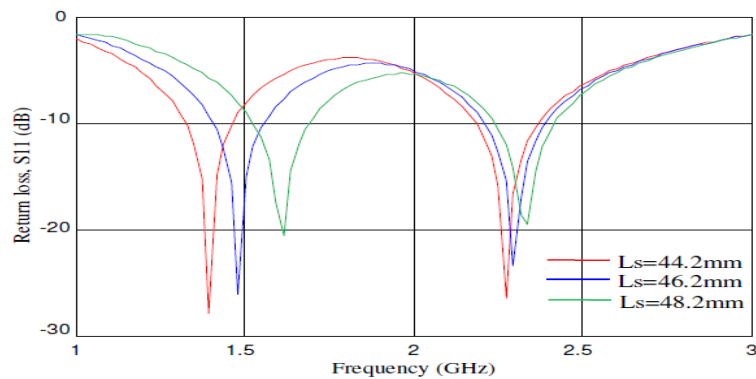


Figure 4. Variation of return loss S11 with frequency obtained from Matlab code for different value of slot length L_s .

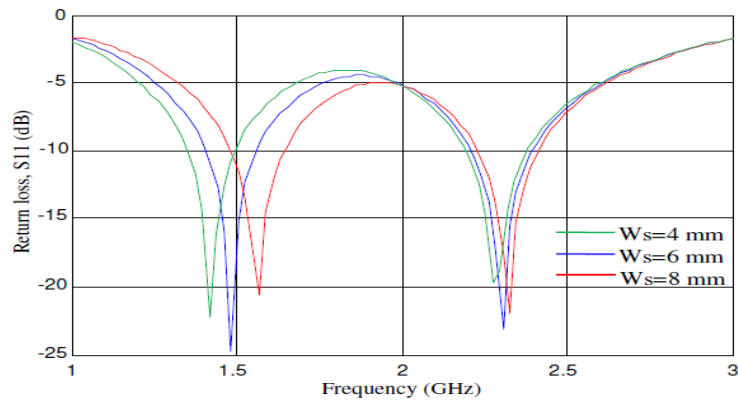


Figure 5. Variation of return loss S_{11} with frequency obtained from Matlab code for different value of slot width W_s .

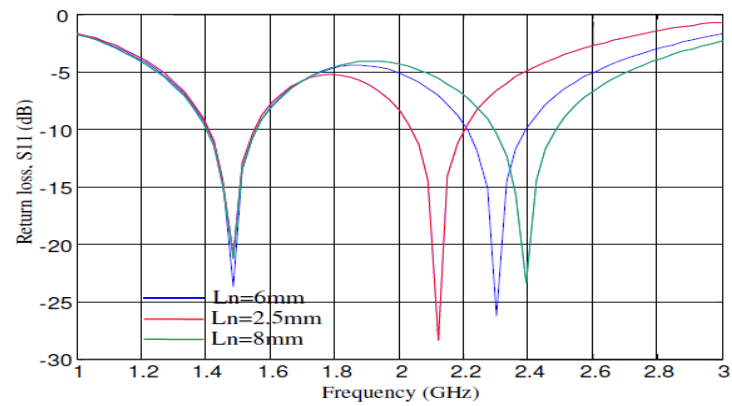


Figure 6. Variation of return loss with frequency for different value of notch length “L”

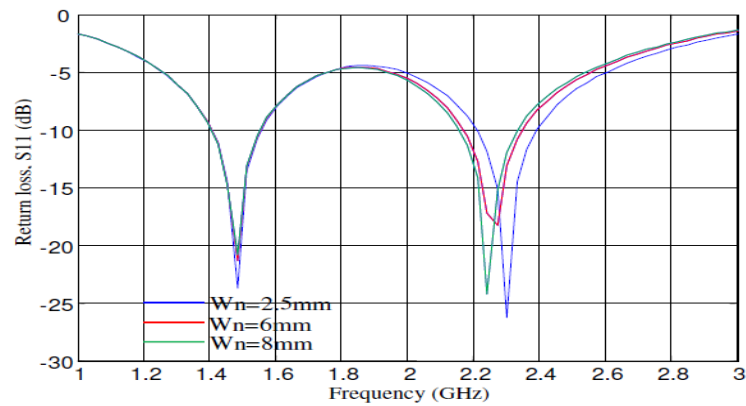


Figure 7. Variation of return loss with frequency for different value of notch width “ W_n ”

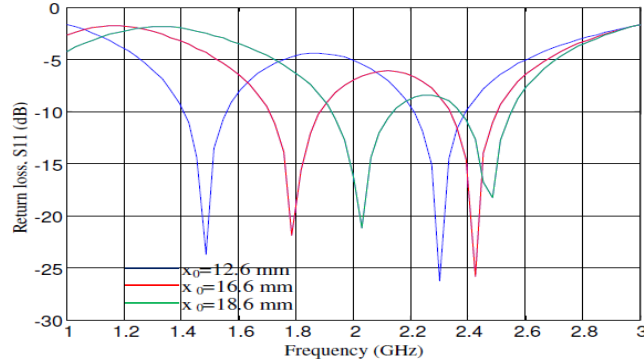


Figure 8. Variation of return loss with frequency for different value of feed point “ x_0 ”

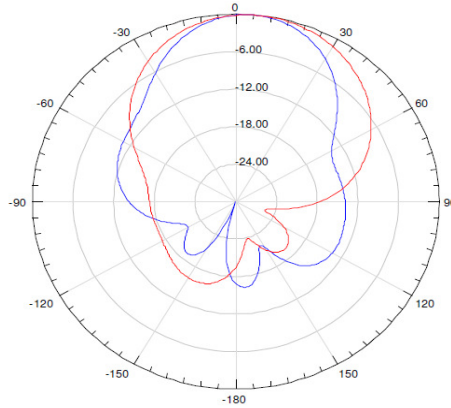


Figure 9. Radiation pattern of L-shaped slot loaded semicircular patch antenna for both upper resonant frequency (blue line) and lower resonant frequency (red line) at E plane

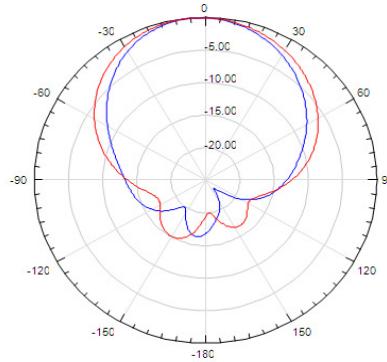


Figure 10. Radiation pattern of L-shaped slot loaded semicircular patch antenna for both upper resonant frequency (blue line) and lower resonant frequency (red line) at H plane

4. CONCLUSION

It is found that the proposed L-shaped slot loaded semicircular patch structure can operate at two resonance frequencies and consequently this proposed antenna can be used for dual band operation, also the effects of different physical parameters on the characteristics of this structure are investigated, the proposed structure can be scaled to meet different frequencies of wireless communication systems just by changing the dimension of the main antenna. Numerical results indicate that both the upper and lower resonant frequencies and the band widths depend on the

size of the slot dimensions, by properly choosing the location of feed point and the slots two bands can be achieved and controlled. Furthermore, the lower resonant frequencies and band widths are highly dependent on the slot dimensions as well as feed locations, however the upper resonant frequencies and band widths are highly dependent on the notch dimensions. In addition, the radiation pattern of both upper and lower resonant frequencies of the proposed antennas are presented in the principal planes E and H.

REFERENCES

- [1] J. A. Ansari, P. Singh and S. K. Dubey, (2008) "H-shaped stacked patch antenna for dual band operation," *Progress In Electromagnetics Research B*, Vol. 5, pp. 291–302.
- [2] A. A. Razaqi, M. Mustaqim and B. A. Khawaja, (2013) "Wideband E-Shaped Antenna Design for WLAN Applications," (ICET), *IEEE 9th International Conference on Emerging Technologies*, 9-10.
- [3] S. Maci and G. B. Gentili, (1997) "Dual-Frequency Patch Antennas" *IEEE Antennas & Propagation Magazine*, Vol. 39, No. 6.
- [4] I-F. Chen, C. M. Peng, and S-C. Liang, (2005) "Single Layer Printed Monopole Antenna for Dual ISM-Band Operation," *IEEE Trans. On Antennas and Propaga.*, Vol.53, N. 4. pp. 1270–1273.
- [5] J A. Ansari, S. K. Dubey, P. Singh, R. U. Khan, B. R. Vishvakarma, (2008) "Analysis of U-slot loaded patch for dualband operation," *International Journal of Microwave and Optical Technology*, Vol. 3, pp. 80-84.
- [6] J. A. Ansari, A. Mishra, (2011) "Half U-slot loaded semicircular disk patch antenna for GSM mobile phone and optical communications," *Progress In Electromagnetics Research C*, Vol. 18, pp. 31-4.
- [7] D. K. Srivastava, J. P. Saini, D. S. Chauhan, (2009) "Broadband stacked H-shaped patch antenna," *International Journal of Recent Trends in Engineering*, Vol. 2, pp. 385-389.
- [8] E. Wang, J. Zheng, (2009) "A novel dual-band patch antenna for WLAN communication," *Progress In Electromagnetics Research C*, Vol. 6, pp. 93-102.
- [9] A. Boufrioua, A. Benghalia, (2006) "Effects of the resistive patch and the uniaxial anisotropic substrate on the resonant frequency and the scattering radar cross section of a rectangular microstrip antenna" *Elsevier, AST, Aerospace Science and Technology*, Vol. 10, pp. 217-221.
- [10] A. Boufrioua, (2014) "Bilayer microstrip patch antenna with loaded with U and half U-shaped slots," (ICMCS), *IEEE International Conference on multimedia computing and systems*, 14-16.
- [11] J. J. Bahl and P. Bhartia, (1980) *Microstrip antennas*, Edited by M. A. Dedham, Artech House.

AUTHORS

Amel Boufrioua was born in Constantine, Algeria; she received the B.S. degree in Electronic Engineering in 1996, the M.S. and Ph.D. degrees in Microwave from Electronics Department, Constantine University, Algeria, in 2000 and 2006 respectively. From February 2002 to December 2003, she was a Research Assistant with Space Instrumentation Laboratory at the National Centre of Space Techniques "CNTS" (Oran, Algeria), and then in November 2003, she was an Assistant Professor at the Electronic Engineering Department (Constantine University). Since 2008, she is a Lecturer with the electronic department, University Constantine 1; her area of interest is microwave and microstrip antennas. Dr. Amel Boufrioua is the corresponding author and can be contacted at: boufrioua_amel@yahoo.fr



INTENTIONAL BLANK

LOW ALTITUDE AIRSHIPS FOR SEAMLESS MOBILE COMMUNICATION IN AIR TRAVEL

Madhu D¹, Santhoshkumar M K¹, Swarnalatha Srinivas² and
Narendra Kumar G¹

¹Bangalore University, Bangalore.

²Visvesvaraya Technological University, Belgaum.

madhudmm@gmail.com, santhoshkumarmksss@gmail.com,
gnarenk@yahoo.com
swarnalatha.ss@gmail.com

ABSTRACT

The Aviation Administration policy prohibits the use of mobile phones in Aircraft during transition for the reason it may harm their communication system due to Electromagnetic interference. In case the user wants to access cellular network at higher altitudes, base station access is a problem. Large number of channels are allocated to a single user moving at high speed by various Base Stations in the vicinity to service the request requiring more resources. Low Altitude Platforms (LAPs) are provided in the form of Base stations in the Airships with antennas projected upwards which has direct link with the Ground Station. LAPs using Long-Endurance Multi-Intelligence Vehicle (LEMVs) equipped with an engine for mobility and stable positioning against rough winds are utilized. This paper proposes a system that allows the passengers to use their mobiles in Aircraft using LAPs as an intermediate system between Aircraft and Ground station. As the Aircraft is dynamic, it has to change its link frequently with the Airships, MANETs using AODV protocol is established in the prototype using NS2 to provide the service and the results are encouraging.

KEYWORDS

Airships, AODV, Seamless Mobile Communication.

1. INTRODUCTION

Mobile phone is one of the active radio transmitter emitting electromagnetic radiations causing interference in the communication between the pilot and air traffic control unit [1]. The radiations also interfere with sensitive galvanometer based displays of older planes. The interference is caused because the cellular towers might be miles below the aircraft and the phone might have to transmit at its maximum power to establish a connection, increasing the risk of interference with Aircraft system. This paper proposes a system that operates at a frequency other than the Aircrafts communication system. Hence, the mobile can easily establish a connection with the Ground Station without having to transmit at its maximum power using Airships as the intermediate Base Stations. Airships are placed at appropriate positions along the path of the Aircraft in the required region, to provide communication services.

2. EXISTING METHOD

GSMOB (GSM On Board) [2], [3], [4] mobile services will allow airline passengers to use their own mobile terminals during certain stages of flight. Passengers are able to make and receive calls, send and receive SMS text messages and use GPRS functionality. The frequencies used for onboard communication are in the GSM1800 band. The main reasons for the selection of these bands is due to the small transmission power for an individual terminal when compared to the 900MHz band and emissions at higher frequencies produce higher path loss. A functional overview of a GSMOB system, Fig. 1.

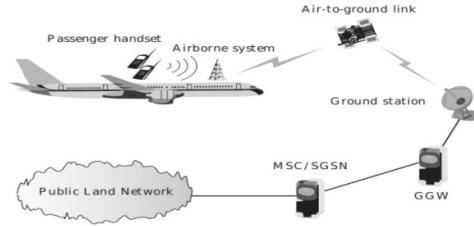


Figure 1. An overview of GSMOB system.

The challenges faced by the GSMOB system is to control the radio emissions of the mobile phones used by passengers, called Aircraft Mobile Stations (AcMS) and the on-board transmitters. AcMS try to connect to the cellular station even when the Aircraft is at cruising altitudes. Hence the AcMS transmit at higher power levels increasing the risk of interference. The log-on procedure used by all mobile phones on the market today is depicted in Fig. 2.

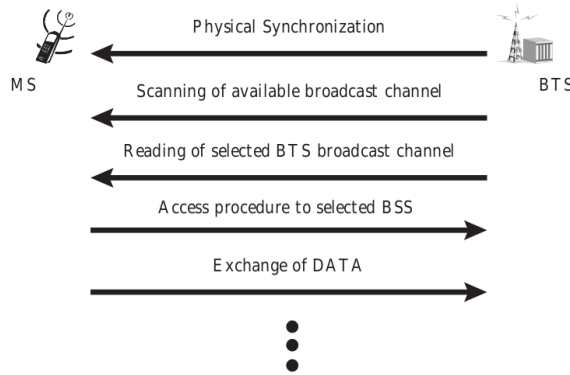


Figure 2. Mobile phones log-on procedure.

A technical approach for controlling the radio emissions aboard the aircraft is by making use of Network Control Unit (NCU) which prevents AcMS from attaching to the cellular network by injecting wideband noise of low power density into the relevant frequency bands, by which signals from cellular networks are effectively screened. Hence the cellular networks become invisible to the AcMS and they can transmit in a controlled manner by connecting to the Aircraft Base Transceiver Station (AcBTS) with the end to end architecture of the GSMOB, Fig 3.

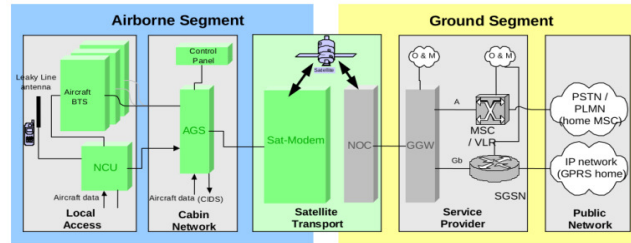


Figure 3. End to end architecture of GSMOB system.

Ground Segment- The ground segment consists of Service Provider Domain which includes Ground Gateway (GGW) and network components such as Mobile Switching Centre (MSC), Visitor Location Register (VLR) and Serving GPRS Support Node (SGSN) etc. The routing of the Aircraft traffic towards terrestrial backbone network of the Public Domain, Billing functions, mobility management are taken care of by the Service Provider Domain. The Public Network Domain of the Ground Segment provides the interconnection of the call, data or signalling communication to the relevant public network end points.

Airborne Segment- The Airborne Segment consists of the Local Access Domain and the Cabin Network Domain. The Local Access Domain contains the AcBTS providing GSM access for passengers AcMS and the NCU. The Cabin Network Domain contains the control panel and an Aircraft GSM Server (AGS). The control panel enables the crew to control the states of the GSMOB system. The AGS combines the GSM software on-board and interconnects the mobile phone system with the satellite modem.

3. PROPOSED MODEL

As the Aircraft enters the cruising altitude it spends much of its time during the flight in the range of 25,000 to 40,000 feet. In order to provide communication between the Aircraft and Ground station an Airship is used which is located at a height of about 20,000 feet. A scenario in which communication services are rendered to multiple Aircrafts which are separated by different Horizontal (minimum of 300m) and Vertical (minimum of 9.26km) distances [5] in region of interest is depicted, Fig. 4.

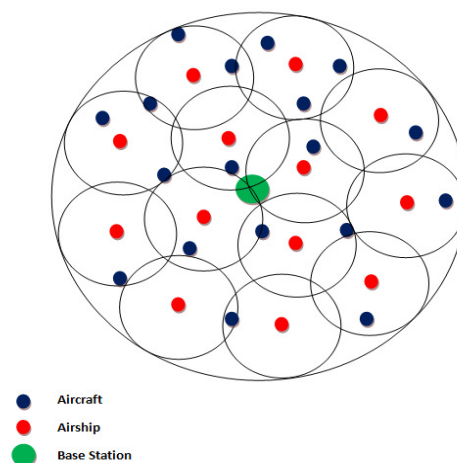


Figure 4. Airship assisted communication system for Aircraft.

3.1. Aircraft Communication System

This system consists of dedicated transceivers for users in Aircraft to communicate with the Ground station. The Passengers can use their own terminals to make and receive calls, send and receive SMS and GPRS functionalities.

3.2. Communication Channel

Airships are located at lower altitudes (around 20,000 feet) compared to satellite, the effects such as signal delay, noise and interference are less. LEMV [6], [7] are large helium-filled balloon like

Airships with an aerodynamic "cigar" shape, about 91m in length, 34m in width, 26m in Height, 38,000 cubic metres of envelope and has 4 x 350HP, 4 litre supercharged V8 diesel engines which can carry payloads of up to 2,750lbs. The altitude of 20,000 feet is high enough to give local coverage of about 30km in diameter and also offers the advantage of minimum wind speeds. LEMV can be optionally manned, remotely piloted or autonomously operated that consumes about 3,500 gallons of fuel to remain aloft continuously for a period of 21 days. The vehicle can fly at a loiter speed of 30kt and a dash speed of 80kt. There will be two types of antenna in airship: Master antenna and Slave antenna. The Airships along the path of the Aircraft are synchronised with one another with the help of the Master antenna to provide a regional coverage. The Master antennas of the Airships is also used to establish a continuous communication link with the ground station antenna. Slave antenna is used to capture the signal from the aircraft, the captured signal is then forwarded to the Master antenna and vice versa.

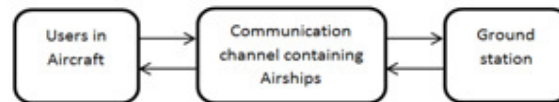


Figure 5. Proposed communication system.

3.3. Ground Station

It is the earth based communication station providing the communication link to the users in Aircraft. The earth station itself is usually an antenna that includes low noise amplifier, a down converter as well as an electronic receiver. The ground station is connected to Mobile Switching Centre for further switching operations.

4. PROCESS

In case a user in the Aircraft initiates a call, MANET routes the call to the Airship. The transceiver antenna in the Airship receives the call request and transmits it to the ground station which passes it to the MSC and connects the call to the desired user. Similarly in case a user on ground initiates the call is established and completed in the reverse direction. In case an Aircraft comes into the coverage region of particular Airship, slave antennae captures the call request signal and forwards it to the Master antenna which is in continuous communication with the ground station. In case the Aircraft comes out of the communication range of particular Airship, Handoff takes place. The process flow of the complete scenario is depicted, Fig. 6.

5. ROUTING

The Ad hoc On Demand Distance Vector (AODV) is a reactive routing Protocol, the routes are determined during requirement. AODV is capable of both unicast and multicast routing. It maintains the routes as long as they are needed by the sources. Each active node periodically broadcasts a Hello message that all its neighbours receive, in case a node fails to receive several Hello messages from a neighbour, a link break is detected. Data transmitted by source to an unknown destination broadcasts a Route Request (RREQ) for that destination. At each

intermediate node a RREQ is received and a route to the source is created. In case the request receiving node is not the destination and does not have current route to the destination, it rebroadcasts the RREQ. In case the receiving node is the destination or has a current route to the destination, it generates a Route Reply (RREP). The RREP is unicast in a hop-by-hop fashion to the source. As the RREP propagates, each intermediate node creates and records a route to the destination. In case multiple RREPs are received by the source, the route with the shortest hop count is chosen and when a link break is detected during data flow, a Route Error (RERR) is sent to the source of the data in a hop-by-hop fashion, invalidates the route and reinitiates route discovery, Fig. 7.

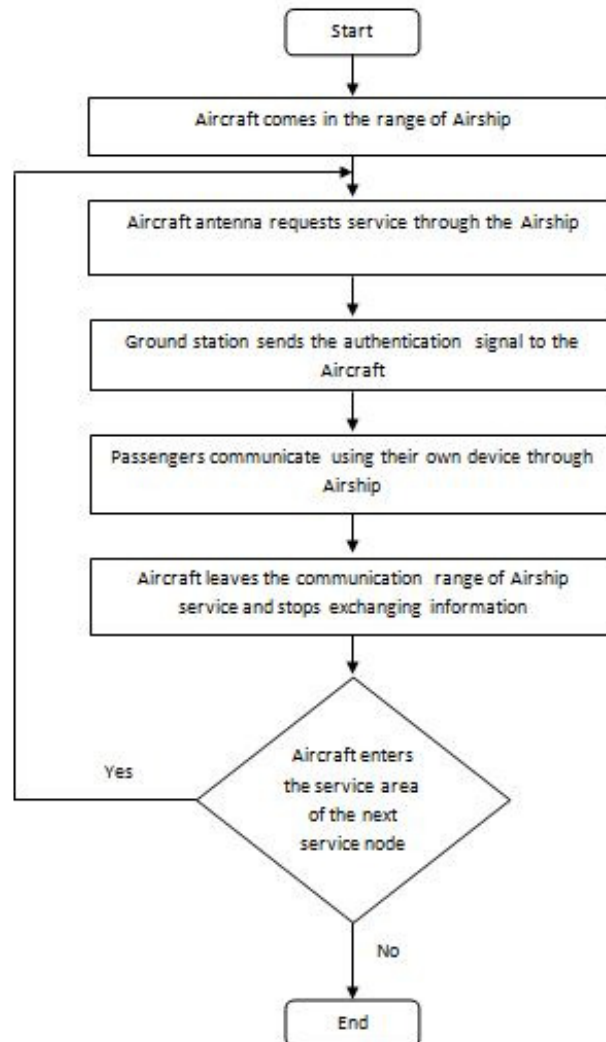


Figure 6. Process flow.

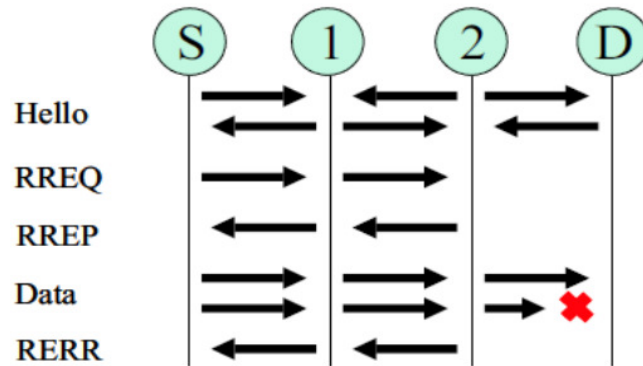


Figure 7. AODV Protocol Messaging.

6. IMPLEMENTATION

In order to provide communication, the Aircraft needs to consist of the subsystems, Fig. 8.

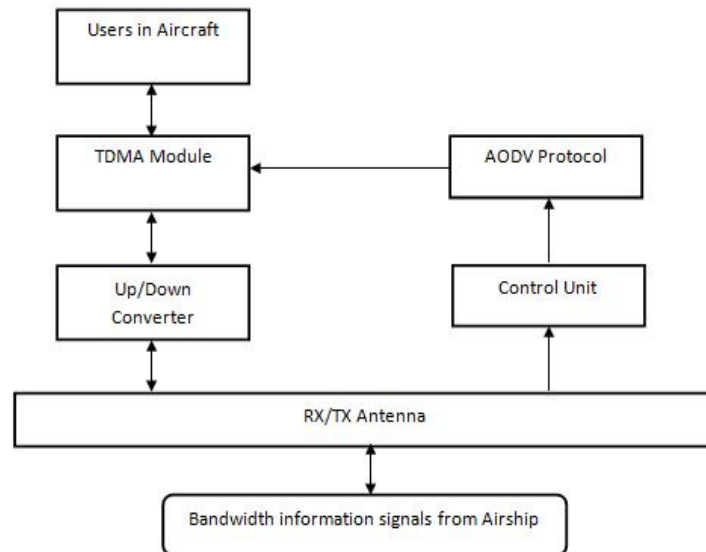


Figure 8. Aircraft system.

The users in the Aircraft requiring communication service can use their mobile devices which can be achieved by using a TDMA module aboard the Aircraft. The TDMA module allocates different time slots to different users to provide the service. The Up/Down converter in the Aircraft up converts the signal to be transmitted to antennas frequency and down converts the received signal to the mobiles operating frequency. The control unit present in the Aircraft receives the bandwidth information signals from the Airship and decides on the bandwidth to be allocated to the users in the Aircraft with the help of the TDMA module implementing AODV protocol to determine the shortest path and provide the point to point communication service to the users.

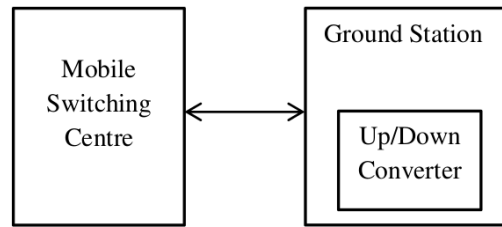


Figure 9. Ground station.

The Airship is the intermediate router between Aircraft and Ground station, also Guidance and Navigation system is used to control the position of the Airship. The Airships in a specified region have their own dedicated Ground station for transmitting and receiving the signal. The Ground stations antenna should be in line of sight with the closest Airships antenna and they should communicate continuously with each other to render the requested service. The ground station is synchronized with the MSC, Fig. 9 and the Up/Down converter in it has the same functionality as that of Aircraft's.

7. SIMULATION AND RESULTS

Basic Model Configuration: The nodes are positioned for Airships, Ground Station and Aircraft. Red colour nodes serve as Airships, Green nodes as Ground stations and Blue nodes as Aircrafts. Since an Airship provides circular coverage of 30km diameter, they are placed consecutively along the path of Aircraft and all of these Airships are in sync with a dedicated Ground station depicted, Fig. 10.



Figure 10. Basic Model of Simulation.

Experimental Analysis and Results: A TCL program is written to simulate the required topology of wireless network in NS2. The wireless simulation related parameters are defined as follows:

- **Channel Type** : Wireless
- **Radio-propagation model** : Two Ray Ground
- **Network interface type** : Wireless Phy
- **MAC type** : Mac/802.11
- **Interface queue type** : Queue/ Drop Tail/ Pri Queue
- **Link layer type** : LL
- **Antenna model** : Antenna/Omni
- **Max packet in ifq** : 50
- **Routing protocol** : AODV

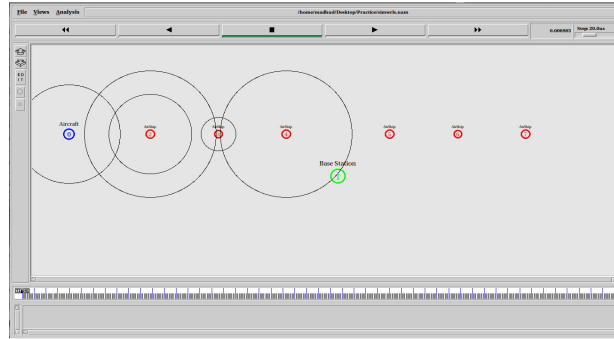


Figure 11. Aircraft linking with Airship.

In case an Aircraft comes within the communication range of an Airship, it sends request message to Ground station and in turn receives an authentication using Airships as an intermediate router. The exchange of information between Aircraft, Airship and the Ground station is depicted, Fig. 11. Gradually when the Aircraft comes out of the communication range of currently linked Airship, the Handoff to the next Airship takes place as depicted, Fig. 12.

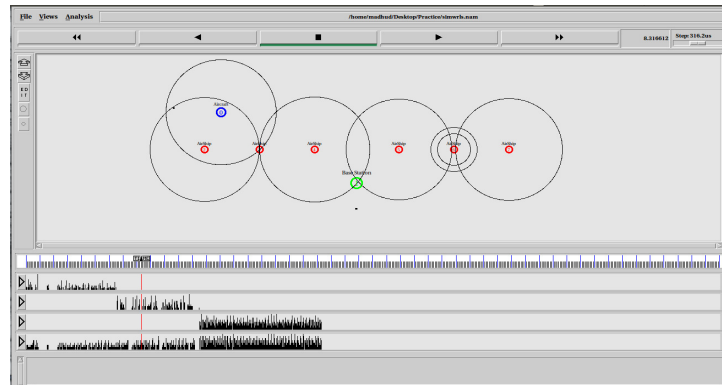


Figure 12. Aircraft changing its link to other Airship.

In case two aircrafts simultaneously arrive in a regional coverage area provided by dedicated Airships and Ground station, both the Aircrafts are provided with the communication services simultaneously, Fig. 13. The continuous communication services provided to passengers in multiple Aircrafts in different regions of interest is depicted, Fig. 14.

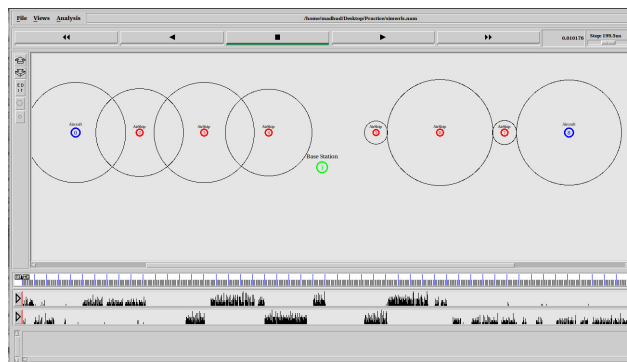


Figure 13. Two Aircrafts in a regional coverage of dedicated Airships and Ground station.

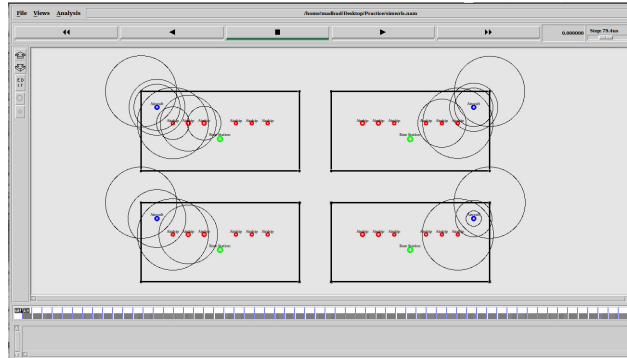


Figure 14. Continuous communication service in different regions of interest.

8. CONCLUSION

The paper aims to provide seamless communication services to the mobile users in the Aircraft, they are enabled to operate their own devices by using the communication system installed in the Aircraft. The Airship is used as an intermediate router to forward the information between Aircraft and ground station and vice-versa. In the proposed model the delay incurred by the signal along its path will be reduced. As communication services can be readily established in regions of interest, the proposed model can be utilized in relief measures for natural disasters like Floods, Earth quakes etc. Since there exists a line of sight between the Airships and the Ground station, the use of Airships to provide communication services to the users on ground is another possibility.

REFERENCES

- [1] See Website- "http://en.wikipedia.org/wiki/Mobile_phones_on_aircraft".
- [2] César Gutiérrez Miguélez, "GSM operation onboard aircraft", ETSI White Paper No. 4, January 2007.
- [3] Carlos Gonzaga Lopez, "GSM ON BOARD AIRCRAFT", December 15, 2008.
- [4] John Mettrop, "GSM On-Board Aircraft", Directorate of Airspace Policy Surveillance & Spectrum Management.
- [5] Aircraft Separation- "http://www.nokaviation.com/PPL_trg/rules.htm".
- [6] See Website- "<http://www.defenseindustrydaily.com/rise-of-the-blimp-the-us-armys-lemv-06438/#>".
- [7] See Website- "<http://www.army-technology.com/projects/long-endurancemulti-intelligence-vehicle/>".
- [8] See Website- "<http://www.slideshare.net/rishikeshims/aircraftcommunicationsystems>".
- [9] ETSI ETS 300 326-1, "Radio Equipment and Systems (RES)", "Terrestrial Flight Telephone System (TFTS)".
- [10] ETSI ETS 300 326-2, "Radio Equipment and Systems (RES)", "Terrestrial Flight Telephone System (TFTS)".
- [11] Network Simulator Tutorial- "<http://www.isi.edu/nsnam/ns/tutorial>".
- [12] NS2 Range Calculation- "<http://mailman.isi.edu/pipermail/nsusers/2012-August/072160.html>".
- [13] Data Transfer in NS2- "<http://csis.bitspilani.ac.in/faculty/murali/resources/tutorials/ns2.htm>".

AUTHORS

Dr. Narendra Kumar G, born in Bangalore on 5th February, 1959. Obtained Masters Degree in Electrical Communication Engineering, (Computer Science & Communication) from Indian Institute of Science, Bangalore, Karnataka, India in 1987. Was awarded PhD in Electrical Engineering(Computer Network) from Bangalore University, Bangalore, Karnataka, India in 2006. Currently Professor in the Department of Electronics & Communication Engineering, University Visvesvaraya College of Engg., Bangalore University, Bangalore, held the positions of Associate Professor, Lecturer and Director of Students Welfare. Research interests include Mobile Communication, Wireless Communication, E-Commerce, Robotics and Computer Networks.



Dr. Swarnalatha Srinivas, born in Bangalore on 22nd October, 1964. Obtained Bachelors Degree in Electrical Engineering from University Visvesvaraya College of Engg., Bangalore, Karnataka, India in 1988. Obtained Masters degree in Power Systems, University Visvesvaraya College of Engg., Bangalore, Karnataka, India in 1992 and was awarded PhD under the guidance of Dr Narendra Kumar G in 2014. Currently Associate Professor in the Department of Electrical Engineering, Bangalore Institute of Technology, VTU, Bangalore.



Madhu D and **Santhoshkumar M K** are research students under the guidance of Prof. Narendra Kumar G.

ON THE MODELING OF OPEN FLOW-BASED SDNS: THE SINGLE NODE CASE

Kashif Mahmood¹, Ameen Chilwan², Olav N. Østerbø¹ and Michael Jarschel³

¹Telenor Research, Norway, ²Department of Telematics, NTNU, Norway, ³Nokia, Germany

¹kashif.mahmood@telenor.com, ¹olav-norvald.osterbo@telenor.com, ²chilwan@item.ntnu.no, ³michael.jarschel@nsn.com

ABSTRACT

OpenFlow is one of the most commonly used protocols for communication between the controller and the forwarding element in a software defined network (SDN). A model based on M/M/1 queues is proposed in [1] to capture the communication between the forwarding element and the controller. Albeit the model provides useful insight, it is accurate only for the case when the probability of expecting a new flow is small.

Secondly, it is not straight forward to extend the model in [1] to more than one forwarding element in the data plane. In this work we propose a model which addresses both these challenges. The model is based on Jackson assumption but with corrections tailored to the OpenFlow based SDN network. Performance analysis using the proposed model indicates that the model is accurate even for the case when the probability of new flow is quite large. Further we show by a toy example that the model can be extended to more than one node in the data plane.

KEYWORDS

OpenFlow, Performance analysis, Queuing theory, Software defined networks.

1. INTRODUCTION

Software defined networking (SDN), an academic lead initiative, has already made a lot of impact in the datacenters. As early as January 2012, Google had their full scaled datacenter WAN running as OpenFlow based SDN [2]. SDN is now all set to roar in the carrier networks domain too. This is because SDN promises network deployment and service upgrade on software which has huge benefits for the network operators because in the future the network operators will not compete on the basis of network coverage alone but on the basis of features and services.

All this has been possible due to the basic architectural principle of SDN which is the separation of the control plane from the data plane. The architecture involves SDN controller(s) residing in the control plane while the forwarding element(s) make the data plane. In order to handle the communication between the control plane and the data plane elements, OpenFlow is the only open, standard protocol[3].

OpenFlow started as a test protocol in Stanford but is now managed and maintained by Open Networking Foundation (ONF)[3]. It started with OpenFlow version 1.0.0 and at the writing of the paper, version 1.4 has been specified[3]. The working principle is the same but each version involves some additional features. For example the version 1.1.0 has support for group tables which was not there in version 1.0.0. The work in this paper is based on OpenFlow version 1.0.0 and we believe that it can be easily extended to the new versions.

Under an OpenFlow network, the controller-to-switch communication takes place as follows: where we use the term switch and node interchangeably to represent the forwarding element in the data plane in an SDN network.

When a *flow* with no specified forwarding instructions comes into a network the following actions are taken:

- i. A packet (or part of the packet) of the flow is sent by the switch to the controller, assuming that the switch is not configured to drop unknown packets.
- ii. The controller computes the forwarding path and updates the required nodes in the data path by sending entries to be added to the flow tables.
- iii. All subsequent packets of the flow are forwarded based on pre-calculated forwarding decisions and do not need any control plane action.

It is important to model the controller-to-switch communication for the performance analysis of OpenFlow(OF)-based SDN networks. The modeling of OpenFlow networks will help us to answer questions such as how much data we can pump into the network, what is the packet sojourn time, when and what (switch or the controller) is the bottleneck in a network.

Most of the work on performance analysis of SDN networks is based on simulations or experimentations. Albeit their benefits, analytical modeling is a time efficient alternative because setting up an SDN experiment or performing a simulation can take hours. But the real strength of analytical model lies in the extent to which it can be used for analysing networks and confidence that could be put in the obtained results.

The analytical model should be able to capture actual OpenFlow working principle and at the same time shall be flexible to handle any amount of query traffic going to the controller. Further, the model shall be readily extendable to more than one node in the data plane.

The analytical modeling of OpenFlow-based networks has only been attempted in a handful of papers before. For example feedback oriented queuing theory has been used in [1] to capture the control plane and data plane interaction where the Markovian servers are assumed for both the controller and the switch. However the model becomes less accurate as the probability of traffic going to the controller increases. Secondly it is not clear how the model can be extended to more than one switch in the data plane.

In [4], a network calculus based approach is used to quantify the packet processing capability of the switch in the data plane. However the feedback between the nodes in the data plane and the controller is not considered. This shortcoming of feedback modeling is addressed in [5]. However the model is depicted only for a single node in the data plane and the time stopping method employed therein has limited real time application. Secondly the framework used in [4] and [5] is based on deterministic network calculus which does not provide any meaningful bounds[6]. To the best of our knowledge apart from these handful analytical works, almost all the other efforts of evaluating performance of OpenFlow-based networks are carried out by simulations or

measurements, for example [7], [8], [9]. Moreover, Cbench tool to benchmark the controller performance is also introduced in [10] and is proved to be instrumental in benchmarking.

It is therefore of paramount importance to have an analytical model which can capture the feedback interaction between the controller and the switch, is able to model any amount of traffic going from switch to controller (and vice versa), and can be easily extended to more than one switch in the data plane. The model proposed in this paper is an attempt in that direction. We model the OpenFlow network as a Jackson network but with a modification to accurately represent the traffic flow from the switch to the controller in an actual OF-based SDN network.

It is highlighted later in the paper (Fig.2) that this modification to the native Jackson network is necessary to capture the OpenFlow working principle. The model is in turn used for performance analysis of OF-based SDN networks to calculate the mean packet sojourn time and to find out how much data we can pump into the network. The main contributions of this work are:

- A model is proposed to capture the feedback interaction between the switch and the controller mimicking an actual OpenFlow based SDN network
- The model is accurate even for the case when large amount of new flows are arriving at the switch.
- The model can be easily extended to more than one switch in the data plane.
- We show mathematically that the packet sojourn time calculated by our proposed model based on Jackson assumption is the same as the one explicitly calculated for OF-based SDN network in [1].

The rest of the paper is organized as follows; we first present the system model along with the limitations and the necessary preliminaries in Section 2. The performance measures are outlined in Section 3, while the numerical results are presented in Section 4. An insight as to how the proposed model can be used for multi-node case is highlighted in Section 5, while the conclusions along with the future research directions are presented in Section 6.

2. MODEL DESCRIPTION

We assume that the overall traffic process at the switch and at the controller follows Poisson process similar to [1] given that the two processes are on a different time scale. Further we assume Markovian servers for the switch and the controller wherein we incorporate the transmission time of the packets from the switch to the controller in the service time of the controller. As for the buffer size we assume infinite buffer for the switch and the controller.

We use Jackson network model to represent the OF-based SDN network. To this end a recap of the Jackson model for open queuing networks[11], in which the nodes behaves locally as single M/M/1 queues, is outlined. Albeit trivial, this is done in order to highlight that Jackson model cannot be used as such for modeling OF-based SDNs.

Let us consider a Jackson network consisting of two nodes I and c connected in a feedback path as shown in Fig. 1(a). The service rates of nodes I and c are exponentially distributed with average values of μ_I and μ_c , respectively. The external traffic arrival to the node I is denoted as λ_I packets per unit time.

Let Γ_I be the net input to node I out of which $\Gamma_c = q_I^{jack} \Gamma_I$ goes to the node c where q_I^{jack} is the probability that the packet goes to the node c .

Assuming that no packet is lost at the controller (infinite buffer at the controller) the balance equation for the system can be written as

$$\lambda_1 = \lambda_1 + q_1^{jack} \lambda_1 \quad (1)$$

It is the term q_1^{jack} which needs modification in order to model OF-based SDN in Fig. 1(b) as a Jackson network in Fig. 1(a). This is because an OF-based SDN, as shown in Fig. 1(b), has the following two salient features:

- i. A packet coming to any node in the data plane will at most visit the controller once.
- ii. Only a fraction of the external traffic λ_1 and not a fraction of the net input traffic Γ_1 will go the controller. (Two lines directly out of the data node in Fig. 1(b) as opposed to one line in Fig. 1(a) are used to represent this phenomenon).

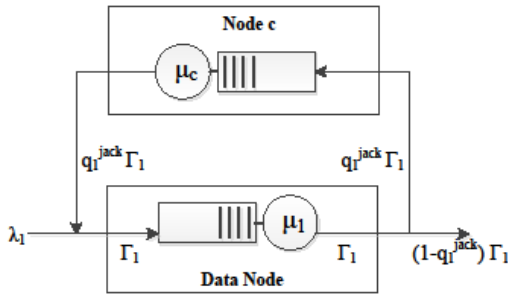


Figure 1(a): Jackson Model

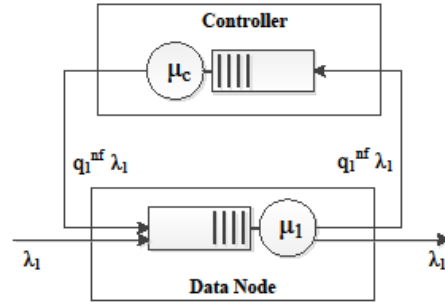


Figure 1(b): Simple OpenFlow Model [1]

In an OpenFlow network, let q_1^{nf} be the probability that the packet goes to the controller in case there is no flow entry in flow-table of the node, then in order to use the Jackson network to represent the OF-based SDN we have to adjust q_1^{jack} by demanding that the input rates to the nodes in both the models are the same. Hence

$$\lambda_1 = \lambda_1 + q_1^{nf} \lambda_1 \quad (2)$$

and

$$q_1^{jack} \lambda_1 = q_1^{nf} \lambda_1 \quad (3)$$

As a result q_1^{jack} can be solved as

$$q_1^{jack} = \frac{q_1^{nf}}{1 + q_1^{nf}} \quad (4)$$

Fig. 2 highlights the need of having a modified Jackson model to represent OF-based SDN networks where we represent mean time spent by a packet in the network (node + controller) as a function of load on the controller.

The curve simulation is obtained from simulating the OpenFlow behavior taking into account the aforementioned two salient features. Further, in the simulation, we assume Poisson arrivals at the input and exponentially distributed service times for the nodes.

The curve denoted by Jackson Model is obtained by using q_1^{jack} as such without modification i.e. $q_1^{jack} = q_1^{nf}$ while the curve Modified Jackson Model is based on q_1^{jack} in (4).

It can be seen that as the percentage of traffic going to the controller dictated by q_l^{nf} increases, the modification to the probability q_l^{jack} in (4) becomes all the more important.

2.1. Limitations

The work in this paper makes the following assumptions:

- The overall traffic arrival process at the switch and the controller is Poisson. Further exponentially distributed service times are used for the switch and the controller. This allows us to use the Jackson network results based on M/M/1 queues.
- Secondly we assume a single queue at the switch instead of a separate queue per line card.
- TCP traffic is used for which only the first packet of the unknown flow is sent to the controller.
- Infinite buffer is assumed at the switch as typically it is quite large.

It needs to be emphasized that the main goal of this work is to develop an analytical model for OF-based SDN networks. The assumptions will be relaxed in the subsequent work.

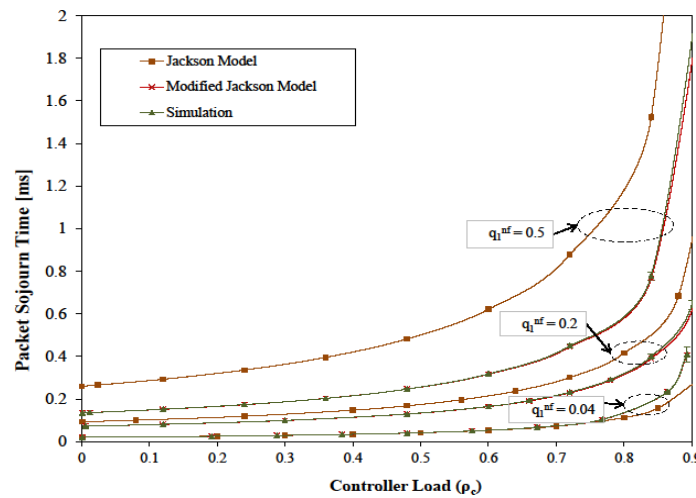


Figure2: Jackson model cannot be used as such

3. PERFORMANCE MEASURES

One of the advantages of our proposed model is that we can leverage the well-established results for performance analysis of Jackson networks for analysing OpenFlow-based SDNs. In this section, we will use the proposed model to find two elementary performance measures, the average packet sojourn time, and the distribution of time spent by a packet in the network.

3.1. Average Packet Sojourn Time

The average packet sojourn time, $E[W^{jack}]$, is defined as the time spent by a packet in the network from the moment it enters the network at its source node, until it leaves through the destination. $E[W^{jack}]$ for the network in Fig. 1 is given as [11]

$$E[W^{jack}] = \frac{1}{\lambda_l} \left(\frac{\rho_l}{1 - \rho_l} + \frac{\rho_c}{1 - \rho_c} \right) \quad (5)$$

where $\rho_l = \Gamma_l/\mu_l$ and $\rho_c = \Gamma_c/\mu_c$ denote the load on the node l and the controller c , respectively. Further, in order to have a stable system, it is assumed that all the loads are less than unity that is $\rho_l < 1$ and $\rho_c < 1$.

Alternatively, we can use the delay formula derived explicitly for the OpenFlow model, depicted in Fig. 1(b), as highlighted in [1].

To this end it needs to be highlighted that a packet arriving at the switch in the data plane of the OpenFlow network is confronted with two conditions. If there is already a flow entry installed in the switch then the packet is forwarded as such after spending time T_l otherwise it goes to the controller, spends time T_c then returns back to the same switch for packet matching where it spends time $T_l^{(2)}$ and is forwarded on the output interface.

So the absolute value of packet sojourn time W^{of} in an OF-based SDN network where node l interacts with the SDN controller c as shown in Fig. 1(b) is given as

$$W^{of} = \begin{cases} T_l & \text{with probability } 1 - q_l^{nf} \\ T_l + T_c + T_l^{(2)} & \text{with probability } q_l^{nf} \end{cases} \quad (6)$$

where T_l and T_c are sojourn times in node l and node c respectively, while $T_l^{(2)}$ is the sojourn time when a packet enters node l the second time after visiting the controller.

Eventually, the mean of W^{of} is given as

$$\begin{aligned} E[W^{of}] &= (1 - q_l^{nf}) E[T_l] + q_l^{nf} (E[T_l] + E[T_c] + E[T_l^{(2)}]) \\ &= (1 + q_l^{nf}) \frac{1}{\mu_l - \Gamma_l} + q_l^{nf} \frac{1}{\mu_c - \Gamma_c} \end{aligned} \quad (7)$$

We draft a short proof in Lemma 1 to show that the mean packet sojourn time calculated by the two methods is indeed the same.

Lemma 1: For the single node case the packet sojourn time calculated in (5) using the standard Jackson assumption is the same as explicitly calculated using (7).

Proof: By rearranging (7) in terms of traffic loads we have:

$$E[W^{of}] = \frac{1 + q_l^{nf}}{\Gamma_l} \left(\frac{1}{1 - \rho_l} \right) + \frac{q_l^{nf}}{\Gamma_c} \left(\frac{1}{1 - \rho_c} \right) \quad (8)$$

Using $\Gamma_l = (1 + q_l^{nf}) \lambda_l$ from (2) and $\Gamma_c = q_l^{nf} \lambda_l$ from (3) we obtain $E[W^{jack}]$, in (5) which proves the Lemma

3.2. Distribution of Time Spent by the Packet

In this section we take a step forward by presenting the probability density function (PDF) and the cumulative density function (CDF) of the time spent by a packet in the node.

Lemma 2: The PDF $w_l^c(t)$ and the CDF $\tilde{W}_l^c(t)$ of the time spent by a packet in the node l are given respectively as

$$w_i^c(t) = b_i^{(1)} a_1 e^{-a_1 t} + b_i^{(2)} a_1 (a_1 t) e^{-a_1 t} + d_1 a_c e^{-a_c t} \quad (9)$$

$$\tilde{W}_i^c(t) = P(W_i^{pf} > t) = (b_i^{(1)} + b_i^{(2)}) e^{-a_1 t} + b_i^{(2)} (a_1 t) e^{-a_1 t} + d_1 e^{-a_c t} \quad (10)$$

where

$$a_1 = \mu_1 - \Gamma_1, \quad a_c = \mu_c - \Gamma_c$$

while

$$b_i^{(1)} = 1 - q_1^{nf} - q_1^{nf} \frac{a_1 a_c}{(a_c - a_1)^2}, \quad b_i^{(2)} = q_1^{nf} \frac{a_c}{a_c - a_1}, \quad d_1 = q_1^{nf} \frac{a_1^2}{(a_c - a_1)^2}$$

Proof: If we assume that the sojourn times T_1 and $T_1^{(2)}$ are independent then the Laplace transform

$W_i^c(s) = E[e^{-sW_i^{pf}}]$ may be written as

$$W_i^c(s) = (1 - q_1^{nf}) \frac{a_1}{a_1 + s} + q_1^{nf} \left(\frac{a_1}{a_1 + s} \right)^2 \left(\frac{a_c}{a_c + s} \right) \quad (11)$$

which can further be written as

$$W_i^c(s) = b_i^{(1)} \frac{a_1}{a_1 + s} + b_i^{(2)} \left(\frac{a_1}{a_1 + s} \right)^2 + d_1 \left(\frac{a_c}{a_c + s} \right) \quad (12)$$

Inverting the Laplace transform proves the Lemma.

4. NUMERICAL RESULTS

In order to verify our proposed model we developed a discrete event simulation model to mimic the queuing behavior in an OF-based SDN. We assume that at the arrival to the node I , packets are queued in the data node before being processed. The processing time of data node I/μ_1 is considered to be exponentially distributed with a mean value of $9.8\mu s$. The value of 9.8 is the average processing time taken by Pronto 3290 switch for forwarding packets of size 1500 bytes[1]. We here assume that TCP uses maximum transmission unit (MTU) of 1500 bytes.

At the controller, the number of responses per second are taken to be 4175 as reported in [1] by using the Cbench tool[10]. Hence this is parameterized as $I/\mu_c = 240\mu s$ in the model.

To enhance confidence in the simulation result, five replications for each value were run and the normally distributed 95% confidence interval is incorporated in the plots.

We first highlight in Fig.3 that our proposed model provides a fix to the results reported by [1]. To this end we first plot the Simulation curve from [1] as a reference. On top of it we plot the analytical curves; Analytical[1]} and Modified Jackson Model, obtained using the model in [1] and our proposed Jackson model, respectively.

It can be seen that the model proposed by [1] performs very well for small loads ($q_1^{nf}=0.2$). However in the cases when there is a large amount of query traffic coming to the controller due to unknown flows the model in [1] falls short. In such cases the proposed modification to the Jackson model is quite accurate as seen for the extreme case of $q_1^{nf}=1.0$.

In Fig. 4, the effect of q_1^{nf} on network throughput is studied where the network throughput is defined as the amount of traffic λ_1 which can be injected into the OF-based SDN for a given delay guarantee. In this case the delay guarantee is the average packet sojourn time. This plot also highlights how the proposed model can be used to dimension the network if packet sojourn time is considered as the design parameter. A striking feature of this plot is that the network throughput saturates after reaching a certain value of packet sojourn time. Subsequently, it can be inferred that even if the network is over-loaded after crossing a certain traffic threshold, the result will be

just increased packet sojourn time without further enhancing the network throughput. Similarly, it is also observed that the critical value for packet sojourn time remains almost the same for all the values of q_1^{nf} but the resulting network throughput for each of them is quite different.

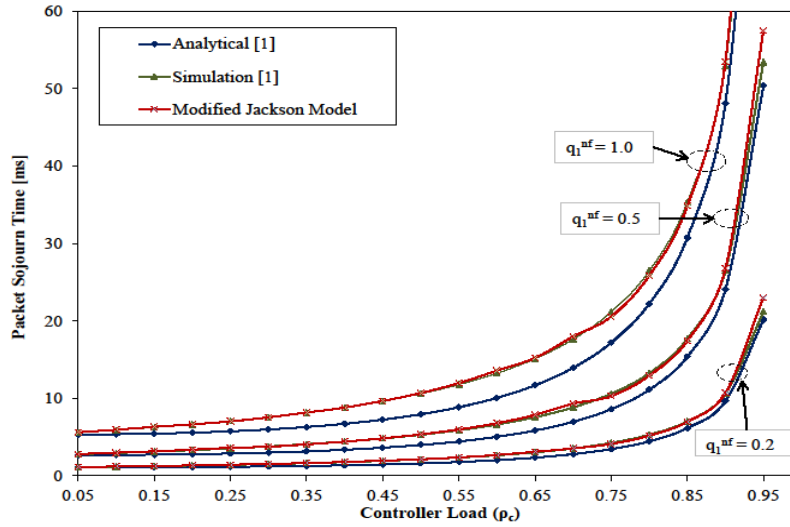


Figure 3: Comparing Modified Jackson Model to Results from [1]

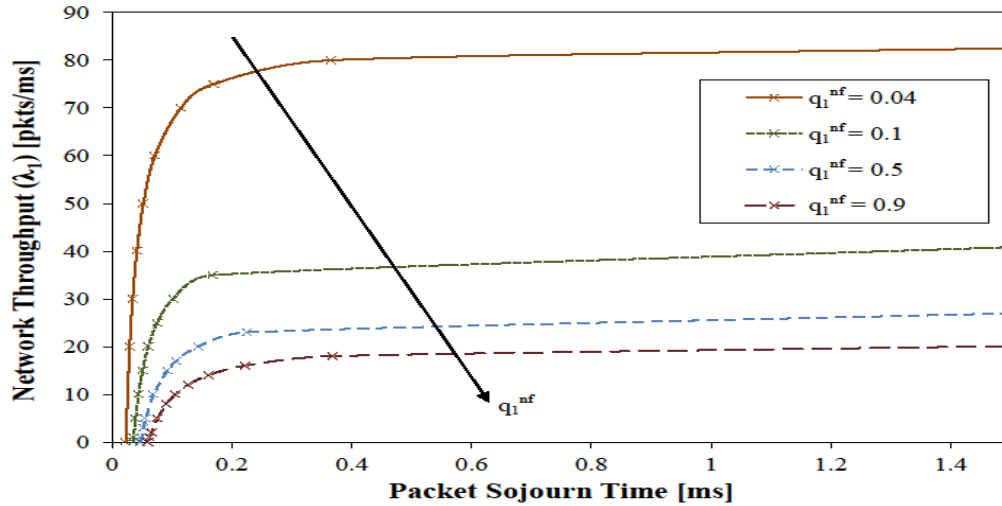


Figure 4: Dimensioning Network Throughput

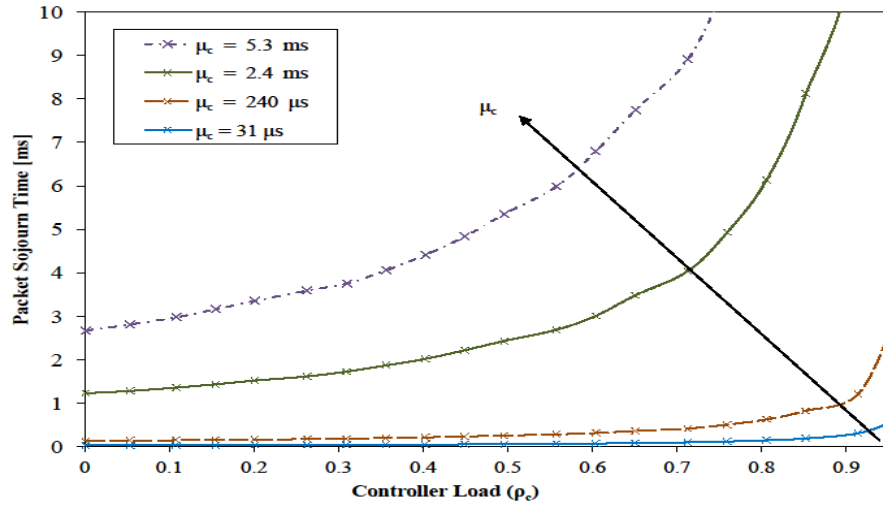


Figure 5: Effect of the controller service rate μ_c

In Fig. 5 a fundamental performance plot is shown in which packet sojourn time is plotted against the controller load ρ_c for differing values of the controller service time μ_c with q_1^{nf} constant at 0.5. Although the plot is mainly for evaluating performance, it can also play a role in designing a network with controller of known average service time and giving guarantees on packet sojourn time by keeping the controller load at a certain level.

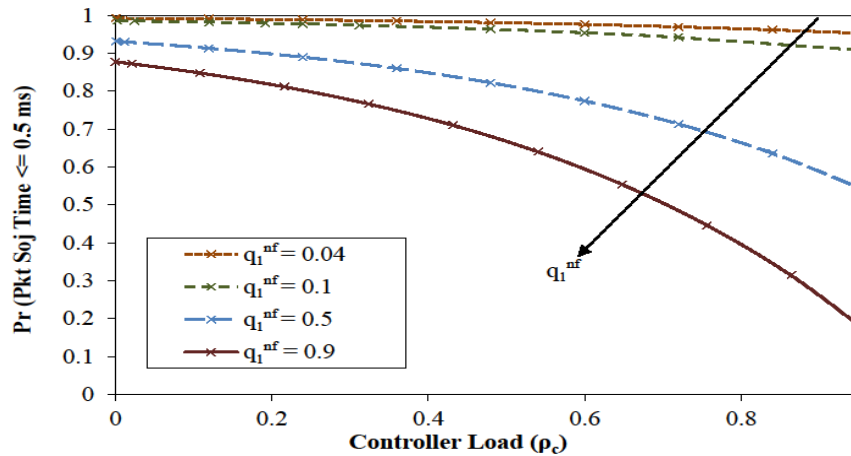


Figure 6: CCDF of packet sojourn time as a function of load on controller

In Fig. 6, the plot shows the probability that packet sojourn time ≤ 0.5 ms for varying values of controller load ρ_c and for different q_1^{nf} values. The plot can be used to determine the maximum load that the controller should reach before its performance is compromised. The plot in Fig. 6 is pilot and similar plots for different values of packet sojourn time can be obtained depending upon the requirements.

It needs to be emphasized that blocking probability p_b was zero for the setup which we had for the simulation and infinitesimal small for the analytical model.

5. THE MULTI-NODE CASE

In a real life SDN deployment, an SDN controller is responsible for more than one node in the data plane. In this section we highlight how the proposed model can be used to model this scenario. To this end we take a toy example in which we only have two nodes in the data plane as shown in Fig. 8. We define q_2^{jack} and q_1^{jack} for node 2 similar to q_1^{jack} and q_1^{nf} defined earlier for node 1.

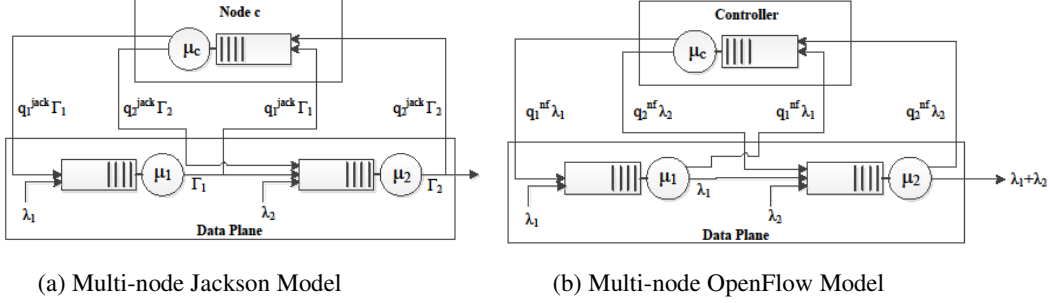


Figure 7: Modeling more than one node in the data plane

In order to leverage the Jackson model in Fig.8(a) for modeling the OF-based SDNs in Fig.8(b), the probabilities q_2^{jack} and q_1^{jack} need to be adjusted. This is accomplished by forcing the rates at all the nodes in both the models to be the same as

$$\text{For node 1: } \lambda_1(1+q_1^{nf}) = \Gamma_1, \quad q_1^{nf}\lambda_1 = q_1^{jack}\Gamma_1 \quad (17)$$

$$\text{For node 2: } \lambda_1+\lambda_2(1+q_2^{nf}) = \Gamma_2, \quad q_2^{nf}\lambda_2 = q_2^{jack}\Gamma_2 \quad (18)$$

Solving (17) we get q_1^{jack} same as (4) while by solving (18) we have q_1^{jack} as

$$q_2^{jack} = \frac{q_2^{nf}}{1 + \frac{\lambda_2}{\lambda_1} (1 + q_2^{nf})} \quad (19)$$

We can then use the modified q_1^{jack} and q_2^{jack} to derive the appropriate performance metrics such as packet sojourn time using existing queuing theory results[11] similar to the single node case.

6. CONCLUSION

In this work we have proposed an analytical model for an OpenFlow enabled SDN based on Jackson network. We have shown that the model is accurate even for the case when the probability of new flows is quite large. The applicability of the model is determined by establishing two performance measures, the average packet sojourn time and the distribution of time spent by a packet in the network, by using the proposed model. Secondly we showed by a toy example that the model can be readily extended to more than one switch in the data plane. Conclusively it is noted, and can be safely stated, that the model proposed in this paper caters for realistic OpenFlow-based SDNs and this argument has readily been validated in this paper.

Furthermore, the effects of key parameters in an SDN network are studied which include the time required by the controller to process a request, amount of traffic going to the controller, average time spent by a packet in a network and the network throughput.

There is more than one direction that the work presented in this paper can be taken forth. First of all, the work presented and validated for a single node can be extended to larger and more realistic topological scenarios, such as fat-tree topology. Secondly, the model in this work is

based on Markovian arrival and service processes which can be generalized and more realistic distributions or traces can be used in modeling. This can be supplemented with simulations for validation and verification of the model. Also, a test-bed study for verifying our model can be performed which will enhance the confidence in the proposed model.

REFERENCES

- [1] Jarschel, Michael & Oechsner, Simon & Schlosser, Daniel & Pries, Rastin & Goll, Sebastian & Tran-Gia, Phuoc (2011) "Modeling and Performance Evaluation of an OpenFlow Architecture", Proceedings of the 23rd International Teletraffic Congress (ITC '11), pp1-7.
- [2] Hoelzle, Urs (2012) "Opening Address: 2012 Open Network Summit", [Online] Available: <http://www.opennetsummit.org/archives/apr12/hoelzle-tue-openflow.pdf>, Date Retrieved: 08/08/2014.
- [3] ONF, Open Networking Foundation, [Online] Available: <https://www.opennetworking.org>.
- [4] Bozakov, Zdravko & Rizk, Amr (2013) "{Taming SDN Controllers in Heterogeneous Hardware Environments", 2nd IEEE European Workshop on Software-Defined Networks (EWSDN).
- [5] Azodolmolky, Siamak & Nejabati, Reza & Pazouki, Maryam & Simeonidou, Dimitra (2013) "An Analytical Model for Software Defined Networking: A Network Calculus-based Approach", IEEE Globecom.
- [6] Ciucu, Florin & Schmitt, Jens (2012) "Perspectives on Network Calculus: No Free Lunch, but Still Good Value", SIGCOMM Computer Communication Reviews, vol. 42, no. 4, pp 311-322.
- [7] Naous, Jad & Erickson, David & Covington, G. Adam & Appenzeller, Guido & McKeown, Nick (2008) "Implementing an OpenFlow Switch on the NetFPGA Platform", Proceedings of the 4th ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS '08), pp 1-9.
- [8] Bianco, Andrea & Birke, Robert & Giraudo, Luca & Palacin, Manuel (2010) "OpenFlow Switching: Data Plane Performance", IEEE International Conference on Communications (ICC) 2010, pp 1-5.
- [9] Khan, Asif & Dave, Nirav (2013) "Enabling Hardware Exploration in Software-Defined Networking: A Flexible, Portable OpenFlow Switch", IEEE 21st Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), 2013.
- [10] Sherwood, Rob "Cbench (Controller benchmarker)", [Online] Available: <http://archive.openflow.org/wk/index.php/Oflops>, Date Retrieved: 13/03/2014.
- [11] Jackson, James R. (1957) "Networks of Waiting Lines", Operations Research, vol. 5, no. 4, pp518-521.

INTENTIONAL BLANK

ENHANCING AN ATL TRANSFORMATION WITH TRACEABILITY

Laura Felice, Marcela Ridao, Maria Carmen Leonardi and
Maria Virginia Mauco

INTIA, Departamento de Computación y Sistemas
Universidad Nacional del Centro de la Provincia de Buenos Aires
Tandil, Argentina

felice@exa.unicen.edu.ar, mridao@exa.unicen.edu.ar,
cleonard@exa.unicen.edu.ar, vmauco@exa.unicen.edu.ar

ABSTRACT

Model transformation is widely recognized as a key issue in model engineering approaches. In previous work, we have developed an ATL transformation that implements a strategy to obtain a set of Raise Specification Language (RSL) modules from Feature Models (FM). In this paper, we present an improvement to this strategy by defining another complementary and independent model, allowing the incorporation of traceability information to the original transformation. The proposed mechanism allows capturing and representing the relationships created by the application of the transformation rules.

KEYWORDS

Raise Specification Language, Feature Models, Traceability, MDD, ATL.

1. INTRODUCTION

As formal methods offer a wide spectrum of possible paths towards designing high-quality software, in the academia and the industry have adopted them as an alternative of development, especially where safety or security is important [1]. By using formal methods early in the software development process, ambiguities, incompleteness, inconsistencies, errors, or misunderstandings can be detected, avoiding their discovery during costly testing and debugging phases.

However, formal specifications are unfamiliar to stakeholders, whose active participation is crucial in the first stages of software development process to understand and communicate the problem. This holds in Domain Analysis (DA), because its first stage is to capture the knowledge of a particular domain, making necessary to have a model that is comprehensible by software engineers and domain experts. To contribute to bridge the gap between DA and formal specifications, we have been working in the integration of domain analysis phase into the RAISE Formal Method [2]. The main purpose is to specify a family of systems to produce qualitative and reliable applications in a domain, promoting early reuse and reducing development costs. Feature models were used to represent domain analysis because they facilitate the customization of software requirements.

The integration between the models implies the definition of rules to derive an initial hierarchy of RSL types from a FM. We use the structural information of the FM to derive RSL (Raise Specification Language) [3] constructs following one of the several proposals that facilitate the construction of FM: the Feature-Oriented Reuse Method (FORM) [4]. In order to fit the main proposal of enhancement of formal developments with the RAISE Method into Model Driven Development (MDD) paradigm [5], an ATL (Atlas Transformation Language) [6] transformation has been developed. This transformation allows the automatic derivation of a first abstract RSL specification of a domain starting from a FM. The ATL rules define how features and relationships between them (the source model) are matched and navigated to produce the RSL specification (the target model) [7]. The rules follow closely the principles proposed in the RAISE Method, so this first and still incomplete specification may be later developed into a more concrete one following the RAISE Method steps. The overall strategy is explained in [8]. In this work, we improve this transformation by providing a simple trace mechanism that creates a trace relationship between the elements of the source and target metamodels.

The paper is organized as follows: In Section 2 we introduce the derivation process. The core of the paper is in section 3, where we describe the ATL transformation that obtain RSL schemes from FM with the incorporation of the trace mechanism, and exemplify it with the case study in section 4. Finally, in Section 5 we present some conclusions and outline possible future work.

2. AN OVERVIEW OF THE DERIVATION PROCESS

We have been working in the integration of DA models and formal specifications, giving a strategy to derive a set of RSL schemes from FM. The strategy begins with the analysis of features into the model to arrive to RSL schemes. Then, different constructions of the FM are analyzed in order to complete the structure of the schemes. Relationships between schemes are modelled from the feature information. The result of these steps is a set of schemes that serves as a basis for the RSL scheme hierarchy, reducing the gap between analysis and specification phases. The full derivation process may be found in [9]. In [8] we presented the rules to define schemes in an automatic way. These rules are a simplification of the derivation process with the objective of defining an automatic transformation aligned with Model Driven Architecture (MDA) framework.

The transformation rules are defined by the following mappings:

- A Feature Model is mapped into one or more RSL modules hierarchies.
- A Feature Diagram (FD) is mapped into a RSL modules hierarchy.
- A Concept Feature (root feature) is mapped into a RSL class expression with a type of interest.
- Features that are not concept features are mapped into RSL modules.
- Relationships
 - Grouping is mapped into RSL schemes relations.
 - Dependencies between features are mapped into RSL axioms expressing the corresponding restriction.

Rule 1: Mapping Feature Model

Description: this transformation declares that a FM will be mapped into one or more RSL modules hierarchy. Each hierarchy will be derived from a FD.

Transformation:

- a FM is transformed into one or more RSL schemes

Rule 2: Mapping Feature Diagram

Description: this transformation declares that an initial hierarchy structure of RSL modules will be derived from a FM diagram. The hierarchy could have modifications later when the evolution of the specification is made.

Transformation:

- each feature of the FD will be mapped into schemes according to the following rules.

Rule 3: Concept2Scheme (mapping Concept feature)

Description: this transformation declares that the most abstract RSL module of the hierarchy will be derived from the feature concept.

Transformation:

- each concept in the FM will be a RSL scheme with the same name
- the RSL scheme will have a type of interest whose name represents the system that is being modeled.

Rule 4: Fea2SchSpec (mapping Mandatory feature)

Description: this transformation declares that the schemes in a hierarchy will be derived from the mandatory features.

Transformation:

- a mandatory feature is transformed into a scheme with the same name as the feature and two clauses:
 - the **value** clause with the following properties:
 - isSolitary with TRUE | FALSE value
 - isMandatory with TRUE value
 - isSelected with TRUE value
 - the **extend** clause with the parent scheme name.

Rule 5: Fea2SchSpec (mapping Optional feature)

Description: this transformation declares that the schemes in a hierarchy will be derived from the optional features selected in a configuration time.

Transformation:

- an optional feature is transformed into a scheme with the same name as the feature and two clauses:

- the **value** clause with the following properties:
 - o isSolitary with TRUE | FALSE value
 - o isMandatory with FALSE value
 - o isSelected with TRUE value
- the **extend** clause with the parent scheme name.

Rule 6: Fea2SchSpec (mapping Parameterized feature)

Description: this transformation declares that a parameterized scheme will be derived from a parameterized feature.

Transformation:

- a parameterized feature is transformed into a parameterized scheme with the same name as the feature and the following clauses:
 - the **context** clause with the list of parameters
 - the **object** clause where the type of the parameters is declared
 - the **value** clause with the following properties:
 - o isSolitary with TRUE | FALSE value
 - o isMandatory with TRUE | FALSE value
 - o isSelected with TRUE value

Rule 7: Aggr2Sch (mapping *Aggregate* relationship)

Description: the statement may express both a composition relationship as well as an aggregate relationship. Thus, a RSL specification will have one or more axioms expressing post-conditions ensuring the relation whole/parts.

Transformation:

- if the feature part is atomic and express simple types, it will be transformed into a component in the scheme corresponding to the parent feature. The component is a **record** into the RSL expression. It has the same treatment that the mandatory, optional or parameterized feature.
- If the feature part is not atomic, the part will be transformed into a RSL expression according to the type of the feature that is being modeled, expressed as an embedded object of the expression that contains it.

Rule 8: GroupOR2Sch (mapping Group OR)

Description: this type of grouping is considered like a 'is-a' relationship among alternative, mandatory or optional features with a parent feature. Thus a RSL hierarchy of modules will be derived from the structure of the set of features.

Transformation:

- the parent feature is modeled as a RSL scheme with at the least one abstract operation. This scheme will define:
 - the **type** clause with the lower and upper values expressing the group cardinality

- the **value** clause with the restriction expressions of the grouping.
- in later refinement of schemes, the consistence restriction will be mapped into a boolean function.

Rule 9: GroupXOR2Sch (mapping Group XOR)

Description: this type of grouping has the same treatment that the OR group.

Transformation:

- the parent feature is modeled as a RSL scheme with at the least one abstract operation. This scheme will define:
 - the **type** clause with the lower and upper values expressing the group cardinality
 - the **value** clause with the restriction expressions of the grouping.
 - in later refinements of schemes, the consistence restriction will be mapped into a boolean function.

Rule 10: FeaReq2SchSpec (Mapping Requires)

Description: this restriction will generate two RSL schemes derived from two features related by the requires relationship.

Transformation:

- the supplier feature is mapped to a RSL scheme with the same name as the supplier feature
- the requester feature is mapped to a RSL scheme with the same name as the requester feature
- an axiom that defines the implication in the requester scheme expressing the requires restriction.

Rule 11: FeaExc2SchSpec (Mapping Excludes)

Description: this restriction will generate a RSL scheme derived from the relations between two features by the exclude relationship.

Transformation:

- the supplier feature is mapped to a RSL scheme with the same name as the supplier feature
- an axiom that defines the implication in the requester scheme expressing the excludes restriction.

3. INCORPORATION OF TRACEABILITY INTO THE ATL TRANSFORMATION

This section describes the ATL transformation that obtains RSL schemes from FM with the incorporation of the trace mechanism. The complete description of the transformation rules may be found in [9]. This ATL definition represents the rules described before and it is a single module involving several ATL Rules (both matched and lazy rules) along with a set of helpers. In order to define the transformation and execute it, we must define the source and target models as

metamodel according to each transformation rule. For example, each RSL scheme of the module hierarchy is related with one feature because the scheme was originated from one of them. Each of the relationships has their own semantic, and there may be more than one relationship between those components, depending on the rules.

Table 1 shows each trace relationship between elements of the source (FM) and target (RSL) metamodels: the left column represents the source elements and the top row, the target elements. Cells with data indicate a trace relationship. Trace shows the relationships that give rise to new elements in the target metamodel from elements in the source metamodel, i.e. forward relationships, but from them backward traceability ones may be also obtained. The trace relationships are originated from the application a particular transformation rule.

For each trace relationship the following item are described in Table 1:

- Cardinality of source: how many elements were used to create the new element (Table 1, in the left side of the parenthesis),
- Cardinality of target: how many elements are created in that relationship (Table 1, in the right side of the parenthesis),
- Name of the rule that originated the trace relationship.

Table 1. Trace Relationship between FM and RSL generated by the application of the rules

Target Source	SchemeSpecification	SchemeTypes	ObjectDeclar ation	SchemeValue
Concept	(1/1)Rule3:Concept2 Scheme	(1/1)Rule3: Concept2Sche me		
Feature mandatory	(1/1)Rule 4:Fea2SchSpec	(1/1)Rule4: Fea2SchSpec		(1/n)Rule4:Fea2Sch Spec
Feature optional	(1/1)Rule 5:Fea2SchSpec	(1/1)Rule 5: Fea2SchSpec		(1/n)Rule4: Fea2SchSpec
Feature parameterized	(1/1) Rule 6: Fea2SchSpec	(1/1)Rule 6: Fea2SchSpec		(1/n)Rule4: Fea2SchSpec
Aggregate	(1/n) Rule 7: Aggr2Sch		(1/n) Rule7: Aggr2Sch	
GroupORAssociati on	(1/n) Rule8:GroupOR2Sch			
GroupXORAssocia tion	(1/n) Rule9:GroupXOR2S ch			
Requires relation	(2/1) Rule10:FeaReq2Sch Spec			
Excludes relation	(2/1) Rule 11: FeaExc2SchSpec			

3.3. An example of a rule application

In order to exemplify part of the derivation strategy, we describe the ATL transformation corresponding to the Aggregate association (Figure 3) present in a FM Model to RSL schemes

explaining each of the defined rules and helpers. The Transformation Process contains a matched primary rule that guide the overall process of this transformation, the rule `aggr2Sch`. This rule allows the matching of all features from the FM and defines a RSL scheme for each of them. For each association, the rules identify the name of the association and collect all of the features clustered under this association. These features will be expressed under the objects declaration into a scheme definition. The helper `getparts` returns the features that are the parts of the aggregate. The lazy rule `Feature2Obj` has the input features that are part of the aggregate (this condition is implemented by the helper `isPartOfAggregate`), and returns the names of these features. The helper `isPartOfAggregate` verifies the type of association will be an aggregation.

```

module fm2rsl;
create OUT: rsl, trace: Trace from IN: FM;
.....

rule aggr2Sch{
from
  A:FM!AggregateAssociation
to
  S:rsl!SchemeSpecification (
    name <-A.name,
    object <- A.getparts()->collect (feature |
thisModule.Feature2Obj (feature))
  ),
  TL1: Trace!traceLink (
    ruleName <-'aggr2Scheme',
    targetElements <-S)
do {
  TL1.refSetValue('sourceElements', A);
}
}

```

Figure 3. AggregateAssociation Rule

3.4 Implementing the Traceability Mechanism

In order to implement traceability in our transformation process, we adopt the proposal of Jouault [10]. In this work a trace mechanism is defined by considering traceability information as a separate model (Figure 4), and the code to generate trace relationship is added directly to the transformation rules. It is a simple mechanism that supports any form of traceability, and it is used in other transformations.

Following this idea, we have defined our trace metamodel, and added the corresponding code to the rules in order to define the trace relationship presented in Table 1. All the matched and lazy rules are modified in order to define the trace information.

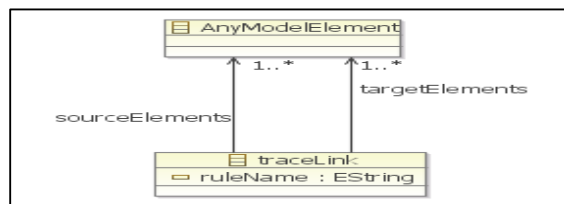


Figure 4: Trace Metamodel

Table 2 shows the implementation of traceability for each transformation shown in Table 1. For our example, during the transformation of a feature to a RSL scheme, trace information is generated for `Feature2Object`, `Feature2ValueIsMandatory`, `Feature2ValueIsSolitary`, `Feature2ValueIsSelected` lazy rules, besides `Aggr2Scheme` rule.

Table 2. Implementation of Trace Relationship

Source \ Target	SchemeSpecification	SchemeTypes	ObjectDeclaration	SchemeValue
Concept	ATL rule: Concept2Scheme	ATL lazy rule: Concept2SType		
Feature mandatory	ATL rule: Fea2Scheme	ATL lazy rule: Feature2Type		ATL Lazy rule: Feature2ValuesMandatory Feature2ValuesSolitary Feature2ValuesSelected
Feature optional	ATL rule: Fea2Scheme	ATL lazy rule: Feature2Type		ATL Lazy rule: Feature2ValuesMandatory Feature2ValuesSolitary Feature2ValuesSelected
Feature parameterized	ATL rule: Fea2Scheme	ATL lazy rule: Feature2Type		ATL Lazy rule: Feature2ValuesMandatory Feature2ValuesSolitary Feature2ValuesSelected
Aggregate	ATL rule: Aggr2Scheme		ATL Lazy rule: Feature2Object	ATL Lazy rule: Feature2ValuesMandatory Feature2ValuesSolitary Feature2ValuesSelected
GroupOR Association	ATL rule: group2Scheme	ATL lazy rule: Feature2Type		ATL Lazy rule: Feature2ValuesMandatory Feature2ValuesSolitary Feature2ValuesSelected
GroupXOR Association	ATL rule: group2Scheme	ATL lazy rule: Feature2Type		ATL Lazy rule: Feature2ValuesMandatory Feature2ValuesSolitary Feature2ValuesSelected
Requires relation	ATL rule: Fea2Scheme	ATL lazy rule: Feature2Type		ATL Lazy rule: Feature2ValuesMandatory Feature2ValuesSolitary Feature2ValuesSelected
Excludes relation	ATL rule: Fea2Scheme	ATL lazy rule: Feature2Type		ATL Lazy rule: Feature2ValuesMandatory Feature2ValuesSolitary Feature2ValuesSelected

4. CASE STUDY

We applied the ATL transformation described in this paper to the case study “e-Shop” [12], as it is a well known one. Figures 4 and 5 show some FM features taken from the FM Model of the complete case study, which are necessary to exemplify how the ATL transformation works. For example, we consider the Eshop-Aggregate Association (Figure 4) for the Aggregate Association StoreFront. The features were defined in a XMI format to be used as source in the derivation of the RSL schemes.

Figure 5 shows the XMI definition for the Trace-Eshop-AggregateAssociation, using the Sample Reflective Ecore Model. This Figure presents an extract of the ADITIONAL trace information produced after the ATL transformation. Each trace link includes a reference to an element in the source model (Feature Model), and another one to an element in the target (RAISE Model). For example, trace link with rule name aggr2Scheme produce the trace link for Feature2Obj for the three object declaration (Catalog, BuyPaths and CustomerService) and the trace link for StoreFront.

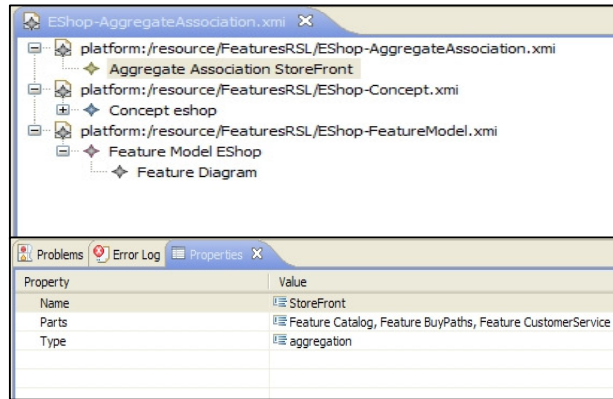


Figure 5. Sample Reflective Ecore Model for Eshop-Aggregate Association

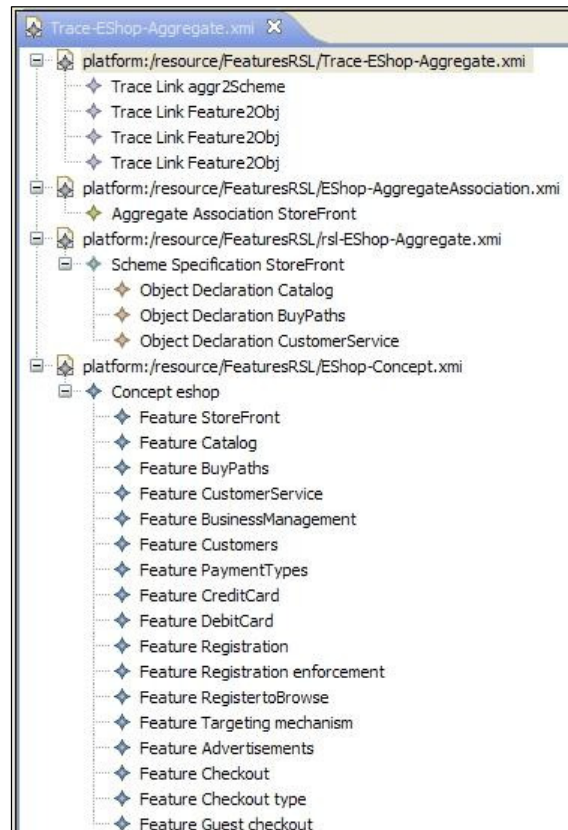


Figure 6. Trace-Eshop-AggregateAssociation

5. CONCLUSIONS AND FUTURE WORK

Traceability plays a crucial role in MDD. The transformation process we have proposed and implemented in the ATL rules allows making a trace between the source and the target models. This concept is quite simple: to follow relationships or links. It is essential for software development because a lot of information is used and produced and it should be kept traceable.

A domain component (source model) is traced forward, for example, when the component is changed and we want to investigate the impact of the change. A scheme is traced backward, for example, when there is a change and we want to understand it, investigating the information used to derive this scheme. Also, we may want to know how a RSL scheme is related with others in a hierarchy, for example, if the derived scheme has a parent or it has some restrictions among features.

In this work, traceability information is easily created but, until the moment it is not managed. There are several difficulties related to the implementation and use of traceability [13]. Wieringa [14] points out that the major problems of realising traceability are organisational, not technical. As future work, we must incorporate trace supporting in order to recorded traces became useful for the entire development process. The traceability generating code is easily added to the ATL code, and can be automated

To help those who are responsible for the specification phase, it is helpful if the work team has traceability policies to traceability information be maintained. Maintaining traceability information is tedious, time-consuming a labour- intensive. Therefore, the policies may be fine, if they cannot be implemented, they are useless [11]. Although in this paper we focused on the transformation rules and the traceability, we know that a lot of information is used and produced and it should be kept related. So, we believe that this approach could be adapted to approach for visualizing traceability in (compositions of) model transformations. The main purpose will be to study the effects of the evolution of a source model, or changes in the transformation model.

REFERENCES

- [1] Streitferdt, D., Riebisch, M. & Philippow, I. (2003) "Formal Details of Relations in Feature Models". Proceedings 10th IEEE Symposium and Workshops on Engineering of Computer-Based Systems, pp: 297-304.
- [2] George, C; Haxthausen, A; Hughes, S; Milne, R; Prehn, S & Pedersen, JS. (1995) The RAISE Development Method. BCS Practitioner Series, Prentice Hall.
- [3] George, C; Haff, P; Havelund, K; Haxthausen, A; Milne, R; Nielsen, CB; Prehn, S. & Wagner, K.R. (1992) The RAISE Specification Language. Prentice Hall.
- [4] Kang, K; Kim, S; Lee, J; Kim, K; E. Shin, E; M. & Huh, M. (1998) "FORM: A feature-oriented reuse method with domain-specific reference architectures". Annals of Software Engineering 5, pp 143-168.
- [5] Mellor, S., Clark, A. & Futagami, T. (2003) "Model-driven Development". IEEE Software Vol. 20, N°5.
- [6] ATL Transformation Language. Available in: <http://www.eclipse.org/atl/>.
- [7] Felice, Laura; Ridaio, Marcela; Mauco, María Virginia & Leonardi, María Carmen. (2011) "Using ATL Transformations to Derive RSL Specifications from Feature Models", Proceedings of the 2011 International Conference on Software Engineering Research & Practice, Volume I, USA, pp: 273 – 279.
- [8] Felice, Laura; Ridaio, Marcela; Mauco, María Virginia & Leonardi, María Carmen. (2014) "Enhancing Formal Methods with Feature Models in MDD". Encyclopedia of Information Science and Technology, Third Edition. Ed. Mehdi Khosrow-Pour. pp:170-183.
- [9] Felice, Laura. (2013) Integration of Domain Analysis Techniques with RSL Specifications. Master Thesis. Facultad de Informática, Universidad Nacional de La Plata, Argentina (<http://sedici.unlp.edu.ar/>)
- [10] Jouault, F. & Kurtev, I. (2005) "Transforming Models with ATL". Proceedings of the Model Transformation in Practice Workshop. MoDELS 2005 Conference.
- [11] Kotonya, G & Sommerville, I. (2010) Requirements Engineering. Processes and Techniques. John Wiley & Sons.
- [12] Mendonca, M; Branco, M & Cowan, D. (2009) "S.P.L.O.T Software Product Lines Online Tools". In Companion to the 24th ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications. OOSPLA 2009. Florida, USA.

- [13] Pinheiro, F. (2004). "Requirement Traceability". In Perspectives on Software Requirements. Ed. Julio C.S.do Prado Leite, J. Doorn.
- [14] Wieringa R.J. (1998). Traceability and Modularity in Software Design. Proceeding of 9th International Workshop in Software Specification and Design. Japan.

Authors

Laura Felice is a computer science assistant professor at the Universidad Nacional del Centro de la Provincia de Buenos Aires from Argentina. She is a member of the Computer Science department. She has a Master's degree in Software Engineering from Universidad Nacional de La Plata, Argentina. Her main research interests include Software development methodologies, domain engineering and model-driven development. She has been member of the program committee of national and international conferences related to software engineering.

Contact her at:

Computer Science Department, UNCPBA,
Campus Universitario. (7000) Tandil. Buenos Aires. Argentina.
e-mail: lfelice@exa.unicen.edu.ar



Marcela Ridaio is a System Engineer and she is a computer science assistant professor at the Universidad Nacional del Centro de la Provincia de Buenos Aires, in Argentina. She has a master's degree in Software Engineering from La Plata University, and she wrote her master's thesis on Patterns used in Scenario Construction Process. Her research interests include Requirements Engineering, Compilers and real time problems. Currently, she is a doctoral student at La Plata University. She's writing her doctoral dissertation on Quantitative Techniques Oriented to Semantic Reuse in Requirements Models.

Contact her at:

Computer Science Department, UNCPBA,
Campus Universitario. (7000) Tandil. Buenos Aires. Argentina.
e-mail: mridao@exa.unicen.edu.ar



María Carmen Leonardi is a computer science assistant professor in Universidad Nacional del Centro de la Provincia de Buenos Aires from Argentina. She is member of the Computer Science Department. She has a Master's degree in Software Engineering from Universidad Nacional de La Plata, Argentina. Her main research interests include software development methodologies, requirements engineering, and model-driven development. She has been member of the program committee of national and international conferences related to software engineering.

Contact her at:

Computer Science Department, UNCPBA,
Campus Universitario. (7000) Tandil. Buenos Aires. Argentina.
e-mail: cleonard@exa.unicen.edu.ar



María Virginia Mauco is a computer science assistant professor in Universidad Nacional del Centro de la Provincia de Buenos Aires from Argentina. She is member of the Computer Science Department. She has a Master's degree in Software Engineering from Universidad Nacional de La Plata, Argentina. Her main research interests include software development methodologies, requirements engineering, and formal methods. She has been member of the program committee of national and international conferences related to software engineering.

Contact her at:

Computer Science Department, UNCPBA,
Campus Universitario. (7000) Tandil. Buenos Aires. Argentina.
e-mail: vmauco@exa.unicen.edu.ar



SOLUTION OF UNSTEADY ROLLING MOTION OF SPHERES EQUATION IN INCLINED TUBES FILLED WITH INCOMPRESSIBLE NEWTONIAN FLUIDS BY DIFFERENTIAL TRANSFORMATION METHOD

Y. Rostamiyan¹, S.D.Farahani², M.R.Davoodabadi³

¹Departments of Mechanical Engineering, Sari branch,
Islamic azad university, Sari, Iran

²Departments of Mechanical Engineering, Tehran University, Tehran, Iran

³Department of Mechanical Engineering, Semnan University, Semnan, Iran

ABSTRACT

In this paper, the unsteady motion of a spherical particle rolling down an inclined tube in a Newtonian fluid for a range of Reynolds numbers was solved using a simulation method called the Differential Transformation Method (DTM). The concept of differential transformation is briefly introduced, and then we employed it to derive solution of nonlinear equation. The obtained results for displacement, velocity and acceleration of the motion from DTM are compared with those from numerical solution to verify the accuracy of the proposed method. The effects of particle diameter (size), continues phase viscosity and inclination angles was studied. As an important result it was found that the inclination angle does not affect the acceleration duration. The results reveal that the Differential Transformation Method can achieve suitable results in predicting the solution of such problems.

KEYWORDS

Spherical particle; Acceleration motion; Inclination angle; Non-linear equation, Differential Transformation Method (DTM).

1. INTRODUCTION

The description of the motion of immersed bodies in fluids is present in several manufacturing processes, e.g. sediment transport and deposition in pipe lines, alluvial channels, chemical engineering and powder process [1-6]. Several works could be found in technical literature which investigated the spherical particles in low and high concentration [7-9]. A particle falling or rolling down a plane in a fluid under the influence of gravity will accelerate until the gravitational force is balanced by the resistance forces that include buoyancy and drag. The constant velocity reached at that stage is called the “terminal velocity” or “settling velocity”. Knowledge of the terminal velocity of solids falling in liquids is required in many industrial applications. Typical examples include hydraulic transport slurry systems for coal and ore transportation, thickeners,

*y.rostamiyan@yahoo.com

mineral processing, solid–liquid mixing, fluidization equipment, drilling for oil and gas, geothermal drilling. The resistive drag force depends upon drag coefficient. Drag coefficient and terminal velocities of particles are most important design parameters in engineering applications. There have been several attempts to relate the drag coefficient to the Reynolds number. The most comprehensive equation set for predicting C_D from Re for Newtonian fluids has been published by Clift et al. [10], Khan and Richardson [11], Chhabra [12] and Hartman and Yutes [13]. Comparing between most of these relationships for spheres, demonstrates quite low deviations [14].

The most of mentioned applications involve the description of the particle position, velocity and acceleration during time e.g. classification, centrifugal and gravity collection or separation, where it is often necessary to determine the trajectories of particle accelerating in a fluid for proposes of design or improved operation [15,16]. Unfortunately, there are few studies in the literature in the filled of rolling particles and the major part of the available investigations are related to the use of a rolling ball viscometer to measure viscosity of liquids [5,6]. Hasan [7] studied the role of wall effect on the rolling velocity of spherical particles in Newtonian media. He found a very limited correlation for $(d/D) > 0.707$ as follows:

$$C_D = \frac{15.717}{\text{Re}} \left(1 - \left(\frac{d}{D} \right) \right)^{-2.5} \quad (1)$$

Where, Reynolds number is defined as follow:

$$\text{Re} = \frac{\rho \cdot u \cdot D}{\mu} \quad (2)$$

Where Re, d, and D, are the particle Reynolds number, particle diameter and tube inner diameter respectively. Chhabra et al. [12] presented a valuable experimental work for drag on spheres in rolling motion in inclined smooth tubes. They used an enough number of sphere made of glass and steel with four smooth walled glass tubes of different diameter. They used numerous aqueous solutions of glycerol and glucose syrup to cover a wide range of Reynolds number. The angles of inclination and sphere-to-tube diameter ratios were varied from 3 to 30 and 0.114 to 0.58, respectively. Therefore, the Reynolds number range was $10^{-6} < \text{Re} < 3000$. They had 900 data points to define their empirical correlations. It was concluded that the sphere-to-wall diameter ratio (d/D) , is not a significant parameter at the 95% confidence level. Consequently, the authors presented a three-part equation as follows:

$$C_D = \frac{225}{\text{Re}}, \text{Re} < 1 \quad (3)$$

$$C_D = 1 + \frac{235}{\text{Re}}, 1 \leq \text{Re} \leq 250 \quad (4)$$

$$C_D = 1.35 + \frac{177.5}{\text{Re}}, 250 \leq \text{Re} \quad (5)$$

Eqs. (3-5) predict the transition points within about 5%, and correlates the experimental set of data within an average error of 8%. To describe a general correlation covering the experimental data we describe a new correlation using Chhabra et al. experimental points [12] as:

$$C_D = 1.2 + \frac{190}{Re} + \frac{1.003 \times 10^{-7}}{Re^2}, 10^{-6} \leq Re \leq 3000 \quad (6)$$

The third term in the right hand side of the Eq. (6) is important in low Reynolds number and its effect vanishes by increasing the Reynolds number and reduction of the drag coefficient. Eq. (6) is in very good agreement with results of the reference [12] and correlates the data with average error of 8.4883%. The maximum difference between values of Eqs. (5) and (6), and experimental data, is related to the transient region where the Reynolds number is in the range of $5.8 \times 10^{-2} \leq Re \leq 4.5$. Aside from mentioned work of Chhabra et al. [12], all other surveys of the rolling motion of the particles are related to open channels [12–15]. In reality, when a sphere is rolling in a tube, the wall exerts an extra retardation effect on it due to upward motion of the fluid through the eccentric annular gap between the particle and the wall this issue distinguishes the mechanism of rolling in tubes from open channels. It is clear that a few studies are performed on rolling motion of particles, especially in tubes while it is an important practical issue both in nature and industry. Moreover, most of the previous studies in particles motion and sedimentations are experimental or numerical. However, an exact analytical expression is more opportune for engineering calculations, and is also the evident starting point for a better understanding of the relationship between the physical properties of the sphere-fluid combination and the accelerated motion of the sphere. In addition, in contrast to steady-state motion of particles much less has been reported about the acceleration motion of spherical particles in incompressible Newtonian fluids. The accelerated motion is relevant to many processes such as particle classification, centrifugal and gravity particle collection and/or separation, where it is often necessary to determine the trajectories of particles accelerating in a fluid [14]. Furthermore, for other particular situations, like viscosity measurement using the falling-ball method or rain-drop terminal velocity measurement it is necessary to know the time and distance required for particles to reach their terminal velocities. In this case study, similarity transformation has been used to reduce the governing differential equations into an ordinary non-linear differential equation. In most cases, these problems do not admit analytical solution, so these equations should be solved using special techniques. The differential transform method is based on Taylor expansion. It constructs an analytical solution in the form of a polynomial. It is different from the traditional high order Taylor series method, which requires symbolic computation of the necessary derivatives of the data functions. The Taylor series method is computationally taken long time for large orders. The differential transform is an iterative procedure for obtaining analytic Taylor series solutions of differential equations. Differential transform has the inherent ability to deal with nonlinear problems, and consequently Chiou [17] applied the Taylor transform to solve non-linear vibration problems. Furthermore, the method may be employed for the solution of both ordinary and partial differential equations. Jang et al. [18] applied the two-dimensional differential transform method to the solution of partial differential equations. Finally, Hassan [19] adopted the Differential Transformation Method to solve some problems. The method was successfully applied to various practical problems [20-21].

The aim of current study is the analytical investigation of acceleration motion of a spherical particle rolling down an inclined boundary with drag coefficient in form of Eq. (3), using the Differential Transformation Method (DTM). Investigation and solution of falling objects' equation is a new application for DTM which was used for some other engineering problems.

2. PROBLEM DEFINITION

Consider a spherical particle of diameter d and density of ρ_s rolling down a smooth tube having angle of inclination α with the horizontal, and filled with an incompressible Newtonian fluid of density ρ . Let u represent the velocity of the sphere at any instant t and g the acceleration due to gravity. Figure. 1 illustrates a schematic view of the present problem.

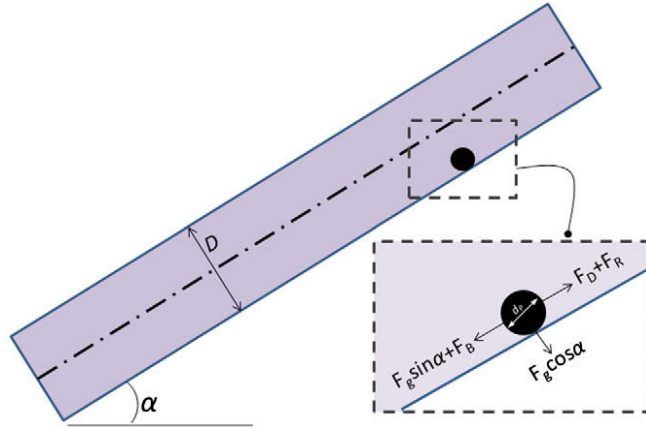


Figure.1 a schematic figure of current problem [15]

The forces acting on the spheres are the fluid-drag, F_D , fluid lift force, F_L , buoyancy force, F_B , gravitational force, F_g , solid–solid resistance force due to rolling, F_R , virtual mass force, F_{VM} due to relative acceleration of the fluid around the particle. It is reasonable to postulate that since smooth walled tubes have been studied, F_R is expected to be negligible (or alternately $F_D + F_R$ can be regarded as the total resistance to sphere motion, which is included in the drag coefficient). The all of detail of problem was explained at [15]. The equation of motion is gained as follow from [15]:

$$\frac{\pi d^3}{6} \rho_s \left(1.4 + 2.0 \frac{\rho}{\rho_s} \right) \frac{du}{dt} = \frac{\pi d^3}{6} \rho_s g \left(1 - \frac{\rho}{\rho_s} \right) \sin \alpha - \frac{15 \pi d_p^3 \rho}{100} u^2 - 2375 \pi d_p \mu u - \frac{1.00 \times 10^7 \pi \mu^2}{8 \rho_c} \quad (7)$$

To simplify the post processes of the problem we had generated four coefficients in the equation above. Therefore, Eq. (7) is reduced to:

$$a \frac{d^2 w(t)}{dt^2} + b \frac{dw(t)}{dt} + c \left(\frac{dw(t)}{dt} \right)^2 - d = 0, \quad w(0) = 0, \quad \frac{dw(0)}{dt} = 0, \quad (8)$$

Where:

$$a = \frac{\pi \cdot d^3}{6} \rho_s \left(1.4 + 2.0 \frac{\rho}{\rho_s} \right) \quad (9)$$

$$b = 23.75 \pi \cdot d_p \cdot \mu \cdot \rho \quad (10)$$

$$c = \frac{15\pi.d_p^3.\rho}{100} \quad (11)$$

$$d = \frac{\pi.d^3}{6} \rho_s g \left(1 - \frac{\rho}{\rho_s}\right) \sin \alpha - \frac{1.003 \times 10^{-7} \pi \mu^2}{8\rho_c} \quad (12)$$

With change of variation as bellow we obtain velocity,

$$u(t) = \frac{dw(t)}{dt} \quad (13)$$

By substituting Eq. (11) into Eq. (6) we will have:

$$a \frac{du(t)}{dt} + bu(t) + c(u(t))^2 - d = 0, \quad u(0) = 0 \quad (14)$$

Eqs.(8) and (14) are non-linear ordinary differential equations which could be solved by numerical techniques such Runge–Kutta method. We employed DTM and compared our results with numerical solution of 4th order Runge–Kutta method using the Maple package.

3. DIFFERENTIAL TRANSFORMATION METHOD

We suppose $x(\tau)$ to be analytic function in a domain D and $\tau = \tau_i$ represent any point in D . The function $x(\tau)$ is then represented by one power series whose center is located at τ_i . The Taylor series expansion function of $x(\tau)$ is in the form of [23]:

$$x(\tau) = \sum_{k=0}^{\infty} \frac{(\tau - \tau_i)^k}{k!} \left[\frac{d^k x(\tau)}{dt^k} \right]_{\tau=\tau_i}, \quad \forall \tau \in D \quad (15)$$

The particular case of Eq. (13) when $\tau_i = 0$ is referred to as the Maclaurin series of $x(\tau)$ and is expressed as:

$$x(\tau) = \sum_{k=0}^{\infty} \frac{\tau^k}{k!} \left[\frac{d^k x(\tau)}{d\tau^k} \right]_{\tau=0}, \quad \forall \tau \in D \quad (16)$$

As explained in [25-31] the differential transformation of the function $x(\tau)$ is defined as follows:

$$X(k) = \sum_{k=0}^{\infty} \frac{H^k}{k!} \left[\frac{d^k x(\tau)}{d\tau^k} \right]_{\tau=0} \quad (17)$$

Where, $x(\tau)$ is the original function and $X(k)$ is the transformed function. The differential spectrum of $X(k)$ is confined within the interval $\tau \in [0, H]$, where H is a constant. The differential inverse transform of $X(k)$ is defined as follows:

$$x(\tau) = \sum_{k=0}^{\infty} \left(\frac{\tau}{H}\right)^k X(k) \tag{18}$$

It is clear that the concept of differential transformation is based upon the Taylor series expansion. The values of function $X(k)$ at values of argument k are referred to as discrete, i.e. $X(0)$ is known as the zero discrete, $X(1)$ as the first discrete, etc. The more discrete available, the more precise it is possible to restore the unknown function. The function $x(\tau)$ consists of the T-function $X(k)$, and its value is given by the sum of the T-function with $\left(\frac{\tau}{H}\right)^k$ as its coefficient. In real applications, at the right choice of constant H , the larger values of argument k the discrete of spectrum reduce rapidly. The function $x(\tau)$ is expressed by a finite series and Eq. (16) can be written as:

$$x(\tau) = \sum_{k=0}^n \left(\frac{\tau}{H}\right)^k X(k) \tag{19}$$

Eq. (19) implies that the value $k = n + 1 \rightarrow \infty$ is negligible.

If $u(t)$ and $v(t)$ are two uncorrelated functions with time t where $U(k)$ and $V(k)$ are the transformed functions corresponding to $u(t)$ and $v(t)$ then we can easily proof the fundamental mathematics operations executed by differential transformation .The fundamental mathematical operations performed by differential transformation method are listed in Table 1 [25-30].

Table 1. The fundamental operations of differential transformation method	
Original function	Transformed function
$x(t) = \alpha f(t) \pm \beta g(t)$	$X(k) = \alpha F(k) \pm \beta G(k)$
$x(t) = \frac{df(t)}{dt}$	$X(k) = (k + 1)F(k + 1)$
$x(t) = \frac{d^2 f(t)}{dt^2}$	$X(k) = (k + 1)(k + 2)F(k + 2)$
$x(t) = t^m$	$X(k) = \delta(k - m) = \begin{cases} 1 & k = m \\ 0 & k \neq m \end{cases}$
$x(t) = f(t)g(t)$	$X(k) = \sum_{l=0}^k F(l)G(k - l)$

4. APPLICATION OF DIFFERENTIAL TRANSFORMATION METHOD

Now we apply Differential Transformation Method into Eq. (8) for find $w(t)$ as displacement. Taking the differential transform of Eq. (16) with respect to t according table 1 gives:

$$a((k+2)(k+1)W_{k+2}) + b((k+1)W_{k+1}) + c\left(\sum_{j=0}^k (k-j+1)W_{k-j+1}(j+1)W_{j+1}\right) - d\begin{cases} 1 & k=0 \\ 0 & \text{otherwise} \end{cases} = 0 \tag{20}$$

By suppose W_0 and W_1 are apparent from boundary conditions by solving Eq. (20) respect W_{k+2} , we will have:

$$W_2 = -\frac{1}{2} \frac{(bW_1 + cW_1^2 - d)}{a} \tag{21}$$

$$W_3 = -\frac{1}{3} \frac{(W_2(b + 2cW_1))}{a} \tag{22}$$

$$W_4 = \frac{-1}{12} \frac{(3bW_3 + 6cW_3W_1 + 4cW_2^2)}{a} \tag{23}$$

$$W_5 = -\frac{1}{5} \frac{(bW_4 + 2cW_4W_1 + 3cW_3W_2)}{a} \tag{24}$$

⋮
⋮
⋮

The above process is continuous. Substituting Eq. (20-24) into the main equation based on DTM, Eq. (19), it can be obtained the closed form of the solutions,

$$w(t) = W_0 + tW_1 - \frac{t^2}{2} \frac{(bW_1 + cW_1^2 - d)}{a} - \frac{t^3}{3} \frac{(W_2(b + 2cW_1))}{a} - \frac{t^4}{12} \frac{(3bW_3 + 6cW_3W_1 + 4cW_2^2)}{a} - \frac{t^5}{5} \frac{(bW_4 + 2cW_4W_1 + 3cW_3W_2)}{a} + \Lambda \tag{25}$$

Substituting Eq. (21-24) into the main equation based on DTM, it can be obtained the closed form of the solutions. In this stage for achieve higher accuracy we use sub-domain technique, i.e. the domain of t should be divided into some adequate intervals and the values at the end of each interval will be the initial values of next one. For example for first sub-domain assume that distance of each interval is 0.005. For first interval, $0 \rightarrow 0.005$ boundary conditions are From boundary conditions in Eq. (8) at point $t = 0$. By exerting transformation, we will have:

$$W_0 = 0 \tag{26}$$

The other boundary conditions are considered as follow:

$$W_1 = 0 \tag{27}$$

As mentioned above for next interval, $0.005 \rightarrow 0.01$, new boundary conditions are:

$$W_0 = w(0.005) \tag{28}$$

The next boundary condition is considered as follow:

$$W_1 = \frac{dw}{dt} (0.2) \quad (29)$$

For this interval function $w(t)$ is represented by power series whose center is located at 0.005, by means that in this power series t convert to $(t - 0.005)$.

As we can see bellow in similar case for achieves the solution for $u(t)$ as velocity we should apply DTM on Eq. (14) to find transformed function.

$$a((k+1)U_{k+1}) + bU_k + c \left(\sum_{j=0}^k U_{k-j} U_j \right) - d \begin{cases} 1 & k=0 \\ 0 & \text{otherwise} \end{cases} = 0 \quad (30)$$

By assuming that U_0 is apparent from boundary condition by solving Eq. (30) respect U_{k+1} , we will have:

$$U_1 = -\frac{(bU_0 + cU_0^2 - d)}{a} \quad (31)$$

$$U_2 = -\frac{1}{2} \frac{(U_1(b + 2cU_0))}{a} \quad (32)$$

$$U_3 = -\frac{1}{3} \frac{(bU_2 + 2cU_2U_0 + cU_1^2)}{a} \quad (33)$$

$$U_4 = -\frac{1}{4} \frac{(bU_3 + 2cU_3U_0 + 2cU_2U_1)}{a} \quad (34)$$

$$U_5 = -\frac{1}{5} \frac{(bU_4 + 2cU_4U_0 + 2cU_3U_1 + cU_2^2)}{a} \quad (35)$$

⋮
⋮
⋮

As mentioned above this process is continuous. By substituting Eq. (31-35) into Eq. (19), closed form of the solutions is,

$$u(t) = U_0 - t \frac{(bU_0 + cU_0^2 - d)}{a} - \frac{t^2}{2} \frac{(U_1(b + 2cU_0))}{a} - \frac{t^3}{3} \frac{(bU_2 + 2cU_2U_0 + cU_1^2)}{a} - \frac{t^4}{12} \frac{(bU_3 + 2cU_3U_0 + 2cU_2U_1)}{a} - \frac{t^5}{5} \frac{(bU_4 + 2cU_4U_0 + 2cU_3U_1 + cU_2^2)}{a} + \Lambda \quad (36)$$

And for achieve higher accuracy we use sub-domain technique as described above. By substituting Eqs. (9-12) into Eq. (25) and Eq. (36), an exact solution for $w(t)$ and $u(t)$ can be obtained which is only related to the particle and the fluid properties.

5. RESULTS AND DISCUSSION

The mentioned method was applied for real combination of solid-fluid. A single Aluminum spherical particle with versus diameter was assumed to roll down a smooth inclined plane in an infinity medium of Ethylene-glycol, glycerin solution and water. Required physical properties of selected materials are given in Table 2.

Material	Density	Viscosity
Water	996.51	0.001
Ethylene-glycol	1111.40	0.0157
Glycerin	1259.90	0.779
Aluminum	2702.0	-

In the modeling, Aluminum with density of $\rho_s = 2702.00 \text{ kg/m}^3$ is used for dispersed phase (particle). Inserting above properties into Eqs.(9) to (12), different combinations are gained which are classified in Table 3.

Fluid	d(mm)	α ($^\circ$)	B	C	D	
Water	1	5	0.000003024123088	0.000074610625	0.0004695804248	7.638667428 e-7
		1	0.000003024123088	0.000074610625	0.0004695804248	0.000002268242935
		2	0.000003024123088	0.000074610625	0.0004695804248	0.000003283167712
		3	0.000003024123088	0.000074610625	0.0004695804248	0.000004382231328
	2	1	0.00002419298470	0.000149221250	0.001878321699	0.00001814594348
		3	0.00008165132337	0.000223831875	0.004226223822	0.00006124255925
Ethylene	1	2	0.000003144432067	0.001171386812	0.0005237194650	0.000003061997753
Glycerin	1	2	0.000003299936317	0.05961388938	0.0005936963775	0.000002776106643

By substituting above coefficients in Eq. (8), and for four different inclination angles, twelve different nonlinear equations are achieved. Inclination angles were selected to be 5° , 15° , 22° and 30° . Differential Transformation Method was applied to gained equations and results were compared with numerical method. The influence of particle size is studied where the diameter of the particles is varied in the range of $1 \text{ mm} < d_p < 3 \text{ mm}$. Figs. (2 - 4) shown the variations of the displacement and velocity and acceleration for three different particles rolling in a tube inclined with the angle of 15° and filled with the water.

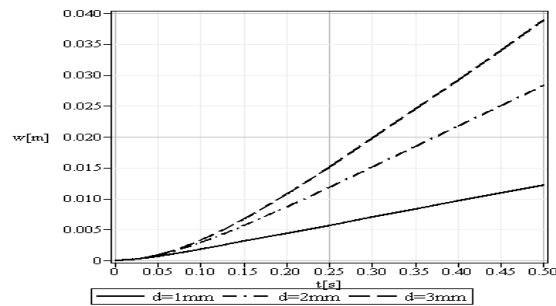


Figure.2 displacement variation for three spherical particles rolling in a tube filled with the water

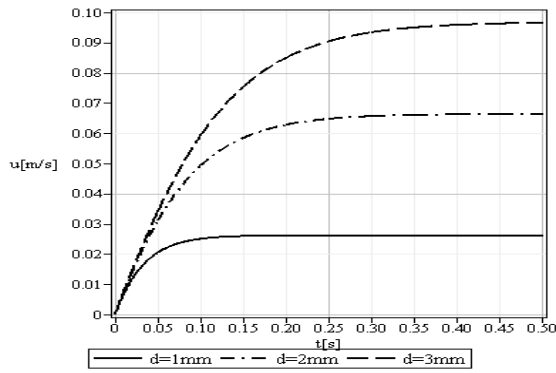


Figure.3 velocity variation for three spherical particles rolling in a tube filled with the water

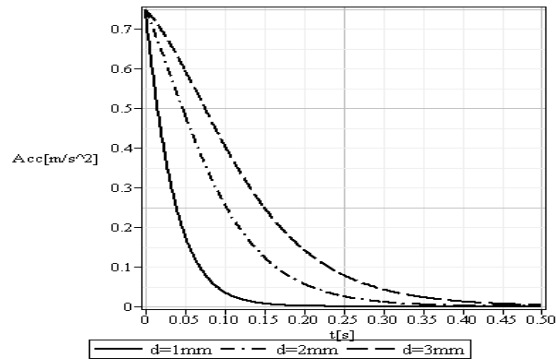


Figure.4 Acceleration variation for three spherical particles rolling in a tube filled with the water

These figures clearly illustrate that how different diameters affect the displacement and velocity and acceleration of particles while other conditions are equivalent. Observably, it is shown that the value of the displacement and velocity and acceleration in a rolling procedure is significantly increased with adding to the particle size. The variation of displacement and velocity and acceleration of the particle versus time for the different inclination angles are shown in Figs. (5-7).

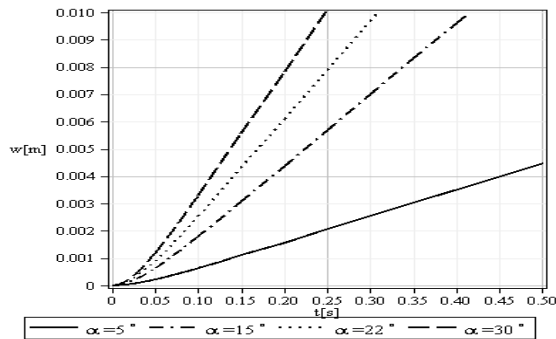


Figure.5 displacement variation of a spherical particle rolling in a tube for different angles

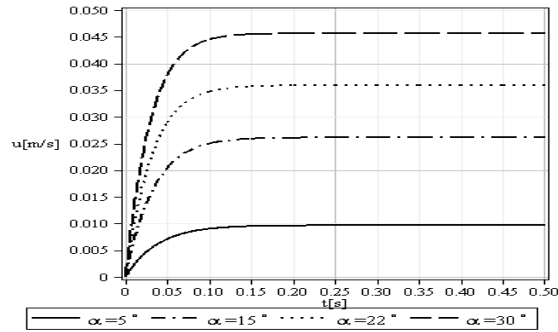


Figure.6 Velocity variation of a spherical particle rolling in a tube for different angles

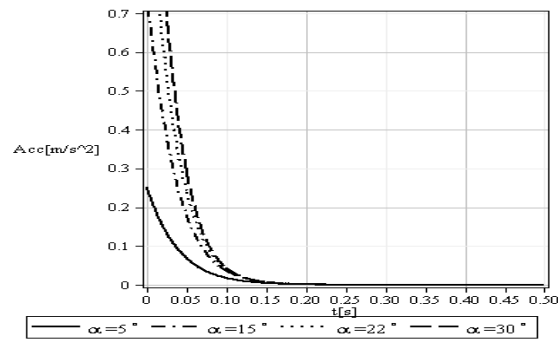


Figure.7 Acceleration variation of a spherical particle rolling in a tube for different angles

For a given the fluid viscosity, by increasing inclination angle, displacement and acceleration duration are increasing. Results show that increasing of inclination angle increases the terminal velocity as well as acceleration and displacement. Outcomes illustrated that higher acceleration is obtained for larger inclination angle. Variable displacement and velocity for sphere which its fluid is water, results of the present analysis are tabulated and comprised with the numerical solution obtained by fourth-order Runge–Kutta method in Table 4 and 5.

Table 4 the $u(t)$ obtained from DTM and NS for water, $\alpha = 15^\circ$, $d=1\text{mm}$			
t	U_{DTM}	U_{NS}	Absolute Error($U_{DTM} - U_{NS}$)
0	0	0	0
0.05	0.02047928110	0.02047930816	2.70611E-08
0.1	0.02499292820	0.02499294991	2.17E-08
0.15	0.02589242595	0.02589245023	2.42859E-08
0.2	0.02606797450	0.02606799339	1.88962E-08
0.25	0.02610209443	0.02610211188	1.74534E-08
0.3	0.02610872071	0.02610873515	1.44461E-08
0.35	0.02611000738	0.02611002114	1.38E-08
0.4	0.02611025723	0.02611026728	1.01E-08
0.45	0.02611030573	0.02611031057	4.84E-09
0.5	0.02611031515	0.02611032008	4.93E-09

Table 5 the $w(t)$ obtained from DTM and NS for water, $\alpha = 15^\circ$, $d=1\text{mm}$

T	W_{DTM}	W_{NS}	Absolute Error($W_{DTM}- W_{NS}$)
0	0	0	0
0.05	0.000629372276	0.000629371444	8.32E-10
0.1	0.001794953256	0.001794952721	5.35E-10
0.15	0.003072942044	0.003072941443	6.01E-10
0.2	0.004373099347	0.004373098915	4.32E-10
0.25	0.005677574236	0.005677573848	3.88E-10
0.3	0.006982887961	0.006982887666	2.95E-10
0.35	0.008288364581	0.008288364307	2.74E-10
0.4	0.009593872830	0.009593872669	1.61E-10
0.45	0.01089938727	0.010899387220	5E-11
0.5	0.01220490284	0.012204902801	3.9E-11

Presented results demonstrate an excellent agreement between DTM and numerical solution. In Figs.(8,9) the agreement between DTM and numerical solution for displacement and velocity of Eq.(8) when the fluid is water, $\alpha= 15^\circ$, $d_p = 1\text{mm}$ is shown.

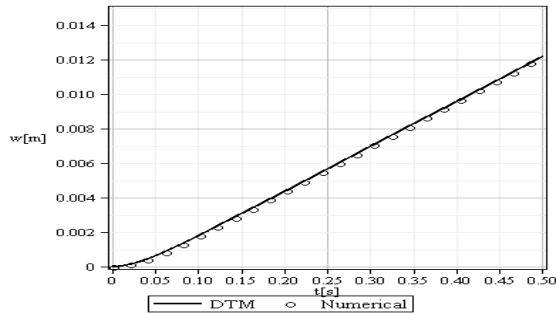


Figure.8 DTM and numerical solutions of Eq. (8) when the fluid is water, $\alpha = 15^\circ$, $d=1\text{mm}$

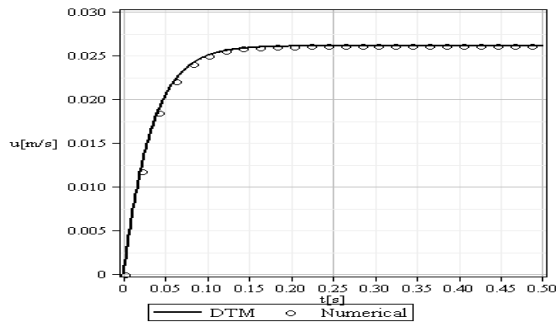


Figure.9 DTM and numerical solutions of Eq. (8) when the fluid is water, $\alpha= 15^\circ$, $d=1\text{mm}$

In this case, a very interesting agreement between the results of two methods is observed which confirms the excellent validity of the DTM.

6. CONCLUSIONS

In this paper, Differential Transformation Method (DTM) is applied to obtain the solution of the unsteady motion of a spherical particle rolling down an inclined tube in a Newtonian fluid. Equation was solved generally and for some real combinations of solid-liquid. Instantaneous velocity, acceleration and position were obtained as results and outcomes were compared with Runge–Kutta method solution. Very good agreement has been seen between numerical and current analytical method. Results show that for a given condition of particle and fluid, an increase in inclination angle, α , results in an increase in terminal displacement and velocity and acceleration. Current work approved the simplicity and capability of Differential Transformation Method. Solution of equation of motion for an object rolling down an inclined boundary is a new application of DTM and could be used in wide area of scientific problems, especially hydraulic and sedimentation engineering.

REFERENCES

- [1] J.W. Delleur, New results and research needs on sediment movement in urban drainage, *J. Water Resour. Plan. Manage*, ASCE; 186–19(2001).
- [2] T. Hvitved-Jacobsen, J. Vollertsen, N. Tanaka, Wastewater quality changes during transport in sewers: An integrated aerobic and anaerobic model concept for carbon and sulfur microbial transformations, *Water Sci. Technol.* 257–264(1998).
- [3] Z.X. Cao, Equilibrium near-bed concentration of suspended sediment, *J. Hydraul. Eng.*, ASCE; 1270–1278(1999).
- [4] J.S. Bridge, S.J. Bennett, A model for the entrainment and transport of sediment grains of mixed sizes, shapes, and densities, *Water Resour. Res.*, 337–363(1992).
- [5] J. Duran, The physics of fine powders: Plugging and surface instabilities, *C.R. Phys.*, 17–227(2002).
- [6] J.G. Yates, *Fundamentals of fluidized-bed processes*, Butterworths, London, (1983).
- [7] M.A. Hasan, The role of wall effects on the rolling velocity of spheres in Newtonian fluids, *Chemical Engineering Journal* 33 -97(1986).
- [8] J. L. Boillat, N.H. Graf, Vitesse de sedimentation de particules spheriques en milieu turbulent, *J. Hydraul. Res.* 395-413(1982).
- [9] D.D. Joseph, Y.L. Liu, M. Poletto, J. Feng, Aggregation and dispersion of spheres falling in viscoelastic liquids, *J. Non-Newtonian Fluid mech*, 54: 45-86 (1994).
- [10] R. Clift, J.R. Grace, M.E. Weber, *Bubbles, Drops and Particles*, Academic Press, New York, (1978).
- [11] A.R. Khan, J.F. Richardson, The resistance to motion of a solid sphere in a fluid, *Chem. Eng. Commun.*; 135– 150(1987).
- [12] R.P. Chhabra, J.M. Ferreira, An analytical study of the motion of a sphere rolling down a smooth inclined plane in an incompressible Newtonian fluid, *Powder Technology* ,130–138(1999).
- [13] M. Hartman, J.G. Yates, Free-fall of solid particles through fluids, *Collect. Czechoslov. Chem. Commun.* 961–982(1993).
- [14] J.M. Ferreira, R.P. Chhabra, Accelerating motion of a vertically falling sphere in incompressible Newtonian media: an analytical solution, *J. Powder Technology*, 97: 6-15(1998).
- [15] C.D. Jan, J.C. Chen, Movements of a sphere rolling down an inclined plane. *J. Hydraulic Res.*, 689–706(1997).
- [16] J.S. Chiou, J.R. Tzeng, Application of the Taylor transform to nonlinear vibration problems, *Transaction of the American Society of Mechanical Engineers, J. Vib.Acoust.* (83–87)(1996).
- [17] C.K. Chen, S.P. Ju, Application of differential transformation to transient advective–dispersive transport equation, *Appl. Math. Comput.* 25–38(2004).
- [18] Y.L. Yeh, C.C. Wang, M.J. Jang, Using finite difference and y to analyze of large deflections of orthotropic rectangular plate problem, *Appl. Math. Comput.* 1146–1156(2007).
- [19] I.H. Abdel-Halim Hassan.: Differential transformation technique for solving higher-order initial value problems. *Appl. Math. Comput.* 299–311(2004).
- [20] M. Jalaal, D.D. Ganji, An analytical study on motion of a sphere rolling down an inclined plane submerged in a Newtonian fluid, *Powder Technology*, 82–92(2010).

- [21] M. Jalaal, D.D. Ganji, G. Ahmadi, Analytical investigation on acceleration motion of a vertically falling spherical particle in incompressible Newtonian media, *Advanced Powder Technology* (2009).
- [22] M. Jalaal, D.D. Ganji, On unsteady rolling motion of spheres in inclined tubes filled with incompressible Newtonian fluids, *Advanced Powder Technology* (2010).

PDD CRAWLER: A FOCUSED WEB CRAWLER USING LINK AND CONTENT ANALYSIS FOR RELEVANCE PREDICTION

Prashant Dahiwale¹, M M Raghuwanshi² and Latesh Malik³

¹Research Scholar , Dept of CSE, GHRCE and
Assist Prof Dept of CSE, RGCER, Nagpur, India

²Department of Computer Science Engineering, RGCER, Nagpur India

³Department of Computer science Engineering, GHRCE, Nagpur, India

¹prashantdd.india@gmail.com, ²m_raghuwanshi@rediffmail.com

³lateshmalik@gmail.com

ABSTRACT

Majority of the computer or mobile phone enthusiasts make use of the web for searching activity. Web search engines are used for the searching; The results that the search engines get are provided to it by a software module known as the Web Crawler. The size of this web is increasing round-the-clock. The principal problem is to search this huge database for specific information. To state whether a web page is relevant to a search topic is a dilemma. This paper proposes a crawler called as “PDD crawler” which will follow both a link based as well as a content based approach. This crawler follows a completely new crawling strategy to compute the relevance of the page. It analyses the content of the page based on the information contained in various tags within the HTML source code and then computes the total weight of the page. The page with the highest weight, thus has the maximum content and highest relevance.

KEYWORDS

Web Crawler, HTML, Tags, Searching Relevance, Metadata

1. INTRODUCTION

The World Wide Web is a huge collection of data. The data keeps increasing week by week, day by day and hour by hour. It is very important to classify data as relevant or irrelevant in accordance with users query. Researchers are working on techniques which would help to download relevant web pages. Researchers say that the huge size of data results in low exposure of complete data while search is performed and it is predicted that only one third of the data gets indexed[1].The web is so large that even the number of relevant web pages that get download is too large to be explored by the user. This scenario generates the need of downloading the most relevant and superior pages first. Web search is currently generating more than 13% of the traffic to Web sites [2]. Many policies have been developed to schedule the downloading of relevant web pages for a particular search topic. Some of the policies are Breadth First Search, Depth First Search, Page Ranking Algorithms, Path ascending crawling Algorithm, Online Page Importance Calculation Algorithm, Crawler using Naïve Bayes Classifier, Focused Web Crawler, Semantic web Crawler etc. Each technique has its pros and cons. Focused Web Crawler is a technique which uses the similarity major to map relatedness among the downloaded page and unvisited

David C. Wyld et al. (Eds) : SAI, CDKP, ICAITA, NeCoM, SEAS, CMCA, ASUC, Signal - 2014

pp. 245–253, 2014. © CS & IT-CSCP 2014

DOI : 10.5121/csit.2014.41123

page. This technique ensures that similar pages get downloaded and hence the name Focused web crawler[3]. Web crawler needs to search for information among web pages identified by URLs. If we consider each web page as a node, then the World Wide Web can be seen as a data structure that resembles a Graph'. To traverse a graph like data structure our crawler will need traversal mechanisms much similar to those needed for traversing a graph like Breadth First Search (BFS) or Depth First Search (DFS). Proposed Crawler follows a simple Breadth First Search'approach. The start URL given as input to the crawler can be seen as a start node' in the graph. The hyperlinks extracted from the web page associated with this link will serve as its child nodes and so on. Thus, a hierarchy is maintained in this structure. Each child can point to its parent is the web page associated with the child node URL contains a hyperlink which is similar to any of the parent node URLs. Thus, this is a graph and not a tree. Web crawling can be considered as putting items in a queue and picking a single item from it each time. When a web page is crawled, the extracted hyperlinks from that page are appended to the end of the queue and the hyperlink at the front of the queue is picked up to continue the crawling loop. Thus, a web crawler deals with the infinite crawling loop which is iterative in nature. Since crawler is a software module which deals with World Wide Web, a few constraints [4] have to be dealt with: High speed internet connectivity, Memory to be utilized by data structures, Processor for algorithm processing and parsing process Disk storage to handle temporary data.

2. RELATED WORK

Mejdl S. Safran, Abdullah Althagafi and Dunren Che in Improving Relevance Prediction for Focused Web Crawlers'[4] propose that a key issue in designing a focused Web crawler is how to determine whether an unvisited URL is relevant to the search topic. this paper proposes a new learning-based approach to improve relevance prediction in focused Web crawlers. For this study, Naïve Bayesian is used as the base prediction model, which however can be easily switched to a different prediction model. Experimental result shows that approach is valid and more efficient than related approaches.

S. Lawrence and C. L. Giles in Searching the World Wide Web'[1] state that the coverage and recency of the major World Wide Web search engines when analyzed, yield some surprising results. Analysis of the overlap between pairs of engines gives an estimated lower bound on the size of the indexable Web of 320 million pages. Pavalam S. M., S. V. Kasmir Raja, Jawahar M., and Felix K. Akorli in Web Crawler in Mobile Systems' [6] propose that with the advent of internet technology, data has exploded to a considerable amount. Large volumes of data can be explored easily through search engines, to extract valuable information. Carlos Castillo , Mauricio Marin , Andrea Rodriguez in Scheduling Algorithms for Web Crawling'[7] presents a comparative study of strategies for Web crawling. It shows that a combination of breadth-first ordering with the largest sites first is a practical alternative since it is fast, simple to implement, and able to retrieve the best ranked pages at a rate that is closer to the optimal than other alternatives. The study was performed on a large sample of the Chilean Web which was crawled by using simulators, so that all strategies were compared under the same conditions, and actual crawls to validate our conclusions. The study also explored the effects of large scale parallelism in the page retrieval task and multiple-page requests in a single connection for effective amortization of latency times.

Junghoo Cho and Hector Garcia-Molina in Effective Page Refresh Policies for Web Crawlers' [8] study how to maintain local copies of remote data sources fresh, when the source data is updated autonomously and independently. In particular, authors study the problem of web crawler that maintains local copies of remote web pages for web search engines. In this context, remote data sources (web sites) do not notify the copies (Web crawlers) of new changes, so there is a need to periodically poll the sources, it is very difficult to keep the copies completely up-to-date. This

paper proposes various refresh policies and studies their effectiveness. First formalize the notion of Freshness of copied data by defining two freshness metrics, and then propose a Poisson process as a change model of data sources. Based on this framework, examine the effectiveness of the proposed refresh policies analytically and experimentally. Results show that a Poisson process is a good model to describe the changes of Web pages and results also show that proposed refresh policies improve the freshness of data very significantly. In certain cases, author got orders of magnitude improvement from existing policies. The Algorithm design Manual [9] by Steven S. Skiena is a book intended as a manual on algorithm design, providing access to combinatorial algorithm technology for both students and computer professionals. It is divided into two parts: Techniques and Resources. The former is a general guide to techniques for the design and analysis of computer algorithms. The Resources section is intended for browsing and reference, and comprises the catalog of algorithmic resources, implementations, and an extensive bibliography. Artificial Intelligence illuminated [10] by Ben Coppin introduces a number of methods that can be used to search, and it discusses how effective they are in different situations. Depth-first search and breadth-first search are the best-known and widest-used search methods, and it is examined why this is and how they are implemented. A look is also given at a number of properties of search methods, including optimality and completeness [11][12], that can be used to determine how useful a search method will be for solving a particular problem.

3. DESIGN METHODOLOGY

In general, a web crawler must provide the features discussed [13] ,

1. A web crawler must be robust in nature Spider traps are a part of the hyperlink structure in the World Wide Web. There are servers that create spider traps, which mislead crawlers to infinitely travel a certain unnecessary part of the web. Our crawler must be made spider trap proof.
2. Web pages relate to web servers, i.e. different machines hosting these web pages and each web page has its own crawling policy, thus our crawler must respect the boundaries that each server draws.

3.1 Existing Methodology [14] [15] [16]

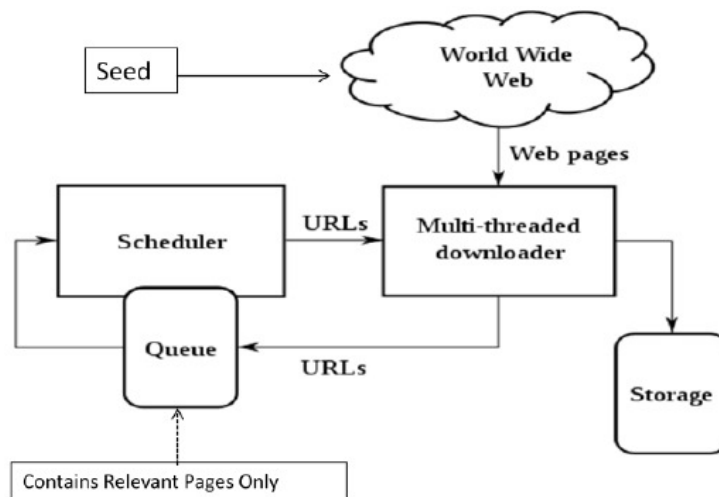


Figure1. High-level architecture of a standard Web crawler

Components of the web crawler are described as [4]:

1. Seed - It is the starting URL from where the where crawler starts traversing the World Wide Web recursively.
2. Multithreaded Downloader: It downloads page source code whose URL is specified by seed.
3. Queue: Contains unvisited hyperlinks extracted from the pages.
4. Scheduler: It is FCFS scheduler used to schedule pages in Queue.
5. Storage: Storage may be volatile or non-volatile data storage component.

3.2 Proposed Approach

1. This paper proposes a PDD crawler, which is both links based and content, based. The content that had been unused by the previous crawlers will also take part in the parsing activity. Thus, Rank Crawler is a link as well as content-based crawler.
2. Since true analysis of the web page is taking place (i.e. the entire web page is searched for the content fired by the user), this crawler is very well suited for business analysis purposes.

3.3 Architecture

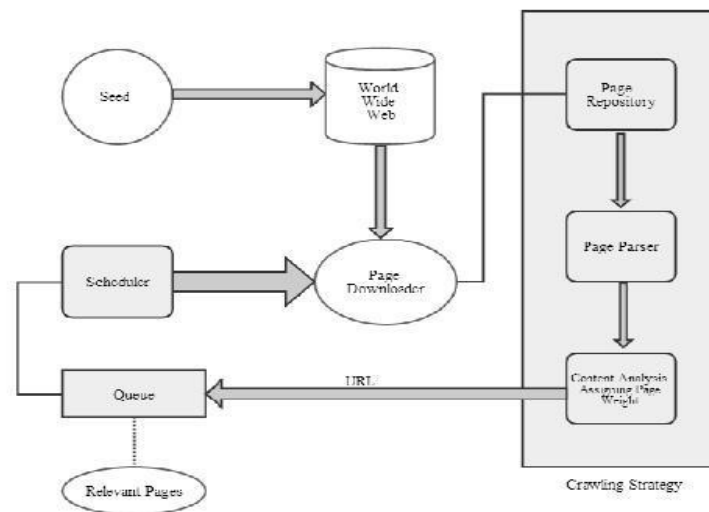


Figure 2. Proposed Architecture

Before describing the actual working of proposed Crawler, it is important to note that:

1. Every URL has got a web page associated with it. Web pages are in the form of HTML source code in which information is divided in various tags. Thus, page parsing and information retrieval are solely based on the tags of HTML source code.
2. Web pages have a certain defined structure, a hierarchy of sorts. This hierarchy defines the importance of tags with respect to a web page. Meta is the most important tag or the tag with the highest priority, whereas body is the one with the lowest priority.

Proposed PDD Crawler' follows these steps:

1. The seed URL used to connect to the World Wide Web. Along with the seed, a search string is to be provided by the user. The search string will act as the query that the user is searching for. If the URL is valid and the string is valid, then the crawler proceeds else it stops.
2. Every Web page has got tags that may contain information. Weights in the form of integers have been assigned to the tags of the page. If the content is found on the page (once or multiple times), the weight is to be multiplied by the number of occurrences in each tag and the total weight is added up to get the total page weight.
3. The Seed downloaded and all the URLs on the seed pages are extracted. The seed page is then checked to see whether it contains any relevant information. If yes, then the weight is calculated and it is set aside for further use, else the page is irrelevant to users query and has to be thrown away.
4. The URLs extracted are then scheduled to be downloaded one by one and the above process is followed (i.e. download a page and check its contents). If it returns a positive page weight, then set it aside and recursively do the same for every page.
5. The page with the most weight has the highest content. Thus, the results have to be descending sorted.

3.4 Architecture Components

1. Page Repository: Repository refers to the web pages that have been traversed and parsed. It may be a temporary storage from where the crawlers can refer the pages or it can be the browsers cache or storage from where it can refer the pages for faster access.
2. Page Parsing Mechanism: Page parsing mechanism takes pages one at a time from the page repository and searches for the keyword or the string in the page and based on that assign weight to the page.
3. Content Analysis: Final phase, which decides whether the page is relevant or whether it has to be discarded. Relevancy is decided on the basis of relevancy threshold which we have maintained in the algorithm.

3.5 Working of 'PDD Crawler

Proposed PDD Crawler' works using the source code of the page i.e. the downloader uses the source code to analyze the tags and contents of the page so as to get the page weight and calculate the degree of relevancy of a page. To make a simple web page you need to know the following tags:

1. < HTML > tells the browser your page is written in HTML format
2. < HEAD > this is a kind of preface of vital information that does not appear on the screen.
3. < TITLE > Write the title of the web page here - this is the information that viewers see on the upper bar of their screen.
4. < BODY > This is where you put the content of your page, the words and pictures that people read on the screen.
5. < META > This element used to provide structured metadata about a Web page. Multiple Meta elements with different attributes are often used on the same page. Meta elements can be used to specify page description, keywords and any other metadata not provided through the other head elements and attributes.

Consider the source code associated with the web page of the URL: <http://www.myblogindia.com/html/default.asp> as given below-

```
<html>
<head>
<meta name="description" content="Free HTML Web tutorials">
<meta name="keywords" content="HTML, CSS, XML">
<meta name="author" content="RG CER">
<meta charset="UTF-8">
<title > HTML title of page</title >
</head>
< body>
    This is my very own HTML page. This page is just for reference.
</body >
</html >
```

The downloader gets this content from a page. Downloading a page refers to the function of getting the source code of the page and analyzing as well as performing some actions on it. This method is what we call Parsing. The analysis part of the source code is as follows:

1. Let the Total weight of the page be t' units
2. The body tag has got the weight B' units
3. The title tag has got the weight T' units
4. The Meta tag has got the weight M' units
5. The heading tag (h1 through h6) has weight H' units
6. The URL has weight U' units
7. The no. of occurrence of the search string in body be Nb
8. The no. of occurrence of the search string in title be Nt
9. The no. of occurrence of the search string in META be Nm
10. The no. of occurrence of the search string in heading be Nh
11. The no. of occurrence of the search string in URL be Nu
- 12.

The total weight of the page would be –

$$t = (Nb*B)+(Nt*T)+(Nm*M)+(Nh*H)+(Nu*U)$$

Assumptions for calculating page weight are defined below: M = 5 units, U = 4 units, T = 3 units, H = 2 units, B = 1 units Suppose the search string the user entered is: html (not case-sensitive). The number of occurrences of html in the following tags is as follows: Nb = 1, Nt = 1, Nm = 2, Nh = 0, Nu = 1 The total weight of the page comes out to be:

$$\begin{aligned} t &= (1*1) + (1*3) + (2*5) + (0*2) + (1*4) \\ t &= 1 + 3 + 10 + 4 \\ t &= 18 \end{aligned}$$

Conclusion:

1. The page weight (t) > 3, thus the page is relevant.
2. Page is Relevant if and only if t > 3.
3. All pages with t <= 3 will be discarded.
4. Will only work for static pages.
5. Static pages are the one which do not change or update or alter their content on regular basis i.e. the data change that is there on pages is either periodic or none what so ever.
6. Some pages are non-crawl able (e.g. the pages with META as robot).

4. RESULT

The proposed PDD Crawler' was implemented on the following hardware and software specifications: OS Name- Microsoft Windows 7 Ultimate, Version-6.1.7600 Build 7600, Processor- Intel(R) Core(TM) i5-3230M CPU @ 2.60 Ghz, 2601.and compared its performance with intelligent web crawler which is link based crawler , after running both the crawlers in same hardware and software environment with same input parameters, observed following rate of precisions for both the crawlers as shown in Table 1. evaluate the performance of proposed framework by comparing it with Intelligent web crawler by applying both the crawlers to the same problem domains such as Book Show, Book to read, Cricket match and Match making. The experiments are conducted on different social aspects database (web) having same words with different meaning as book meaning reserving some resource for the domain Book showl while as book meaning a physical entity made up by binding pages for the domain book to read. From the above test cases it is vibrant that if correct Seed URLs are provided according to domain sense of the word then:

1. The precision range comes out to be 20 to 70% which is fairly acceptable for the crawl.
2. The test results show that the search is now narrowed down and very specific results are obtained in sorted manner.
3. The sorting of the results cause the most relevant link to be displayed first to the user so as to save his valuable search time

Table 1: Result Comparisons

			Intelligent Crawler	PDD Crawler
Sr No	Seed URL	Query	Percentage Precision	
1.	www.bookmyevent.com	Book show	22.14%	59.37%
	www.ticketnew.com			
	www.bookmyshow.com			
2.	www.best sellers.about.com	Book to read	33.64%	67.53%
	www.publicbookshelf.com			
	www.goodreads.com			
3.	www.espnricinfo.com	Cricket match	32.43%	65.87%
	www.starsports.com			
	www.icc-cricket.com			
4.	www.astrosage.com	Match making	42.65%	69.89%
	www.mangalsutra.com			
	www.drikpanchang.com			

4. The testing of the framework is carried out on live web, thus by precision reported accuracy of proposed Focused Web Crawler is promising.

Above test results completely depend on Seed URLs. The Seed URLs provide the correct domain for the search. Seed URLs help to initiate and guide the search in interested domain. Later on the Focused nature of the proposed Crawler will always deviate the search towards target links.

5. CONCLUSION AND FUTURE WORK

The main advantage of proposed crawler over the intelligent crawler and other Focused Crawlers is that it does not need any Relevance Feedback or training procedure in order to act intelligently; two kinds of change were found after comparing result of both the crawlers.

1. The number of extracted documents was reduced. Link analyzed, and deleted a great deal of irrelevant web page.
2. Crawling time is reduced. After a great deal of irrelevant web page is deleted, crawling lode is reduced.

In conclusion, after link analysis and page analysis in proposed crawler, crawling precision is increased and crawling rate is also increased. This will be an important tool to the search engines and thus will facilitate the newer versions of the search engines

REFERENCES

- [1] S. Lawrence and C. L. Giles. Searching the World Wide Web. *Science*, 280(5360):98.100,1998.
- [2] StatMarket. Search engine referrals nearly double worldwide. <http://websidestory.com/pressroom/-pressreleases.html?id=181>, 2003.
- [3] Rajshree Shettar, Dr. Shobha G, Web Crawler on Client Machine, Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol II, IMECS 2008, 19-21 March, 2008, Hong Kong.
- [4] Mejdil S. Safran, Abdullah Althagafi and DunrenChe Improving Relevance Prediction for Focused Web Crawlers, in the proceeding of 2012 IEEE/ACIS 11th International Conference on Computer and Information Science.
- [5] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze, An Introduction to Information Retrieval, Cambridge University Press, Cambridge, England.
- [6] Pavalam S M, Jawahar M, Felix K Akorli, S V Kashmir Raja Web Crawler in Mobile Systems in the proceedings of International Conference on Machine Learning (ICMLC 2011),
- [7] Carlos Castillo, Mauricio Marin, Andrea Rodriguez,—Scheduling Algorithms for Web Crawling in the proceedings of WebMedia and LA-Web, 2004.
- [8] Junghoo Cho and Hector Garcia-Molina Effective Page Refresh Policies for Web Crawlers *ACM Transactions on Database Systems*, 2003.
- [9] Steven S. Skiena —The Algorithm design Manual Second Edition, Springer Verlag London Limited, 2008, Pg 162.
- [10] Ben Coppin Artificial Intelligence illuminated Jones and Barlett Publishers, 2004, Pg 77.
- [11] Alexander Shen Algorithms and Programming: Problems and solutions Second edition Springer 2010, Pg 135
- [12] NarasinghDeo Graph theory with applications to engineering and computer science PHI, 2004 Pg 301
- [13] DebashisHati, BiswajitSahoo, A. K. "Adaptive Focused Crawling Based on Link Analysis," 2nd International Conference on Education Technology and Computer (ICETC), 2010.
- [14] Brian Pinkerton, Finding what people want: Experiences with the Web Crawler, Proceedings of first World Wide Web conference, Geneva, Switzerland, 1994.
- [15] Gautam Pant, Padmini Srinivasan, Filippo Menczer, Crawling the Web, pp. 153-178, Mark Levene, Alexandra Poulouvassilis (Ed.), *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*, Springer-Verlag, Berlin, Germany, November 2004.
- [16] B. Brewington and G. Cybenko, How dynamic is the web?, Proceedings of the 9th World Wide Web, Conference (WWW9), World Wide Web, Conference , 2000.

AUTHORS

Prashant Dahiwale: Is a Doctoral research fellow at G H Raisonni college of Engineering and research , Nagpur under RTM Nagpur University Nagpur in computer science and engineering. Having 5 years of teaching experience to teach M.tech and B.E comp Sc Engg courses and 1 year of industrial experience. His area of research is information retrieval and web mining. Completed Post PG diploma from CDAC in 2010. M.Tech in Comp Sc Engg form Visvesvaraya National Institute of Technology, (VNIT) Nagpur in 2009,B.E from RTM Nagpur university Nagpur in 2006. Published various research papers at International and National journals and conferences. Also author is a four time finalist of Asia regional ACM ICPC since 2010.organised various workshop on programming language fundamentals and advance concepts.



Dr M M Raghuvanshi: Professor dept of Comp sc Engg ,RG CER, Nagpur India . having total 25 plus years of teaching experience and 5 years of industrial experience. His research area is genetic algorithm, he has chaired multiple international and national conferences, he is IEEE reviewer, Ph D form Visvesvraya National Institute of Technology, Nagpur, and M.Tech from IIT Kharagpur. he is a author of text book on Algorithm and Data structure , He thought multiple subjects to M.Tech and B.E courses. he has organised multiple workshops on Theory of computations and Genetic algorithm. Published various research papers at International and National journals and conferences.



INTENTIONAL BLANK

POLICY OVERLAP ANALYSIS TO AVOID POLICY CONFLICT IN POLICY-BASED MANAGEMENT SYSTEMS

Abdehamid Abdelhadi Mansor¹, Wan Mohd Nasir Wan Kadir² and
Ahmed Mohammed Elswawi²

¹Department of Computer Science, Faculty of Mathematical Sciences,
University of Khartoum, Khartoum, Sudan

abhamidhn@uofk.edu

²Department of Software Engineering, Faculty of Computing, University
Technology Malaysia, Johor, Malaysia

wnasir@cs.utm.my, elswawi@gmail.com

ABSTRACT

A management policy evolves over time by addition, deletion and modifications of rules. Policies authored by different administrators may be merged to form the final system management policy. These operations cause various problems such as policy overlap. Static and dynamic conflicts are considered as two classes of conflict which need to be understood and independently managed. Furthermore, the distinction between these two classes is important; as detecting and resolving of conflict can be computationally intensive, time consuming and hence, costly. However, a dynamic conflict is quite unpredictable, in that it may, or may not, proceed to a state of a realized conflict. In this paper we present static analyses to address the overlap cases when there are two or more policies are enforced simultaneously. Moreover, the paper provides temporal specification patterns to avoid each type of conflicts, and to ensure that policies are enforced correctly.

KEYWORDS

Policy-conflict, overlap, policy-based management, static analysis

1. INTRODUCTION

Policy-based management is a well-established approach where policies are specified as Event-Condition-Action (ECA) rules which specifying the management actions to be performed when certain situations occur. However, the main challenge limits the development of policy-based approach is the policy conflicts. Conflicts may arise in the set of policies and also may arise during the refinement process, between the high-level goals and the implementable policies [1]. The system must be able to handle conflicts such as exceptions to normal authorization policies. For instance, in a large distributed system there will be multiple human administrators specifying policies which are stored on distributed policy servers. Conflict detection between management

policies can be performed statically for a set of policies in a policy server as part of the policy specification process or at run-time [2].

In policy-based management system, policies are specified by the system manager to govern system behavior. Such governing policies evolve over time by policy composition, rule modifications and due to system dynamisms [2]. Multiple policies can become simultaneously eligible for enforcement in a situation and the order of enforcement may determine the final system state. However, in such cases, changes in policies at run time may result in system instability, overlap and cycles among rules which are lead to policy conflicts [3]. The resulting conflicts can be detected and avoided during system design. In this paper, the relationship between policies is a crucial to our discussion of policy conflicts as it is our contention.

Overlap may arise in many situations related to sharing of resources for which both domains have applied to policies that have management responsibilities on the same set of object [4]. Overlapping domains reflect the fact that multiple managers can be responsible for that multiple policies apply to the object or an object. Obviously this can lead to conflicts between policies or managers. Existing approaches to conflict detection are limited in scope and can only detect conflicting actions if they are explicitly stated. In addition, current techniques do not detect overlaps in management policies. This paper presents static analyses to specify and detect potential overlap in order to be avoided earlier since the design time, where most of the required specification can be detected. Moreover, the paper provides temporal specification patterns to avoid the potential overlaps, and to ensure that policies are enforced correctly.

In section II of this paper, briefly presents an introduction of PobMC, and the smart mall system. Section III presents system architecture management. In section IV overlap Analysis is presented and discussed. Sections V provides and discusses conflict analysis. Related work is presented in VII. Finally, conclusions and further work are discussed in section VIII.

2. THE APPLICATION OF POBMC

A. Introduction to PobMC

In PobMC [5], policies are used to control the system behavior. Policies provide a high-level of abstraction and allow us to decouple the adaptation concerns from the application code. Thus, we can change the system behaviour as well as adaptation schemas by changing policies. PobMC is composed of a set of modules; called Self-Managed Module (SMM) is the policy-based building block of PobMC. In our case study, we consider three SMMs including SenModule, LocModule and SecModule to manage sensors, locations and security constraints respectively. An SMM is a set of actors which can manage their behavior autonomously according to predefined policies. PobMC supports interactions of an SMM with other SMMs using well-defined interfaces in the model

Each module in PobMC consists of managers and actors. Actors, which are manage their behaviour autonomously according to predefined policies. Managers, which are manage and provide autonomic behaviour to corresponding actors. Interaction among managers is supported in PobMC.

B. Illustrative Case Study

This section, briefly presents an example that we use throughout the paper to illustrate the supported concepts of adaptation and the underlying modelling techniques.

The Smart Mall System (SMALLS) is a system that allows users to navigate their location in the mall. The users could be able to query the place that they are heading such as baby area, shoes area, food area, banking services area etc. The system directs user how to find the area. SMALLS operation can be summarized as follows. Each user carries a mobile device such as a smart phone as well as a wireless sensor. In addition, locations in the environment shopping area or services area are associated with their own wireless sensors. The sensors determine which area is closest to the user at a given moment and pass this information to a server, which provides specific Web services for each individual object.

SMALLS is required to adapt its behaviour according to the changes of the environment. To achieve this aim, we suppose that the system runs in normal, vacation and failure modes and in each context it enforces various sets of policies to adapt to the current conditions. For the reason of area, here we only identify policies defined for sensing control module while the system runs in normal or failure modes.

3. SYSTEM ARCHITECTURE MANAGEMENT

The main concepts of our system architecture management are grouping objects “domain”, support the specification “a policy service” and reflect the organizational structure “storage of policies and roles”, responsibilities and relationships between the system components “managers”, “coordinator”, and “managed elements”.

Domains are used to group system objects according to object type, responsibility and authority. A sub-domain is a member of another domain “parent domain”. However subdomain is not subset of the parent domain, an object or subdomain may be member of multiple parent domains, figure 1 illustrates the relationship between SMALLS domains. In Figure 1, all the objects in sub-domains SenModule, LocModule and SecModule are members of parent domain SMALLS_Management which therefore overlap.

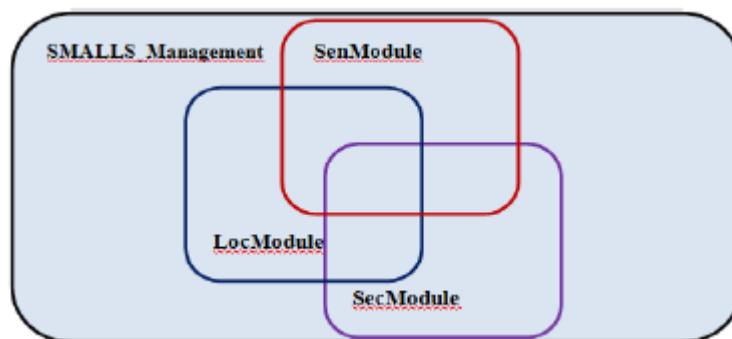


Figure 1. SMALLS Domains

Domains can be identified by path name such as:

/SMALLS_Management/SenModule , // which identify the domain SenModule

Policy applying to domain SenModule will also apply to members of domain SMALLS_Management. Some policies which applying to domain SenModule can be applied to domains SecModule and LocModule. However, all system policies are applying to domain SMALLS_Management.

Using union \cup , intersection \cap and difference $-$ operators, domains can be combined to form a new set of objects for applying a policy. The advantage of combining domains is that deletion and addition of objects from/to the domains can be done without changing the policies.

Two domains overlap if there are objects which are members of both domains, for example SenModule & LocModule in figure 1. Overlap arises in many situations related to sharing of resources for which both domains have applied to policies that have management responsibilities on the same set of object. Overlapping domains reflect the fact that multiple managers can be responsible for that multiple policies apply to the object or an object. Obviously this can lead to conflicts between policies or managers.

A policy, whether it is concerned with obligations or with authority, has the following attributes [6]:

- (i) Modality; each policy has positive or negative modality (see figure 2), which are important and adequate for the analysis in this work,
- (ii) Policy subjects; which define a set of users to whom the policy is directed,
- (iii) Policy target objects; define the set of objects at which the policy is directed,
- (iv) Policy goals; can be expressed as high-level goals which specify what the manager should achieve in abstract terms which do not identify how to achieve the goals, and
- (v) Policy constraints can be expressed in terms of system properties, such as extent or duration, or some other condition.

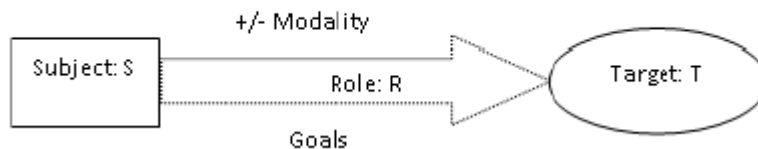


Figure 2. Policy Configuration

For any set of policies $\{p_i, q_j, O_k\}$ has been enforced in the system, the term policy conflict can be defined as follow:

Tow policies p_i and q_j are in conflict if and only if one of the following cases takes place:

- (i) p_i and q_j have been enforced simultaneously, then the system cannot choose a policy to enforce.
- (ii) The execution of p_i violates the action of q_j .

- (iii) Executing p_i makes q_j impossible to be enforced and vice versa (eg. turn-on and turn-off for the same device simultaneously).
- (iv) Executing of p_i before q_j while it must be executed after q_j (the ordering). For instance, q_j is “authorize the user” and p_i is “download the system files”.

While in the system specification, the system must authorize the user before he gets the system files.

In order to detect the conflicting policies first we must identify and define conflicting actions explicitly. Then, the simultaneous triggering of those policies should be investigated. Second, the ordering of events and actions should be identified clearly. Third, the inconsistent policies should be identified to prevent them from simultaneous execution. Finally, all system policies should be checked to identify whether policies make the action of another policies by violating their conditions. For instance, in our SMALLS example if Ok is the policy that “identify the mobile phone location”, while the mobile phone is currently attached to the corresponding APs, no policy that disable the database server must be applied before the policy that “send the required information to the mobile phone”.

4. OVERLAP ANALYSIS

In this work, static analysis is used to determine whether an event specified in the policy condition matches received event. A trigger graph is created after the policy compilation to identify the overlap between set of subjects, targets and actions (see figure 3), in simultaneously triggered policies. Furthermore, specifying the overlap will eventually avoid modality conflicts and multimanager conflicts, thereby improves system scalability. Static analysis is capable to evaluate only potential conflicts rather than actual conflicts. However, static analysis is limited to evaluate policy constraints, because of that constrains are completely depending on run-time state; moreover domain membership may change at run-time.

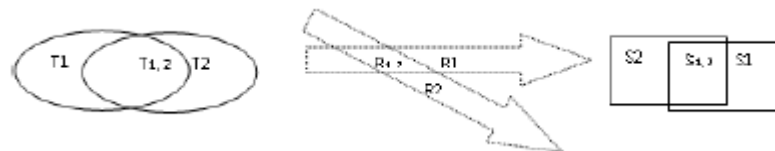


Figure 3. Overlapping Target (T) Role (R) and Subject (S)

A. Overlapping of Subjects

This occurs when the subject of two or more obligations or authority policies overlap, this means that it is expected in some cases the same subject may manage different group of targets. Figure 4 shows that P_1 applies $\{s_1, t_1, r_1\}$ and p_2 applies $\{s_2, t_2, r_2\}$, while there are some subjects $\{s'\}$ are included in both P_1 and P_2 , this means that, the subject of p_1 is $(s' \cup s_1)$ and the subject of P_2 is $(s' \cup s_2)$. Both p_1 and p_2 applies $\{s', t_1, r_1\}$ and $\{s', t_2, r_2\}$ respectively.

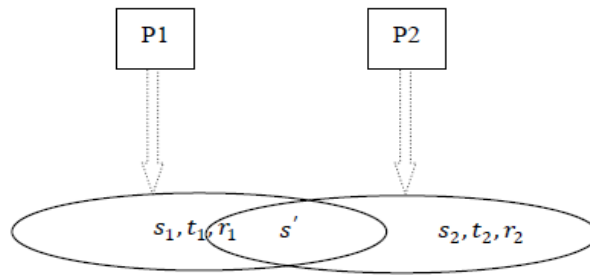


Figure 4. Overlap of Subjects

Example1 in our SMALLS scenario the same manager may enforce two different policies, the first policy to govern a group of Wi-Fi access points, while the other policy is to govern a group of users in the SMALLS active area as follows.

P1: “turn off all the sensors in the supermarket shopping area from 12:00 pm to 7:59 am”

P2: “Users with the description name Security are allowed to perform any action on any resource at anytime from anywhere in the mall”

B. Overlap of Roles

This occurs when the roles of two or more obligations “O” or authority “A” policies overlap, this means that it is expected in some cases the same object may be directed by different actions. The roles of such policies are in conflict if-and-only-if for any two policies $p1$ and $p2$ in one of these forms $\{O-/O+, A-/A+, O+/A-\}$, such that (+) indicates that the policy is permitted and (-) indicates that the policy is forbidden. Figure 5 shows that P1 applies $\{s1, t1, r1\}$ and $p2$ applies $\{s2, t2, r2\}$, while there are some roles $\{r'\}$ are included in both P1 and P2, this means that, the role of $p1$ is $(r' \cup r1)$ and the role of P2 is $(r' \cup r2)$. Both $p1$ and $p2$ applies $\{s1, t1, r'\}$ and $\{s2, t2, r'\}$ respectively.

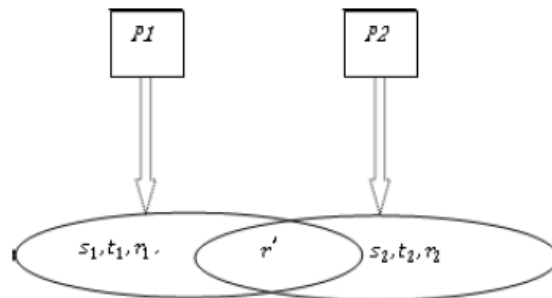


Figure 5. Overlap of Roles

C. Overlap of Targets

Similarly, when the targets of two or more obligations “O” or authority “A” policies overlap, means that it is expected in some cases the same target may be managed by different set of policies. The targets of such policies are in conflict when there are some constraints on the target. Figure 6.6 shows that P1 applies $\{s1, t1, r1\}$ and $p2$ applies $\{s2, t2, r2\}$, while there are some

targets $\{t'\}$ are included in both P1 and P2, this means that, the role of p1 is $(t' \cup t1)$ and the role of P2 is $(t' \cup t2)$. Both p1 and p2 applies $\{s1, t', r1\}$ and $\{s2, t', r2\}$ respectively.

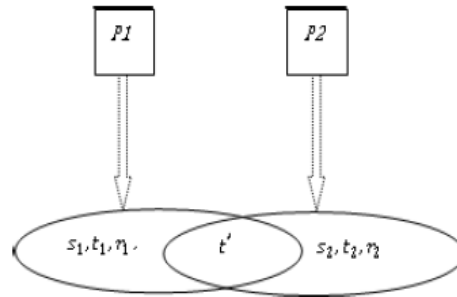


Figure 6. Overlap of Targets

5. CONFLICT ANALYSIS

This work is concentrating on static conflict analyses which assist to specify policies, roles and relationships. In the following sections, we show how to avoid the modality conflicts, inconsistency and multi manager conflicts.

We discuss overlap as the most important factor to policy conflicts. There are several possibilities for overlapping between policies (see figure 7), triple overlap “the set of subjects, targets and actions, of two or more policies with modality of opposite sign to the same subjects, actions and targets overlap”; double overlap “both the subjects and the target of the policies overlap”; subjects-targets overlap “the subjects of one policy and the target of another policy overlap”, target overlap and subjects overlap.

Modality conflict is expected when there is a triple overlap; here we give some example using our case study to show that modality conflict can arise due to different modalities in the set of policies.

In SMALLS system there are different managers (managers and coordinators), managers’ tasks are coordinated by a coordinator. Moreover, each manager has some responsibilities such as manages and governs system services (positioning service), resources (Wi-Fi access points, computer servers and smart phones). Both managers and coordinators use policies to manage and control the system. However some policies enforced by coordinators may restrict the managers from performing their tasks, and managers’ policies may restrict each other. When two or more policies applying to a tuple, there is a potential conflict and the policies can be checked to see whether there is an actual conflict “positive and negative policy with the same subjects, targets and actions”.

Let $S = \{s0, s1, s2, \dots, sn\}$ be a list of subjects, $T = \{t0, t1, t2, \dots, tm\}$ be a list of targets and $R = \{r0, r1, r2, \dots, rk\}$ be a list of roles in the system. For any two policies P1 and P2, let P1 being negative and P2 is positive. Figure 7 below show the overlapping between $p1(s1, t1, r1)$ and $p2(s2, t2, r2)$ for common subjects, actions and roles.

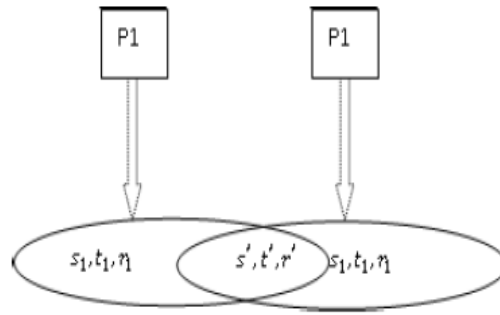


Figure 7. Triple Overlap Between Two Policies

Modality conflicts are expected if and only if a triple overlap between the policies -P1 and +P2 occurs and create the following tuples to which different sets of policies apply: P1 applies $(s_1, t_1, r_1) + (s', t', r')$, while P2 applies $(s_2, t_2, r_2) + (s', t', r')$, what mean that both P1 and P2 applies (s', t', r') . Example2, in SMALLS system there will be some policies pertaining to all staff and customers as well as some more specific policies related to part of the staff. Let us assume P1 and P2 be three polices assigned by different managers.

P1: “Users are not authorized to send requests or perform any action on any resource from 12:00 pm to 7:59 am”

P2: “Users with the description name ADMIN are allowed to perform any action on any resource at any time”

Modality conflicts can be avoided either by changing one policy or block system managers from the managed objects [6]. However, changing policies is not desired in the system, due to the fact that rewriting a policy is time consuming and may not be convenient or desirable in the general case. Blocking system managers means preventing them from controlling objects for a while, however this way is not desirable because the system is completely governed by policies enforced by managers. Thus, the best way to solve the conflicts between P1 and P2 in example 2, is allowing both of them as they enforced and determine which policy should be enforced first. Therefore we can identify the priority for each single policy of each conflicting pair. The assigned priority helps to determine which policy should be ignored when at least one of the policies actions is constrained

Definition 1, Definition 6.1 let $\rho_i = \{pr_i, ei, ci, ai\}$ and $\rho_j = \{pr_j, ej, cj, aj\}$ are two policies, where pr is the priority, e is the event, c is the condition and a is the action. The action $x.m$ means the message m is sent to object x, where $ai = xi .mi$ and $aj = xj .mj$. We assume ρ_i and ρ_j which have simple actions are enforced by manager Mgrk, and $Trig\rho_i$ denotes the triggering of policy ρ_i .

$$Trig\rho_i \Rightarrow (ei \wedge ci) \wedge O(\neg ei) \dots \dots \dots (1)$$

Formula (6.1) illustrates that policy ρ_i is triggered if and only if its event and conditions are true, $O(\neg ei)$ is to reset event after enforcing the policy.

To identify where a conflicts occurs and where potential conflict should be resolved in a set of policies, we should enumerate all subjects, targets which have a different set of policies applying

to them. Moreover, overlapping area should be determined explicitly. Therefore, the simultaneous triggering of p_i and p_j policies should be investigated. The LTL formula 2 requires policies p_i and p_j not to be triggered simultaneously by enforcing the higher priority policy first.

$$[(pr_i > pr_j) ? p_i : p_j] \text{-----} (2)$$

The overlap detection algorithm in figure 8, marks the triggered events of all managers to prevent calling the conflicting rules twice.

```

Algorithm Modality_Conflict (q ,P[],Ev[], Dom[])
1: Let Pr: array[1..MaxDomSize] of
   priority;
2: Let overlap:=false; mc:=0;
3: while q is not empty
4: trigger(Ev[i], Ev[j], Ev[k] ); //push
   event in the queue q
5: Let Ev[i]:=i;
6: Let Ev[j]:=j;
7: Let Ev[k]:=k; //mark triggered events
8: for all m in the domains do
9: Trigger(p[i], p[i+1]);
10:if pr[i+1]>pr[i] then
11:Let overlap:=true;
12:Let mc:=mc+1; //increment of
   conflicts mc
13:end if;
14:end for;
15:Let Ev[i]:=i+1; Ev[j]:=j+1; Ev[k]:=k+1;
16:end while;
17:return(mc);

```

Figure 8. Overlap Detection Algorithm

6. RESULTS AND DISCUSSION

The algorithm in figure 8, was executed using ponder language under the Linux redhat operating system, for three sets containing 100 "SecModule", 150 "LocModule", 50 "SenModule" rules. The output reports 91.8% of the conflicts between managers. The execution was repeated for different number of policies.

Each of the evaluation was measured four times, assuming the number of policies in the location is the same throughout the execution. In figure 9, we can see that the average time required for the 4 times executions according to the execution stages are as follow:

- generate the object file 0.7888s,
- send a query to managers 0.5278s,
- retrieve context information 0.5677s, and
- send back result to the mobile 0.575s.

The amount of time required to perform static conflict avoidance at compile time is 2.46s.

The evaluation result shows that the performance of PobMC is better than the previous works. Less than a second was the enough to perform every task as individual. Furthermore, by this evaluation, it is possible to compare PobMC to other existing approaches in term of its avoiding overlap and policy conflicts.

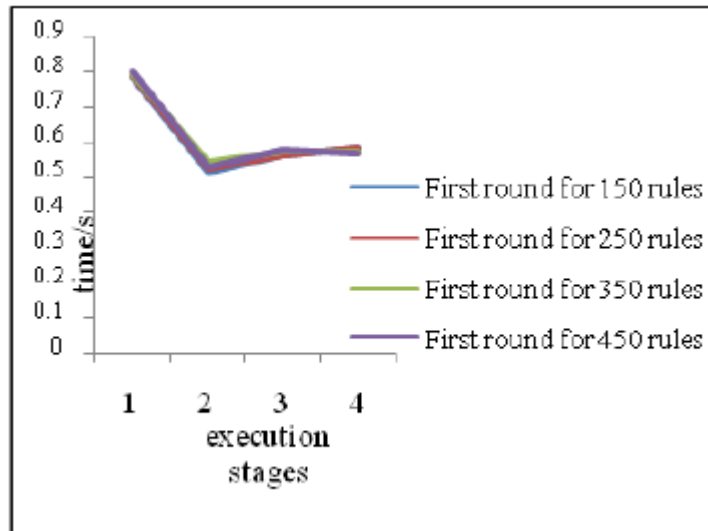


Figure 9. Performance Results for Overlap Algorithm

7. RELATED WORK

There are some techniques to static conflict detection discussed in the literature. Shiva [7] proposed an extended model of Event-Condition-Action (ECA) called ECA-Post condition to enable developers and administrators to annotate actions with their effects. The ECA-P framework uses static and dynamic conflict detection techniques to detect failure in policy execution by using post condition to verify successful completion of policy actions. However, Policy actions may not execute to completion due to various reasons such as changing active space configuration, device and component failure or software errors. Also Khakpour et al. [8] presented an analysis using Rebeca [9] which is an actor-based language for modelling concurrent asynchronous systems which allows to model the system as a set of reactive objects called rebecs, interacting by message passing. In order to introduce this, a new classification of conflicts may occur during governing policies. Moreover, they introduced a number of correctness properties of the adaptation process in the context of their models. Then, they used static analysis of adaptation policies in addition to model checking technique to verify those properties. While their system includes many different managers, there may be more than event,

While a considerable attempt at static and dynamic conflict detection has been presented in previous work, the very complex and crucial issue of dynamic conflict detection in policy-based management has gone largely unresolved. Moreover, current research has revealed that there is still a large class of policy conflict which simply cannot be determined statically. The current state of the art in policybased approach suffers from two main limitations. Firstly, they have

limited ways of detecting and resolving conflicts in policies. Secondly, they do not have mechanisms to ensure that policies are enforced or executed correctly. These limitations severely limit the effectiveness of policies as a way of managing ubiquitous computing environments.

One approach to avoid conflicts in authorization rules is presented by Yu et al, in [10]. They argue that a large number of rules may apply to a service and detecting and resolving conflicts in real time can be a daunting task. However, their system is completely static and assumes that it is always possible to determine priorities ahead of time and avoid conflicts. Another approach for avoiding conflicts in policy specification is proposed by Agrawal, et al, for defining authorization policies for Hippocratic databases [11-13]. Their system allows system administrators to specify system policies for administration and regulatory compliance and these policies have the highest priority. Moreover, the system allows users to manage their privacy preference as their policies do not conflict with the system policies.

While a considerable attempt at static and dynamic conflict detection has been presented in previous work, the very complex and crucial issue of dynamic conflict detection in policy-based management has gone largely unresolved. Moreover, current research has revealed that there is still a large class of policy conflict which simply cannot be determined statically. The current state of the art in policy based approach suffers from two main limitations. Firstly, they have limited ways of detecting and resolving conflicts in policies. Secondly, they do not have mechanisms to ensure that policies are enforced or executed correctly. These limitations severely limit the effectiveness of policies as a way of managing ubiquitous computing environments.

In our framework, the potential cycles specified and avoided earlier since the design time, here most of the requirement can be detected and catch during the analysis. The users policies may override other polices or be overridden based on context information.

8. CONCLUSION AND FUTURE WORK

The analysis of the policy conflicts is implemented using policy specification language called PONDER to check and detect policy cycles. Same as other existing approaches described in sthis paper, PobMC is evaluated using SMALL case study, based on PobMC modeling and policy conflict results. Our experiments show that the PobMC framework leads to effective policy-based management and is a feasible approach. In addition, our evaluation with PobMC has the ability to enhance the existing approaches to support software adaptation. PobMC which enables the coordination among system managers in order to adapt to system changes and avoid the potential overlap is the main contribution of this paper.

Our future work includes the static analysis to avoid inconsistency when a set of rules is enforced by different managers which are managing the same system.

REFERENCES

- [1] E. Lupu, et al., "Autonomous pervasive systems and the policy challenges of a small world!," in 8th IEEE International Workshop on Policies for Distributed Systems and Networks, POLICY 2007, June 13, 2007 - June 15, 2007, Bologna, Italy, 2007, pp. 3-7.
- [2] E. Lupu and M. Sloman, "Conflict analysis for management policies," 1997, pp. 430-443.
- [3] E. Lupu and M. Sloman, "A policy based role object model," 1997, p. 36.

- [4] C. Shankar and R. Campbell, "A Policy-based Management Framework for Pervasive Systems using Axiomatized Rule-Actions," in *Network Computing and Applications*, Fourth IEEE International Symposium on, 2005, pp. 255-258.
- [5] A. Mansor, et al., "Policy-based Approach for Dynamic Architectural Adaptation: A Case Study on Location-Based System," pp. 171-176, 12-14 December 2011.
- [6] E. C. Lupu and M. Sloman, "Conflicts in policy-based distributed systems management," *Software Engineering*, IEEE Transactions on, vol. 25, pp. 852-869, 1999.
- [7] S. Chetan Shiva, et al., "An ECA-P policy-based framework for managing ubiquitous computing environments," in *The Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services*, 2005. *MobiQuitous 2005.* , 2005, pp. 33-42.
- [8] N. Khakpour, et al., "Formal analysis of policy-based self-adaptive systems," in *25th Annual ACM Symposium on Applied Computing*, SAC 2010, March 22, 2010 - March 26, 2010, Sierre, Switzerland, 2010, pp. 2536-2543.
- [9] M. Sirjani, et al., "Modeling and verification of reactive systems using Rebeca," *Fundamenta Informaticae*, vol. 63, pp. 385-410, 2004.
- [10] W. D. Yu and E. Nayak, "An algorithmic approach to authorization rules conflict resolution in software security," 2008, pp. 32-35.
- [11] R. Agrawal, et al., "Managing disclosure of private health data with hippocratic databases," *IBM Research White Paper*, Januray, 2005.
- [12] D. Agrawal, et al., "Policy-based management of networked computing systems," *Communications Magazine*, IEEE, vol. 43, pp. 69-75, 2005.
- [13] R. Agrawal, et al., "Extending relational database systems to automatically enforce privacy policies," 2005, pp. 1013-1022.

CONTROL OF LINEAR SYSTEMS USING DYNAMIC OUTPUT CONTROLLERS

Anna Filasova and Dusan Krokavec

Department of Cybernetics and Artificial Intelligence, Technical University of
Košice, Faculty of Electrical Engineering and Informatics,
Letna 9, 042 00 Kosice, Slovakia

ABSTRACT

The paper deals with the problem of control of continuous-time linear systems by the dynamic output controllers of order equal to the plant model order. The design procedure is based on a solution of the set of linear matrix inequalities and ensures the closed-loop stability using Lyapunov approach. Numerical examples are given to illustrate the design procedure and relevance of the methods as well as to validate the performances of the proposed approach

KEYWORDS

Linear systems, dynamic output controllers, Lyapunov function, linear matrix inequality

1. INTRODUCTION

In practice, online measurements of all state variables of a process are rarely available and since only their observable outputs are accessible, the output feedback control laws have to be considered. Since, really, the system dynamic may be affected by unmeasurable disturbances the H_∞ approach is proposed to be used in the static and dynamic output feedback control law design.

The static output feedback problem seems to be one of the most important questions in linear control system design, see, e.g. [1], [2], [3], [4] and the reference therein. Because of the importance of these kind control systems, considerable attention was dedicated to the study of suitable design methods. Reflecting the fact that the static output feedback stabilization is a concave-convex problem [5], the design conditions based on solution of various mutually coupled matrix equations or coupled linear matrix inequalities (LMI) was discussed, e.g., in [6], [7].

Exploiting the approaches which potentially allow converting dynamic output controller synthesis into an LMI optimization problem, LMI computational technique has brought a tool to solve also this task. An iterative algorithm for designing the linear time-invariant dynamic output controllers of the prescribed structure was presented in [8], formulating the solution as an optimization based on LMIs in which either the Lyapunov matrix or the controller parameter matrix are alternately regarded as the optimization variables. Another iterative approach was proposed in [9], where a convexifying function is reduced in each iteration step to zero to guaranty the feasibility of the problem. Applying the controller parameter transformation and a mix of performance measures, the no recursive approach noted as the multi-objective synthesis of linear dynamic output-feedback controllers is presented in [10], [11] where each objective is formulated relative to a

variety of the closed-loop transfer function and more relaxed sufficient conditions are derived in terms of LMIs.

The aim of this paper is to compare the existing results in design of non-proper and proper dynamic output controllers, but above all to formulate a new design conditions based on the set LMIs and, as yet, one linear matrix equality (LME) the non-proper dynamic output control as well as to extend the methodology for the proper dynamic output control [12]. Applying to the multi-input and multi-output linear systems convexifying assumptions are solved by modifying the H_2 control problem. The stability of the closed-loop system is ensured by finding a suitable Lyapunov matrix within a resolution of the proposed LMIs and LME structure.

The paper is organized in six sections. Following the introduction in Sec. 1, the considered structures of the dynamic output controllers are presented in Sec. 2. The main results are outlined in Sec. 3 and 4, formulating stability analysis and suitable design methods for the given types of output control by use of LMIs. In Sec. 5 the numerical example is given in order to discuss the performances and limitations of the proposed design methods and the last section draws some concluding remarks.

Throughout the paper, the notations are narrowly standard in such a way that x^T , X^T denotes the transpose of the vector x and matrix X , respectively, $X = X^T > 0$ means that X is a symmetric positive definite matrix, $rank(\cdot)$ remits the rank of a matrix, $diag[\cdot]$ designates a block diagonal matrix, the symbol I_n indicates the n -th order unit matrix, R denotes the set of real numbers and R^n , $R^{n \times r}$ refer to the set of all n -dimensional real vectors and $n \times r$ real matrices, respectively.

2. PROBLEM FORMULATION

The systems under consideration are continuous-time linear MIMO systems, described in the state-space form by the set of equations

$$\dot{q}(t) = Aq(t) + Bu(t) \quad (1)$$

$$y(t) = Cq(t) \quad (2)$$

where $q(t) \in R^n$, $u(t) \in R^r$ and $y(t) \in R^m$ are vectors of the system, input and output variables, respectively and the matrices $A \in R^{n \times n}$, $B \in R^{n \times r}$, $C \in R^{m \times n}$ are real matrices, provided that (A, B) is stabilisable and (A, C) is detectable.

It is assumed that the system is stabilized by the full order time-invariant be-proper dynamic output controller

$$\dot{p}(t) = Jp(t) + Ly(t) \quad (3)$$

$$u(t) = Mp(t) + Ny(t) \quad (4)$$

and by the full order time-invariant strictly proper dynamic output controller

$$\dot{p}(t) = Jp(t) + Ly(t) \quad (5)$$

$$u(t) = Mp(t) \quad (6)$$

where $p(t) \in R^n$ is the vector of the controller state variables, the controller matrices

$$K^o = \begin{bmatrix} J & L \\ M & N \end{bmatrix}, \quad K^* = \begin{bmatrix} J & L \\ M & 0 \end{bmatrix} \quad (7)$$

$K^o \in R^{(n+r) \times (n+m)}$, $K^* \in R^{(n+r) \times (n+m)}$ has the prescribed structure with respect to the real matrices $J \in R^{n \times n}$, $L \in R^{n \times m}$, $M \in R^{r \times n}$ and $N \in R^{r \times m}$ or $0 \in R^{r \times m}$.

Considering that the plant (1), (2) is square, i.e., $r = m$, the objective is to present design conditions to expose the above described matrix parameters of the dynamic controllers.

3. BI-PROPER DYNAMIC OUTPUT CONTROLLER

To analyse the stability of the closed-loop system structure with the bi-proper dynamic output controller, the following form of the closed-loop system description can be introduced

$$\begin{bmatrix} \dot{q}(t) \\ \dot{p}(t) \end{bmatrix} = \begin{bmatrix} A + BNC & BM \\ LC & J \end{bmatrix} \begin{bmatrix} q(t) \\ p(t) \end{bmatrix} \quad (8)$$

$$y(t) = \begin{bmatrix} 0 & I_m \end{bmatrix} \begin{bmatrix} 0 & I_n \\ C & 0 \end{bmatrix} \begin{bmatrix} q(t) \\ p(t) \end{bmatrix} \quad (9)$$

After introducing the notations

$$q^{oT}(t) = \begin{bmatrix} q^T(t) & p^T(t) \end{bmatrix} \quad (10)$$

$$A^o = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix}, \quad B^o = \begin{bmatrix} 0 & B \\ I_n & 0 \end{bmatrix}, \quad C^o = \begin{bmatrix} 0 & I_n \\ C & 0 \end{bmatrix}, \quad I^o = \begin{bmatrix} 0 & I_m \end{bmatrix} \quad (11)$$

where, in general, $A^o \in R^{2n \times 2n}$, $B^o \in R^{2n \times (n+r)}$, $C^o \in R^{(n+m) \times 2n}$, $I^o \in R^{m \times (n+m)}$, the closed-loop state-space equations takes the form

$$\dot{q}^o(t) = (A^o + B^o K^o C^o) q^o(t) \quad (12)$$

$$y^o(t) = I^o C^o q^o(t) \quad (13)$$

In the sequel, so it is supposed that (A^o, B^o) is stabilizable, (A^o, C^o) is detectable [15] and the matrix product $C^o B^o$ is nonzero matrix.

Theorem 1

The closed-loop system consisting of the plant (1), (2) with the be-proper dynamic output controller (3)-(4) is stable if there exist a symmetric positive definite matrix $Q^o \in R^{2n \times 2n}$, a regular matrix $H^o \in R^{(n+m) \times (n+m)}$ and a matrix $Y^o \in R^{(n+r) \times (n+m)}$ such that

$$Q^o = Q^{oT} > 0 \quad (14)$$

$$A^o Q^o + Q^o A^{oT} + B^o Y^o C^o + C^{oT} Y^{oT} B^{oT} < 0. \quad (15)$$

$$C^o Q^o = H^o C^o \quad (16)$$

When the above conditions hold, the common control law gain matrix is given by the equation

$$K^o = Y^o (H^o)^{-1} \quad (17)$$

Proof: Defining the Lyapunov function as follows

$$v(q^o(t)) = q^{oT}(t) P^o q^o(t) > 0 \quad (18)$$

where $P^o = P^{oT} > 0$, then

$$\dot{v}(q^o(t)) = \dot{q}^{oT}(t) P^o q^o(t) + q^{oT}(t) P^o \dot{q}^o(t) < 0 \quad (19)$$

Substituting (12) and (13) into (19) it yields

$$\dot{v}(q^o(t)) = q^{oT}(t) (A_c^{oT} P^o + P^o A_c^o) q^o(t) < 0 \quad (20)$$

where

$$A_c^o = A^o + B^o K^o C^o \quad (21)$$

and (20) implies

$$A_c^{oT} P^o + P^o A_c^o < 0 \quad (22)$$

Since P^o is positive definite, it also yields

$$Q^o A_c^{oT} + A_c^o Q^o < 0 \quad (23)$$

where $Q^o = (P^o)^{-1}$ and writing (23) in the open form, it is obtained

$$Q^o (A^o + B^o K^o C^o)^T + (A^o + B^o K^o C^o) Q^o < 0 \quad (24)$$

Analysing the matrix product in (24) it can be set

$$B^o K^o C^o Q^o = B^o K^o H^o (H^o)^{-1} C^o Q^o \quad (25)$$

where H^o is a regular square matrix of appropriate dimension.

Defining the following equality

$$(H^o)^{-1} C^o = C^o (Q^o)^{-1} \quad (26)$$

and using the notation

$$Y^o = K^o H^o \quad (27)$$

then

$$B^o K^o C^o Q^o = B^o Y^o C^o \quad (28)$$

(24) implies (15) and (26) implies (16). This concludes the proof.

In practice, the case with $r = m$ (square plants) is often encountered, where it is generally associated with each output signal a reference signal, which is expected to influence as wanted this output. Such regime, reflecting nonzero set working points, is called the forced regime.

Definition 1

The forced regime for (1), (2) with the bi-proper dynamic output controller (3),(4) is given by the control policy

$$\dot{p}(t) = Jp(t) + Ly(t) \quad (29)$$

$$u(t) = Mp(t) + Ny(t) + Ww(t) \quad (30)$$

where $r = m$, $w(t) \in R^m$ is desired output signal vector, and $W \in R^{m \times m}$ is the signal gain matrix.

Theorem 2

If the system (1), (2) is stabilizable by the control policy (29), (30) and [13]

$$\text{rank} \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} = n + m \quad (31)$$

then the matrix W in (29), designed by using the static decoupling principle, takes the form

$$W = -(C(A - BMJ^{-1}LC + BNC)^{-1}B)^{-1} \quad (32)$$

Proof: In a steady state which corresponds to $\dot{q}(t) = \mathbf{0}$, $\dot{p}(t) = \mathbf{0}$ the equality $y_o = w_o$ must hold. Denoting $q_o \in R^n$, $y_o, w_o \in R^m$ as the vectors of steady state values of $q(t)$, $y(t)$, $w(t)$, respectively, then (1) – (3) and (30) imply

$$0 = Aq_o + Bu_o \quad (33)$$

$$0 = Jp_o + LCq_o \quad (34)$$

$$y_o = Cq_o \quad (35)$$

$$u_o = Mp_o + Ny_o + Ww_o \quad (36)$$

Since now (34) – (36) implies

$$u_o = (-MJ^{-1}LC + NC)q_o + Ww_o \quad (37)$$

then, substituting (37) into (33), it yields

$$0 = (A - BMJ^{-1}LC + BNC)q_o + BWw_o \quad (38)$$

$$q_o = -(A - BMJ^{-1}LC + BNC)^{-1}BWw_o \quad (39)$$

respectively, and with (35)

$$y_o = -C(A - BMJ^{-1}LC + BNC)^{-1}BWw_o \quad (40)$$

Thus, considering $y_o = w_o$, then (40) implies (32). This concludes the proof.

The W matrix is nothing else than the inverse of the closed-loop static gain matrix. This gain matrix can be obtained so by setting $s = 0$ in the state-space expression of the transfer function matrix of the closed-loop system with respect to the forced input. Note, the static gain realized by the W matrix is ideal in control only if the plant parameters, on which the value of W depends, are known and do not vary with time.

The forced regime is basically designed for constant references and is very closely related to shift of origin. If the command value $w(t)$ is changed "slowly enough," the above scheme can do a reasonable job of tracking, i.e., making $y(t)$ follow $w(t)$ [14].

In most cases the control using the non-proper dynamic output control is practically equivalent to the static output control, as the control law component defined by the output direct part $Ny(t)$ is dominant. So the static output control is preferred, or the proper dynamic output control is fitted.

4. STRICTLY PROPER DYNAMIC OUTPUT CONTROLLER

Considering the strictly proper dynamic output controller, the following form of the closed-loop system description is obtained

$$\begin{bmatrix} \dot{q}(t) \\ \dot{p}(t) \end{bmatrix} = \begin{bmatrix} A & BM \\ LC & J \end{bmatrix} \begin{bmatrix} q(t) \\ p(t) \end{bmatrix} \quad (41)$$

$$y(t) = \begin{bmatrix} 0 & I_m \end{bmatrix} \begin{bmatrix} 0 & I_n \\ C & 0 \end{bmatrix} \begin{bmatrix} q(t) \\ p(t) \end{bmatrix} \quad (42)$$

Using the notations(10), (11) the closed-loop state-space equations takes the form

$$\dot{q}^\circ(t) = A_c^\bullet q^\circ(t) \quad (43)$$

$$y^\circ(t) = I^\circ C^\circ q^\circ(t) \quad (44)$$

where I° , C° are given in (11) and

$$A_c^\bullet = \begin{bmatrix} A & BM \\ LC & J \end{bmatrix} \quad (45)$$

Note, only the square system ($r = m$) is considered in the following.

Theorem 3

The strictly proper dynamic output controller (5), (6) to the system (1), (2) exists if there exist symmetric positive definite matrices $Q_1, S_1 \in R^{n \times n}$ and matrices $L_1 \hat{=} R^{n \times m}$, $M_1 \hat{=} R^{r \times n}$ such that

$$Q_1 = Q_1^T > 0, \quad S_1 = S_1^T > 0, \quad \begin{bmatrix} S_1 & I \\ I & Q_1 \end{bmatrix} > 0 \quad (46)$$

$$AQ_1 + Q_1A^T + BM_1 + M_1^TB^T < 0 \quad (47)$$

$$S_1A + A^TS_1 + L_1C + C^TL_1^T < 0 \quad (48)$$

When the above conditions hold, the control law gain matrices are given as follows

$$J = J_1(R_1 - S_1^{-1})^{-1}, \quad L = -S_1^{-1}L_1, \quad M = M_1(R_1 - S_1^{-1})^{-1} \quad (49)$$

where

$$J_1 = S_1^{-1}(A^T + (S_1 A + L_1 C)Q_1 + S_1 B M_1) \quad (50)$$

Proof: Defining the Lyapunov function as in (18) then analogously can be obtained

$$A_c^{\circ T} P^{\circ} + P^{\circ} A_c^{\circ} < 0 \quad (51)$$

$$Q^{\circ} A_c^{\circ T} + A_c^{\circ} Q^{\circ} < 0 \quad (52)$$

respectively, where $Q^{\circ} = (P^{\circ})^{-1}$. Considering that

$$Q^{\circ} = Q^{\circ T} = \begin{bmatrix} Q_1 & Q_2 \\ Q_2^T & Q_3 \end{bmatrix} > 0 \quad (53)$$

then the Schurov complement of (53) (with respect to Q_3) is

$$S_1^{-1} = Q_1 - Q_2 Q_3^{-1} Q_2^T > 0, \quad Q_3 = Q_3^T > 0 \quad (54)$$

and it can be set

$$Q_1 - S_1^{-1} = Q_2 Q_3^{-1} Q_2^T > 0 \quad (55)$$

Using S_1^{-1} , the following transform matrices can be defined

$$T_1^{\circ} = \begin{bmatrix} S_1 & S_1 \\ I & 0 \end{bmatrix}, \quad T_2^{\circ} = \begin{bmatrix} I & 0 \\ 0 & -Q_2 Q_3^{-1} \end{bmatrix} \quad (56)$$

and it yields

$$\begin{aligned} T_1^{\circ T} T_2^{\circ T} Q^{\circ} T_2^{\circ} T_1^{\circ} &= \\ &= \begin{bmatrix} S_1 & S_1 \\ I & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -Q_2 Q_3^{-1} \end{bmatrix} \begin{bmatrix} Q_1 & Q_2 \\ Q_2^T & Q_3 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -Q_3^{-1} Q_2^T \end{bmatrix} \begin{bmatrix} S_1 & I \\ S_1 & 0 \end{bmatrix} = \\ &= \begin{bmatrix} S_1 & -S_1 Q_2 Q_3^{-1} \\ I & 0 \end{bmatrix} \begin{bmatrix} Q_1 & Q_2 \\ Q_2^T & Q_3 \end{bmatrix} \begin{bmatrix} S_1 & I \\ -Q_3^{-1} Q_2^T S_1 & 0 \end{bmatrix} = \\ &= \begin{bmatrix} S_1 & -S_1 Q_2 Q_3^{-1} \\ I & 0 \end{bmatrix} \begin{bmatrix} I & Q_1 \\ 0 & Q_2^T \end{bmatrix} = \begin{bmatrix} S_1 & I \\ I & Q_1 \end{bmatrix} > 0 \end{aligned} \quad (57)$$

where also (55) yields. Analogously,

$$\begin{aligned} T_1^{\circ T} T_2^{\circ T} A_c^{\circ} Q^{\circ} T_2^{\circ} T_1^{\circ} &= \begin{bmatrix} S_1 & -S_1 Q_2 Q_3^{-1} \\ I & 0 \end{bmatrix} \begin{bmatrix} A & B M \\ L C & J \end{bmatrix} \begin{bmatrix} I & Q_1 \\ 0 & Q_2^T \end{bmatrix} = \\ &= \begin{bmatrix} S_1 A - S_1 Q_2 Q_3^{-1} L C & S_1 B M - S_1 Q_2 Q_3^{-1} J \\ A & B M \end{bmatrix} \begin{bmatrix} I & Q_1 \\ 0 & Q_2^T \end{bmatrix} \end{aligned} \quad (58)$$

and denoting

$$J_1 = Q_2 Q_3^{-1} J Q_2^T, \quad L_1 = -S_1 Q_2 Q_3^{-1} L, \quad M_1 = M Q_2^T \quad (59)$$

(58) can be rewritten as follows

$$T_1^{\circ} T_2^{\circ} A^{\circ} Q^{\circ} T_2^{\circ T} T_1^{\circ T} = \begin{bmatrix} S_1 A + L_1 C & (S_1 A + L_1 C) Q_1 + S_1 B M_1 - S_1 J_1 \\ A & A Q_1 + B M_1 \end{bmatrix} \quad (60)$$

Finally, it yields

$$T_1^{\circ} T_2^{\circ} (A_c^{\circ} Q^{\circ} + Q^{\circ} A_c^{\circ T}) T_2^{\circ T} T_1^{\circ T} = \begin{bmatrix} S_1 A + A^T S_1 + L_1 C + C^T L_1^T & U \\ U^T & A Q_1 + Q_1 A^T + B M_1 + M_1^T B^T \end{bmatrix} < 0 \quad (61)$$

where

$$U = A^T + (S_1 A + L_1 C) Q_1 + S_1 (B M_1 - J_1) \quad (62)$$

Setting $U = 0$, i.e.,

$$S_1 J_1 = A^T + (S_1 A + L_1 C) Q_1 + S_1 B M_1 \quad (63)$$

then (61) imply (47), (48). It is evident that (47), (48) are conditioned by the inequalities (46) and (63) implies (50).

Writing as follows

$$Q_1 - S_1^{-1} = Q_2 Q_3^{-1} Q_2^T = (Q_1 - S_1^{-1})(Q_1 - S_1^{-1})^{-1}(Q_1 - S_1^{-1}) \quad (64)$$

then with respect to (55) it can be set

$$Q_2 Q_3^{-1} Q_2^T = (Q_1 - S_1^{-1})(Q_1 - S_1^{-1})^{-1}(Q_1 - S_1^{-1}) \quad (65)$$

which leads to

$$Q_2 = Q_2^T = Q_1 - S_1^{-1}, \quad Q_3 = Q_1 - S_1^{-1}, \quad Q_2 Q_3^{-1} = I \quad (66)$$

and, using(59), then (59) gives

$$J_1 = J(Q_1 - S_1^{-1}), \quad L_1 = -S_1 L, \quad M_1 = M(Q_1 - S_1^{-1}) \quad (67)$$

Thus, (67) implies (49). This concludes the proof.

Definition 2

The forced regime for (1), (2) with the strictly proper dynamic output controller (5), (6) is given by the control policy

$$\dot{p}(t) = Jp(t) + Ly(t) \quad (68)$$

$$u(t) = Mp(t) + Ww(t) \quad (69)$$

where $w(t) \in R^m$ is desired output signal vector, and $W \in R^{m \times m}$ is the signal gain matrix.

Theorem 4

If the system (1), (2) is stabilizable by the control policy (68), (69) and satisfies (31) then the matrix W in (69), designed by using the static decoupling principle, takes the form

$$W = -(C(A - BMJ^{-1}LC)^{-1}B)^{-1} \quad (70)$$

Proof: Setting $N = 0$ in (32) then (70) is obtained. This concludes the proof.

5. ILLUSTRATIVE EXAMPLE

The features of the considered schemes and the effectiveness of the proposed design conditions are presented using the illustrative example.

The state space representation, describing the chemical reactor model [16], consists of the following matrices

$$A = \begin{bmatrix} 1.380 & -2.080 & 6.715 & -5.676 \\ -0.581 & -4.290 & 0.000 & 0.675 \\ 10.672 & 4.273 & -6.654 & 5.893 \\ 0.482 & 4.273 & 1.343 & -2.104 \end{bmatrix}, \quad B = \begin{bmatrix} 0.000 & 0.000 \\ 5.679 & 0.000 \\ 1.136 & -3.146 \\ 1.136 & 0.000 \end{bmatrix}, \quad C^T = \begin{bmatrix} 2 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and this system in the considered closed loop structures under the non-proper dynamic output control (3), (4) as well as under the proper dynamic output control (5), (6) was used in the presented simulations. Note, the pair (A, B) is controllable and the pair (A, C) is observable.

Within the above system parameters, solving (14)-(16) using the SeDuMi package for Matlab [17], the matrix parameters of the non-proper dynamic output controller were as follows

$$J = \text{diag}[-0.7195 \quad -0.7195 \quad -0.7195 \quad -0.7195], \quad N = \begin{bmatrix} 1.2788 & -3.8548 \\ 2.8385 & -5.9951 \end{bmatrix}$$

$$M = 10^{-10} \begin{bmatrix} -0.1464 & -0.0096 & -0.0244 & -0.0446 \\ -0.1889 & -0.1022 & -0.0432 & -0.0684 \end{bmatrix}, \quad L = 10^{-10} \begin{bmatrix} -0.3131 & 0.5737 \\ -0.0831 & 0.0799 \\ -0.2439 & 0.6788 \\ -0.2996 & 0.8917 \end{bmatrix}$$

and the resulting global closed-loop system eigenvalues spectrum was

$$\rho(A_c^\circ) = \{-0.0387, -0.0461, -0.0752 \pm 0.0770i, -0.0072, -0.0072, -0.0072, -0.0072\}$$

It is evident that in this case the static output control part, determined by the matrix N , is dominant.

Applying the same toolbox to solve LMIs (46)-(48) the obtained set of the proper dynamic controller matrix parameters was as follows

$$J = \begin{bmatrix} -2.0338 & 0.2364 & 11.4248 & -6.4693 \\ -0.2450 & -0.0236 & -1.7606 & -9.5833 \\ -7.0794 & 7.1070 & -7.4777 & -2.5268 \\ 1.1450 & 7.2147 & 4.2238 & -5.9784 \end{bmatrix}$$

$$M = \begin{bmatrix} 0.2797 & 1.0404 & -0.3355 & -1.5332 \\ 2.9243 & -0.3763 & -2.6195 & 2.2187 \end{bmatrix}, \quad L = \begin{bmatrix} 1.2432 & -3.8002 \\ 0.6469 & 2.6497 \\ 0.4235 & 4.7377 \\ -0.7341 & -1.3839 \end{bmatrix}$$

The both dynamic controller design methods, previously described, were applied to the simulation benchmark. The conditions in simulations were specified for system in the forced regimes, where

$$p^T(0) = \mathbf{0}, \quad q^T(0) = [0.1 \quad 0.1 \quad 0.0 \quad 0.0], \quad w^T(t) = [2 \quad 1]$$

and the signal gain matrices W_{dn} , W_{dp} were computed using (32), (70), respectively, as follows

$$W_{dn} = \begin{bmatrix} -1.2128 & 3.8408 \\ -2.3680 & 5.9708 \end{bmatrix}, \quad W_{dp} = \begin{bmatrix} -0.0518 & 0.3335 \\ -0.3510 & 0.3335 \end{bmatrix}$$

Since the same desired output variables have been utilized to assess the each controller ability response and to demonstrate performance with respect to asymptotic properties, the results of the both proposed design method can be immediately compared.

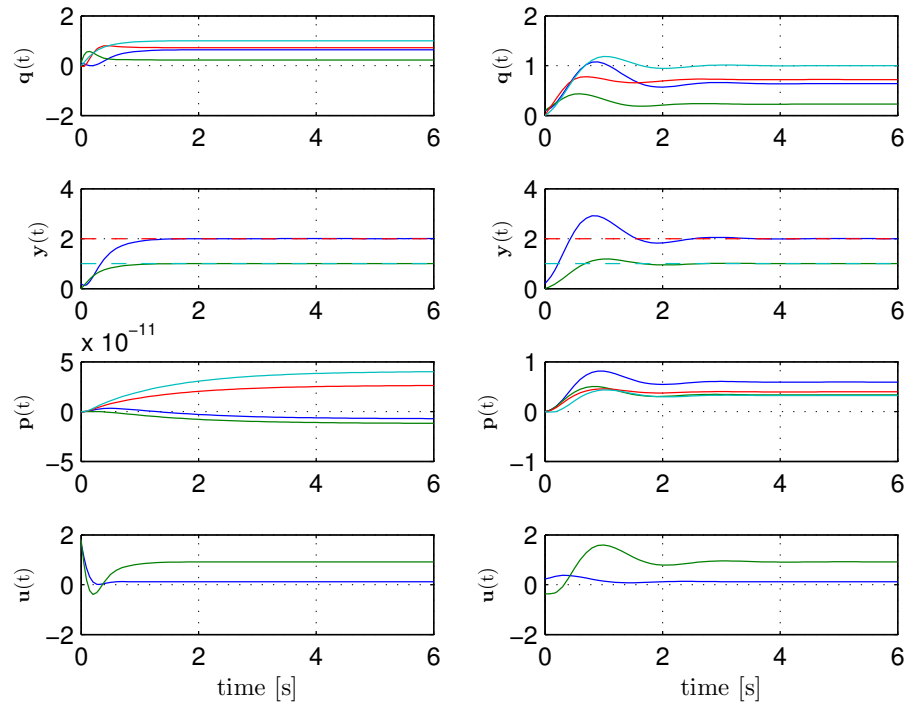


Figure 1: system variable responses using the dynamic output controllers

The left part of Fig. 1 shows the closed-loop system response with the dynamic output controller (3), (4) and the control policy (29), (30), which matrix parameter was obtained solving (14)-(16), (32). Using the dynamic output controller (5), (6) with the gain matrix parameters satisfying the conditions (46)-(48), (70), the right part of Fig. 1 shows the system response of the closed-loop system for the same system initial conditions and the control policy (68), (69). It is obvious from these figures that both controllers which parameters were obtained using the solutions of the LMI

problems specified by Theorem 1 and Theorem 3 can successfully provide for the closed-loop system steady-state properties and asymptotic dynamics.

6. CONCLUDING REMARKS

New approach for output dynamic feedback control design is presented in this paper. By the proposed procedure the control problem is parameterized in such LMIs set with one additional LME which admit more freedom in guaranteeing the output feedback control performance for a bi-proper dynamic controller and by LMIs set only for a strictly proper dynamic output controller. Sufficient conditions of the controller existence manipulating the stability of the closed-loop systems imply the control structure, which stabilize the system in the sense of Lyapunov and the controller design tasks is a solvable numerical problem. An additional benefit of the method is that controller uses minimum feedback information with respect to desired system output and the approach is enough flexible to allow the inclusion of additional design condition bounds.

ACKNOWLEDGMENTS

The work presented in the paper was supported by VEGA, the Grant Agency of the Ministry of Education and the Academy of Science of Slovak Republic, under Grant No. 1/0348/14. This support is very gratefully acknowledged.

REFERENCES

- [1] Chadli, M., Maquin, D. & Ragot, J. (2002) "Static output feedback for Takagi-Sugeno systems: An LMI approach", In Proceedings of the 10th Mediterranean Conference on Control and Automation MED2002, Lisbon, Portugal.
- [2] Syrmos, V.L., Abdallah, C. & Dorato, P. (1994) "Static output feedback: A survey," In Proceedings of the 33rd IEEE Conference on Decision and Control, Lake Buena Vista, FL, USA, pp. 837-842.
- [3] Syrmos, V.L., Abdallah, C., Dorato, P. & Grigoriadis, K. (1997) "Static output feedback: A survey", *Automatica*, Vol. 33, No. 2, pp. 125-137.
- [4] Krokavec, D. & Filasova, A. (2013) "Design of fault residual functions for systems stabilized by static output feedback", In Proceedings of the 2nd International Conference on Control and Fault-Tolerant Systems SysTol'13, Nice, France, pp. 596-600.
- [5] Astolfi, A. & Colaneri, P. (2004) "The static output feedback stabilization problem as a concave-convex programming problem", In Proceedings of the 2004 American Control Conference, Boston, MA, USA, pp. 2139-2141.
- [6] Lewis, F.L. & Syrmos, V.L. (1995) *Optimal Control*, New York, NY, USA, John Wiley & Sons.
- [7] Crusius, C.A.R. & Trofino, A. (1999) "Sufficient LMI conditions for output feedback control problems", *IEEE Transactions on Automatic Control*, Vol. 44, No. 5, pp. 1053-1057.
- [8] El Ghaoui, L. & Balakrishnan, V. (1994) "Synthesis of fixed-structure controllers via numerical optimization", In Proceedings of the 33rd Conference on Decision and Control, Lake Buena Vista, FL, USA, pp. 2678-2683.
- [9] Oliveira de, M.C., Camino, J.E. & Skelton, R.E. (2000) "A convexifying algorithm for the design of structured linear controllers", In Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, pp. 2781-2786.
- [10] Scherer, C., Gahinet, P. & Chilali, M. (1997) "Multiobjective output-feedback control via LMI optimization", *IEEE Transactions on Automatic Control*, Vol. 42, No. 7, pp. 896-911.
- [11] Jungers, M., Castelan, E.B., Moraes, V.M. & Moreno, U.F. (2013) "A dynamic output feedback controller for NCS based on delay estimates", *Automatica*, Vol. 49, No. 3, pp. 788-792.
- [12] Krokavec, D. & Filasova, A. (2007) *Dynamic System Diagnosis*. Kosice, Slovakia, Elfa. (in Slovak)
- [13] Wang, Q.G. (2003) *Decoupling Control*. Berlin, Germany, Springer-Verlag.
- [14] Kailath, T., (1980) *Linear Systems*, Englewood Cliffs, NJ, USA, Prentice-Hall.

- [15] Doyle, J.C., Glover, K., Khargonekar, P.P. & Francis, B.A. (1989) "State-space solutions to standard H_2 and H_∞ control problems", IEEE Transactions on Automatic Control, vol. 34, no. 8, pp. 831-847, 1989.
- [16] Kautsky, J., Nichols, N.K. & Van Dooren, P. (1985) "Robust pole assignment in linear state feedback", International Journal of Control, Vol. 41, No.5, pp. 1129-1155.
- [17] Peaucelle, D., Henrion, D., Labit, Y. & Taitz, K. (2002) User's Guide for SeDuMi Interface 1.04, Toulouse, France, LAAS-CNRS.

AUTHORS

Anna Filasova graduated in technical cybernetics and received M.Sc. degree in 1975, and Ph.D. degree in 1993 both from the Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Slovakia. In 1999 she was appointed Associated Professor from the Technical University in Kosice in technical cybernetics. She is with the Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, and she has been working with as an Assistant Professor from 1975 to 1999. Her main research interests are in robust and predictive control, decentralized control, large-scale system optimization, and control reconfiguration.

Dusan Krokavec received M.Sc. degree in automatic control in 1967 and Ph.D. degree in technical cybernetics in 1982 from the Faculty of Electrical Engineering, Slovak University of Technology in Bratislava, Slovakia. In 1984 he was promoted Associated Professor from the Technical University in Kosice, Slovakia, and in 1999 he was appointed Full Professor in automation and control. He is with the Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Kosice. In the long term, he specializes in stochastic processes in dynamic systems, digital control systems and digital signal processing, and in dynamic system fault diagnosis.

ON OBSERVER DESIGN METHODS FOR A CLASS OF TAKAGI-SUGENO FUZZY SYSTEMS

Dusan Krokavec and Anna Filasova

Department of Cybernetics and Artificial Intelligence, Technical University of Kosice, Faculty of Electrical Engineering and Informatics, Letna 9, 042 00 Kosice, Slovakia

ABSTRACT

The generalized design principle of TS fuzzy observers for one class of continuous-time nonlinear MIMO systems is presented in this paper. The problem addressed can be indicated as a descriptor system approach to TS fuzzy observers design, implying the asymptotic convergence of the state observer error. A new structure of linear matrix inequalities is outlined to possess the observer asymptotic dynamic properties closest to the optimal.

KEYWORDS

Thau observer, TS fuzzy observer, convex optimization, linear matrix inequalities

1. INTRODUCTION

As is well known, observer design is a hot research field owing to its particular importance in observer-based control, and fault diagnosis. The nonlinear system theory using Lipschitz conditions has emerged as a method capable of use in state estimation design for nonlinear systems [1], although Lipschitz condition is a restrictive condition which many classes of systems may not satisfy. However this principle used in state estimators design means that the observer satisfies a sufficient condition for the asymptotic stability of error system, but in fact there is not a straightforward method for selecting the observer gain to fill such conditions [2]. Because the Takagi-Sugeno (TS) fuzzy approach is a suitable representation of certain class of nonlinear dynamic systems [3], employing the fuzzy modelling approach to approximate sector-bounded nonlinear systems, other well-known nonlinear observers are based on Takagi-Sugeno (TS) fuzzy models [4], [5]. To design TS fuzzy observers, usually the technique utilizing the linear matrix inequalities is used [6].

Although the state observers for linear and nonlinear systems received considerable attention, the descriptor design principles have not been studied extensively. Adapting the descriptor observer design principle [7], the first result giving sufficient design conditions, but for linear time-delay systems, can be found in [8]. Reflecting the same problems concerning the observers for descriptor time-delay nonlinear systems represented by TS fuzzy models, an LMI method was presented in [9], but a hint of this methodology can be found only in [10].

Adapting the results on the TS fuzzy observers for bilinear systems [11] as well as their potential extensions, the main issue of this paper is to use the descriptor principle in TS fuzzy observer design. Preferring LMI formulation, although partly conservative, the stability condition proofs use standard arguments on H_2 approach to obtain the design conditions requiring only solving of LMIs without additional constraints. To the best author's knowledge, the proposed LMI structure in design conditions formulation were not fully addressed yet in the previous works.

The paper is organized as follows. In Sec. 2, the TS fuzzy model is briefly described and the TS fuzzy observer design problem for given class of nonlinear systems is formulated in Sec. 3. The new LMI structure, describing the TS fuzzy observer design conditions, is presented in Sec. 4 and analysed and algorithmically explained in Sec 5. Finally, Sec. 6 draws conclusions and some future directions.

The notations throughout the paper are narrowly standard in such a way that \mathbf{x}^T , \mathbf{X}^T denotes the transpose of the vector \mathbf{x} and matrix \mathbf{X} , respectively, $\mathbf{X} = \mathbf{X}^T > 0$ means that \mathbf{X} is a symmetric positive definite matrix, the symbol \mathbf{I}_n indicates the n -th order unit matrix, R denotes the set of real numbers R^n , and $R^{n \times r}$, refer to the set of all n -dimensional real vectors and $n \times r$ real matrices, respectively.

2. TAKAGI-SUGENO FUZZY MODELS

The systems under consideration are from the class of multi-input and multi-output nonlinear (MIMO) continuous-time dynamic systems, represented in TS form as

$$\dot{q}(t) = \sum_{i=1}^s h_i(\theta(t))(A_i q(t) + B_i u(t)) \quad (1)$$

$$y(t) = Cq(t) \quad (2)$$

where $q(t) \in R^n$, $u(t) \in R^r$, $y(t) \in R^m$, are vectors of the state, input, and output variables, $A \in R^{n \times n}$, $B \in R^{n \times r}$, $C \in R^{m \times n}$ are real finite values matrices, $m, r < n$ and $h_i(\theta(t))$ is averaging weight for the i -th rule, representing the normalized grade of fuzzy membership (membership function). By definition, the membership functions satisfy the convex sum properties

$$0 \leq h_i(\theta(t)) \leq 1, \quad \sum_{i=1}^s h_i(\theta(t)) = 1 \quad \text{for all } i \in \{1, \dots, s\} \quad (3)$$

where s is the number of linear models (fuzzy rules) and

$$\theta(t) = [\theta_1(t) \quad \theta_2(t) \quad \dots \quad \theta_p(t)] \quad (4)$$

is p dimensional vector of the premise variables. It is assumed that the premise variable is a system state variable, or a measurable external variable, while none of the premise variables does not depend on any element of the input variables vector $u(t)$. In the above sense, the fuzzy model of a system can be interpreted as a combination of s linear models through the set of membership functions $\{h_i(\theta(t)), i = 1, 2, \dots, s\}$. More details can be found, e.g., in [6], [12].

It is supposed that the couples (A_i, C) are observable for all $i = 1, 2, \dots, s$, as well as the matrix C occurs in all local models and the number of input variables r is equal to the number of output variables m (the dynamic system is a square system).

3. TAKAGI-SUGENO FUZZY OBSERVER DESIGN

The conventional fuzzy observer can be constructed as follows

$$\dot{q}_e(t) = \sum_{i=1}^s h_i(\theta(t))(A_i q_e(t) + B_i u(t) + J_i(y(t) - y_e(t))) \quad (5)$$

$$y_e(t) = C q_e(t) \quad (6)$$

where $q_e(t) \in R^n$ is estimation of the system state vector (the fuzzy observer state vector) and $J_i \in R^{n \times m}$, $i=1,2,\dots,s$ is the set of the observer gain matrices.

Lemma 1

The fuzzy observer (5), (6) is stable if there exist a positive definite symmetric matrix $P \in R^{n \times n}$ and matrices $Y_i \in R^{n \times m}$ such that for all $i=1,2,\dots,s$

$$P = P^T > 0 \quad (7)$$

$$P A_i + A_i^T P - Y_i C - C^T Y_i^T < 0 \quad (8)$$

When the above conditions hold, i.e. if Y_i and the non-singular matrix P are solutions of (7), (8), the set of the observer gain matrices J_i is given by the following equations

$$J_i = P^{-1} Y_i \quad (9)$$

Proof: (compare, e.g., [11]) Introducing the error between the system state vector and the observer state vector as follows

$$e(t) = q(t) - q_e(t) \quad (10)$$

and performing the time derivative of the error $e(t)$, then exploiting (1) and (10) it is

$$\dot{e}(t) = \dot{q}(t) - \dot{q}_e(t) = \sum_{i=1}^s h_i(\theta(t))(A_i(q(t) - q_e(t)) - J_i(y(t) - y_e(t))) \quad (11)$$

which can be written using (2), (11) as follows

$$\dot{e}(t) = \sum_{i=1}^s h_i(\theta(t)) A_{ei} e(t) \quad (12)$$

where

$$A_{ei} = A_i - J_i C \quad (13)$$

Defining the Lyapunov function of the form

$$v(e(t)) = e^T(t) P e(t) > 0 \quad (14)$$

where $P = P^T > 0$, then evaluating the time derivative of (14) it yields

$$\dot{v}(e(t)) = \dot{e}^T(t)P e(t) + e^T(t)P \dot{e}(t) \quad (15)$$

Substituting (12), (13) into (15) gives

$$\dot{v}(e(t)) = e^T(t) \sum_{i=1}^s h_i(\theta(t)) (P(A_i - J_i C) + (A_i - J_i C)^T P) e(t) \quad (16)$$

$$P(A_i - J_i C) + (A_i - J_i C)^T P < 0 \text{ for all } i \quad (17)$$

respectively.

Therefore, setting

$$P J_i = Y_i \quad (18)$$

(17) implies (8). This concludes the proof.

Considering the affine properties of the TS fuzzy models, to reduce the conservatism in solution the enhanced design criterion can be derived by using two slack matrices.

Theorem 1

The fuzzy observer (5), (6) is stable if for given positive scalar $\delta \in R$ there exist a symmetric positive definite matrix $P \in R^{n \times n}$ and matrices $S_3 \in R^{n \times n}$, $Y_i \in R^{n \times m}$ such that for all $i = 1, 2, \dots, s$

$$P = P^T > 0 \quad (19)$$

$$\begin{bmatrix} A_i^T S_3 + S_3^T A_i - Y_i C - C^T Y_i^T & * \\ P - S_3 + \delta S_3^T A_i - \delta Y_{iC} & -\delta(S_3 + S_3^T) < 0 \end{bmatrix} < 0 \quad (20)$$

When the above conditions hold, the set of the observer gain matrices J_i is given by the equations

$$J_i = (S_3^T)^{-1} Y_i \quad (21)$$

Here and hereafter * denotes the symmetric item in a symmetric matrix.

Proof : Since the property of (3) and (12) asserts that

$$\sum_{i=1}^s h_i(\theta(t)) (A_{ei} e(t) - \dot{e}(t)) = 0 \quad (22)$$

using arbitrary square slack matrices $S_3, S_4 \in R^{n \times n}$ it yields

$$(q^T(t) S_3^T + \dot{q}^T(t) S_4^T) \sum_{i=1}^s h_i(\theta(t)) (A_{ei} e(t) - \dot{e}(t)) = 0 \quad (23)$$

Adding (23) and transposition of (23) to (15) gives

$$\begin{aligned}
\dot{v}(e(t)) &= \dot{e}^T(t)Pe(t) + e^T(t)P\dot{e}(t) + \\
&+ (e^T(t)S_3^T + \dot{e}^T(t)S_4^T) \sum_{i=1}^s h_i(\theta(t))(A_{ei}e(t) - \dot{e}(t)) + \\
&+ \sum_{i=1}^s h_i(\theta(t))(e^T(t)A_{ei}^T - \dot{e}^T(t))(S_3e(t) + S_4\dot{e}(t)) < 0
\end{aligned} \tag{24}$$

Then, introducing the notation

$$e^{\circ T}(t) = [e^T(t) \quad \dot{e}^T(t)] \tag{25}$$

after straightforward computation it can be obtained

$$\dot{v}(q(t)) = \sum_{i=1}^s h_i(\theta(t))e^{\circ T}(t)Q_i^{\circ}e^{\circ}(t) < 0 \tag{26}$$

where

$$Q_i^{\circ} = \begin{bmatrix} (A_i - J_i C)^T S_3 + S_3^T (A_i - J_i C) & P - S_3^T + (A_i - J_i C)^T S_4 \\ P - S_3 + S_4^T (A_i - J_i C) & -S_4 - S_4^T \end{bmatrix} < 0 \tag{27}$$

$$S_4 = \delta S_3, \quad Y_i = S_3^T J_i \tag{28}$$

where $\delta > 0$, $\delta \in R$, then (28) implies (20). This concludes the proof.

The importance of Theorem 1 is that the Lyapunov matrix P is separated from the system matrices A_i, C , i.e. there are no terms containing product of P and any of them. This enables to derive design conditions with respect to natural affine properties of TS models.

It is evident, that Theorem 1 can be simple reformulated considering a symmetric matrix $S_3 = S_3^T$.

4. DESCRIPTOR PRINCIPLE BASED DESIGN METHOD

The results given by Theorem 1 can be generalized using descriptor principle and are formulated as the following theorem.

Theorem 2

The fuzzy observer (22), (23) is stable if for given positive scalar $\delta \in R$ there exist a symmetric positive definite matrix $P_1 \in R^{n \times n}$ and matrices $P_2, P_3 \in R^{n \times n}$, $Y_i \in R^{n \times m}$ such that for all $i = 1, 2, \dots, s$

$$P_1 = P_1^T > 0, \quad P_2 + P_2^T > 0 \tag{29}$$

$$\begin{bmatrix} A_i^T P_3 + P_3^T A_i - Y_i C - C^T Y_i^T & * \\ P_1 - P_3 + \delta P_3^T A_i - \delta Y_i C & P_2 + P_2^T - \delta (P_3 + P_3^T) \end{bmatrix} < 0 \tag{30}$$

When the above conditions hold, the set of the observer gain matrices J_i is given by the set of the equations

$$J_i = (P_3^T)^{-1} Y_i \tag{31}$$

Proof : Using the equality (22), then with the identities

$$\dot{e}(t) = \dot{e}(t), \quad \mathbf{0} = \mathbf{0} \quad (32)$$

an equivalent form of (22) can be written as

$$\begin{bmatrix} I_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \dot{e}(t) \\ \ddot{e}(t) \end{bmatrix} = \begin{bmatrix} \dot{e}(t) \\ \mathbf{0} \end{bmatrix} = \sum_{i=1}^s h_i(\theta(t)) \begin{bmatrix} \mathbf{0} & I_n \\ A_{ei} & -I_n \end{bmatrix} \begin{bmatrix} e(t) \\ \dot{e}(t) \end{bmatrix} \quad (33)$$

or more generally

$$E^\circ \dot{e}^\circ(t) = \sum_{i=1}^s h_i(\theta(t)) A_{ei}^\circ e^\circ(t) \quad (34)$$

where $e^\circ(t)$ is given in (25) and

$$E^\circ = E^{\circ T} = \begin{bmatrix} I_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad A_{ei}^\circ = \begin{bmatrix} \mathbf{0} & I_n \\ A_{ei} & -I_n \end{bmatrix} \quad (35)$$

Defining the Lyapunov function of the form

$$v(e^\circ(t)) = e^{\circ T}(t) E^{\circ T} P^\circ e^\circ(t) > 0 \quad (36)$$

where

$$E^{\circ T} P^\circ = P^{\circ T} E^\circ \geq 0 \quad (37)$$

then the derivative of (36) becomes

$$\dot{v}(e^\circ(t)) = \dot{e}^{\circ T}(t) E^{\circ T} P^\circ e^\circ(t) + e^{\circ T}(t) P^{\circ T} E^\circ \dot{e}^\circ(t) < 0 \quad (38)$$

Inserting (34) in (38) it yields

$$\dot{v}(e^\circ(t)) = e^{\circ T}(t) \sum_{i=1}^s h_i(\theta(t)) (P^{\circ T} A_{ei}^\circ + A_{ei}^{\circ T} P^\circ) e^\circ(t) < 0 \quad (39)$$

$$P^{\circ T} A_{ei}^\circ + A_{ei}^{\circ T} P^\circ < 0 \text{ for all } i \quad (40)$$

respectively. Defining

$$P^\circ = \begin{bmatrix} P_1 & P_2 \\ P_3 & P_4 \end{bmatrix} \quad (41)$$

then (35), (37) implies

$$P_1 = P_1^T > 0 \quad (42)$$

and using (35) within (13) in (40) it yields

$$\begin{bmatrix} \mathbf{0} & (A_i - J_i C)^T \\ I_n & -I_n \end{bmatrix} \begin{bmatrix} P_1 & P_2 \\ P_3 & P_4 \end{bmatrix} + \begin{bmatrix} P_1^T & P_3^T \\ P_2^T & P_4^T \end{bmatrix} \begin{bmatrix} \mathbf{0} & I_n \\ A_i - J_i C & -I_n \end{bmatrix} < 0 \quad (43)$$

After some algebraic manipulations (43) takes the following form

$$\begin{bmatrix} (A_i - J_i C)^T P_3 + P_3^T (A_i - J_i C) & (A_i - J_i C)^T P_4 + P_1^T - P_3^T \\ P_1 - P_3 + P_4^T (A_i - J_i C) & P_2 + P_2^T - P_4 - P_4^T \end{bmatrix} < 0 \quad (44)$$

Setting

$$P_4 = \delta P_3, \quad Y_i = P_3^T J_i \quad (45)$$

where $\delta > 0$, $\delta \in R$, then (44) implies (30). This concludes the proof.

Remark 1

It is naturally to point out that Theorem 2 is an extension and generalization of Theorem 1, since setting

$$P_2 = 0, \quad P_1 = P, \quad P_3 = S_3 \quad (46)$$

(29), (30) implies (19), (20), respectively. The extension of (30) reflects the Krasovskii theorem property [13] allowing either to consider (24) in the following form

$$\begin{aligned} \dot{v}(e(t)) = & \dot{e}^T(t) P e(t) + e^T(t) P \dot{e}(t) + \\ & + (e^T(t) S_3^T + \dot{e}^T(t) S_4^T) \sum_{i=1}^s h_i(\theta(t)) (A_{ei} e(t) - \dot{e}(t)) + \\ & + \sum_{i=1}^s h_i(\theta(t)) (e^T(t) A_{ei}^T - \dot{e}^T(t)) (S_3 e(t) + S_4 \dot{e}(t)) < -\dot{e}^T(t) (S_2 + S_2^T) \dot{e}(t) < 0 \end{aligned} \quad (47)$$

or, equivalently, to define the Lyapunov function in the proof of Theorem 2 as follows

$$v(e(t)) = e^T(t) P e(t) + \int_0^t \dot{e}^T(\tau) (S_2 + S_2^T) \dot{e}(\tau) d\tau > 0 \quad (48)$$

and, as initially, to apply (24) in the proof and, finally, to compare the obtained result with (29), (30) setting

$$P = P_1, \quad S_3 = P_3, \quad S_2 = P_2 \quad (49)$$

Corollary 1

Considering

$$P_2 = 0, \quad P_4 = 0, \quad P_1 = P_3 \quad (50)$$

then (44) reduces to

$$\begin{bmatrix} (A_i - J_i C)^T P_1 + P_1^T (A_i - J_i C) & 0 \\ 0 & 0 \end{bmatrix} \leq 0 \quad (51)$$

which implies

$$(A_i - J_i C)^T P_1 + P_1^T (A_i - J_i C) < 0 \quad (52)$$

It is obvious that with

$$P = P_1 = P_1^T, \quad Y_i = P_1^T J_i = P J_i \quad (53)$$

(52) implies (8).

These modifications give the possibility to achieve the degree of conservatism that is most appropriate for a TS system.

5. ILLUSTRATIVE EXAMPLE

The considered system is represented by the TS fuzzy model (1), (2) with $s = 3$ and the system model parameters

$$A_1 = \begin{bmatrix} -1.0522 & -1.8666 & 0.5102 \\ -0.4380 & -5.4335 & 0.9205 \\ -0.5522 & 0.1334 & -0.4898 \end{bmatrix}, A_2 = \begin{bmatrix} -1.0565 & -1.8661 & 0.5116 \\ -0.4380 & -5.4359 & 0.9214 \\ -0.5565 & 0.1339 & -0.4884 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} -1.0602 & -1.8657 & 0.5133 \\ -0.4381 & -5.4353 & 0.9216 \\ -0.5602 & 0.1343 & -0.4867 \end{bmatrix}, B = \begin{bmatrix} 3 & 1 \\ 1 & -1 \\ 3 & 0 \end{bmatrix}, C^T = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$$

where matrices B_i are the same for all i and the premise variable and the membership functions for approximation of $f(q_1(t))$ in the prescribed sector are given as

$$\theta(t) = \begin{bmatrix} \theta_1(t) \\ \theta_2(t) \\ \theta_3(t) \end{bmatrix}, \theta_i = \begin{cases} \theta_1(t) & \text{if } q_1(t) \text{ is about } 5, \\ \theta_2(t) & \text{if } q_1(t) \text{ is about } 0, \\ \theta_3(t) & \text{if } q_1(t) \text{ is about } -5, \end{cases}$$

$$h_1(\theta_2(t)) = 1 - \frac{1}{5} |\theta_2(t) - 5|, \quad h_2(\theta_1(t)) = 1 - \frac{1}{5} |\theta_1(t)|, \quad h_3(\theta_3(t)) = 1 - \frac{1}{5} |\theta_3(t) + 5|$$

Solving the variables $P, Y_i, i = 1, 2, 3$ satisfying (7), (8) via the LMI technique using toolbox SeDuMi [14] gave the following results

$$P = \begin{bmatrix} 0.6353 & -0.1200 & -0.0377 \\ -0.1200 & 0.2368 & 0.0521 \\ -0.0377 & 0.0521 & 0.6776 \end{bmatrix}$$

$$Y_1 = \begin{bmatrix} -0.0969 & -0.1124 \\ -0.3185 & -0.1973 \\ 0.0101 & 0.1690 \end{bmatrix}, Y_2 = \begin{bmatrix} -0.0992 & -0.1120 \\ -0.3173 & -0.1979 \\ 0.0089 & 0.1698 \end{bmatrix}, Y_3 = \begin{bmatrix} -0.1002 & -0.1118 \\ -0.3170 & -0.1977 \\ 0.0082 & 0.1709 \end{bmatrix}$$

and, besides, the fuzzy observer gain matrices were obtained as follows

$$J_1 = \begin{bmatrix} -0.4474 & -0.3634 \\ -1.5966 & -1.0859 \\ 0.1126 & 0.3127 \end{bmatrix}, J_2 = \begin{bmatrix} -0.4502 & -0.3632 \\ -1.5925 & -1.0888 \\ 0.1104 & 0.3140 \end{bmatrix}, J_3 = \begin{bmatrix} -0.4517 & -0.3627 \\ -1.5914 & -1.0881 \\ 0.1092 & 0.3157 \end{bmatrix}$$

guaranteeing the stable eigenvalues spectra of the local observer system matrices in such a way that

$$\rho(A_{e1}) = \{-0.7560, -1.7011 \pm 1.1316i\}, \quad \rho(A_{e2}) = \{-0.7575, -1.7029 \pm 1.1269i\}$$

$$\rho(A_{e3}) = \{-0.7599, -1.7034 \pm 1.1265i\}$$

Applying the same toolbox to solve LMIs (19), (20) conditioned by $\delta = 1$, the obtained set of matrix variables was as follows

$$P = \begin{bmatrix} 0.6415 & -0.1152 & 0.0157 \\ -0.1152 & 0.7158 & -0.0994 \\ 0.0157 & -0.0994 & 0.7004 \end{bmatrix}, S_3 = \begin{bmatrix} 0.3577 & -0.0386 & -0.0348 \\ -0.1063 & 0.1468 & 0.0100 \\ 0.0595 & 0.0316 & 0.3587 \end{bmatrix}$$

$$Y_1 = \begin{bmatrix} -0.0188 & 0.0224 \\ -0.1142 & -0.0324 \\ -0.1452 & 0.1856 \end{bmatrix}, Y_2 = \begin{bmatrix} -0.0191 & 0.0248 \\ -0.1155 & -0.0323 \\ -0.1465 & 0.1828 \end{bmatrix}, Y_3 = \begin{bmatrix} -0.0227 & 0.0237 \\ -0.1138 & -0.0327 \\ -0.1458 & 0.1833 \end{bmatrix}$$

so that the local observers gain matrices were given as

$$J_1 = \begin{bmatrix} -0.2067 & -0.1318 \\ -0.7449 & -0.3663 \\ -0.4039 & 0.5149 \end{bmatrix}, J_2 = \begin{bmatrix} -0.2097 & -0.1226 \\ -0.7543 & -0.3616 \\ -0.4076 & 0.5079 \end{bmatrix}, J_3 = \begin{bmatrix} -0.2171 & -0.1270 \\ -0.7449 & -0.3652 \\ -0.4067 & 0.5088 \end{bmatrix}$$

This set of gains embedded the eigenvalues spectra of the local observer system matrices as follows

$$\rho(A_{e_1}) = \{-4.1633, -1.0047 \pm 0.0879i\}, \quad \rho(A_{e_2}) = \{-4.1600, -1.0016 \pm 0.0804i\}$$

$$\rho(A_{e_3}) = \{-4.1703, -0.9967 \pm 0.0917i\}$$

Finally, solving LMIs (29), (30) conditioned by $\delta = 1$, a feasible solution produced the following LMI variables

$$P_1 = \begin{bmatrix} 1.0236 & -0.2562 & 0.0263 \\ -0.2562 & 1.0633 & -0.1593 \\ 0.0263 & -0.1593 & 1.1439 \end{bmatrix}, P_3 = \begin{bmatrix} 0.7114 & -0.1530 & 0.1444 \\ -0.1842 & 0.2562 & -0.0228 \\ -0.1177 & 0.1345 & 0.7273 \end{bmatrix}$$

$$Y_1 = \begin{bmatrix} -0.1798 & 0.0085 \\ -0.1937 & -0.0745 \\ -0.3158 & 0.1619 \end{bmatrix}, Y_2 = \begin{bmatrix} -0.1816 & 0.0096 \\ -0.1938 & -0.0745 \\ -0.3186 & 0.1637 \end{bmatrix}, Y_3 = \begin{bmatrix} -0.1833 & 0.0106 \\ -0.1937 & -0.0744 \\ -0.3210 & 0.1655 \end{bmatrix}$$

where, for simplicity, P_2 is not listed. This result provides TS fuzzy state observer with the following local gain matrices

$$J_1 = \begin{bmatrix} -0.5430 & -0.0672 \\ -0.8942 & -0.4476 \\ -0.3544 & 0.2220 \end{bmatrix}, J_2 = \begin{bmatrix} -0.5462 & -0.0653 \\ -0.8950 & -0.4475 \\ -0.3576 & 0.2241 \end{bmatrix}, J_3 = \begin{bmatrix} -0.5491 & -0.0635 \\ -0.8951 & -0.4470 \\ -0.3604 & 0.2262 \end{bmatrix}$$

and with the eigenvalues spectra of the local observer system matrices

$$\rho(A_{e_1}) = \{-4.0034, -0.6547 \pm 0.1088i\}, \quad \rho(A_{e_2}) = \{-4.0056, -0.6553 \pm 0.1090i\}$$

$$\rho(A_{e_3}) = \{-4.0060, -0.6557 \pm 0.1095i\}$$

Applying the designed to the TS fuzzy system model with the initial condition

$$q_e^T(0) = 0, \quad u^T(t) = 0, \quad q^T(0) = [0.3 \quad 0.6 \quad 0.9]$$

the simulation results are stated in Fig. 1 to Fig. 3 to illustrate the estimated output behaviour of the system sequentially as the observers parameters were computed using Lemma 1, Theorem 1 and Theorem 2. It is evident that the best compromise in the settling time and overshooting gives the result of Theorem 2.

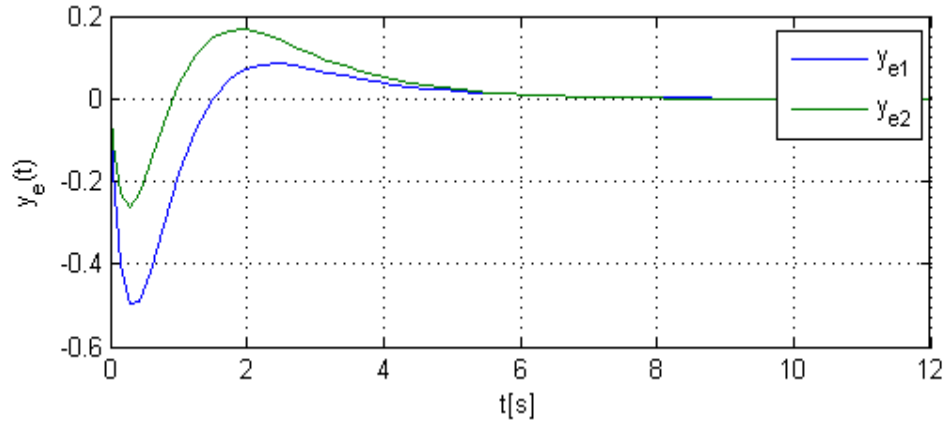


Figure 1: TS fuzzy observer output variables response (based on Lemma 1 results)

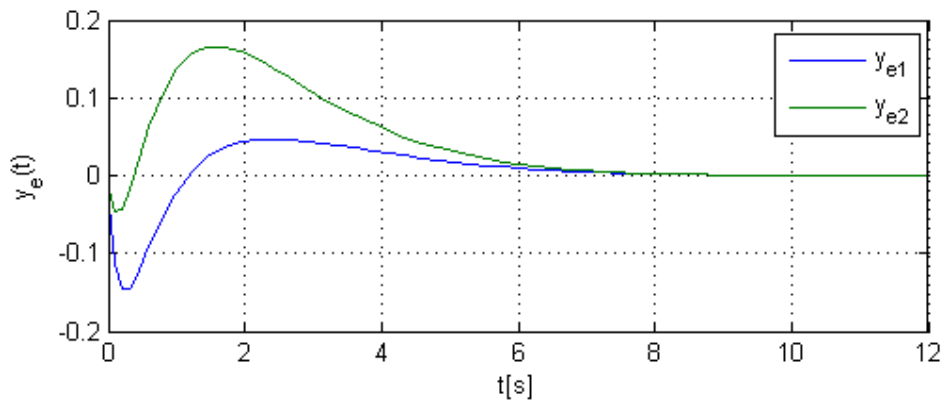


Figure 2: TS fuzzy observer output variables response (based on Theorem 1 results)

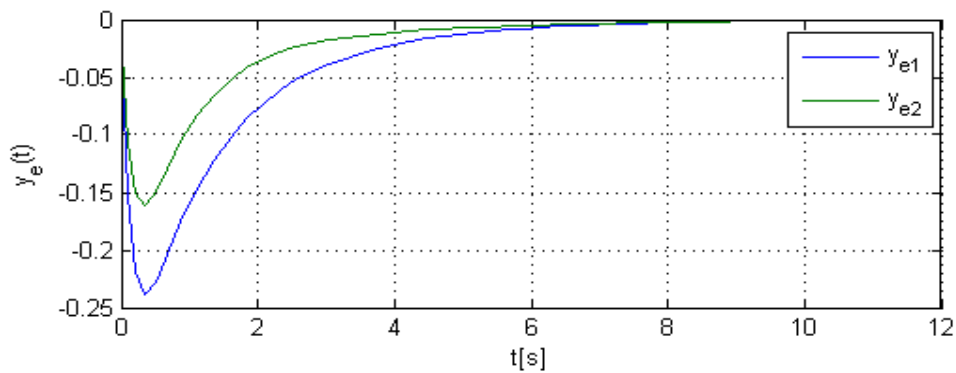


Figure 3: TS fuzzy observer output variables response (based on Lemma 2 results)

6. CONCLUDING REMARKS

New approach for output dynamic feedback control design is presented in this paper. By the proposed procedure the control problem is parameterized in such LMIs set with one additional LME which admit more freedom in guaranteeing the output feedback control performance for a bi-proper dynamic controller and by LMIs set only for a strictly proper dynamic output controller. Sufficient conditions of the controller existence manipulating the stability of the closed-loop systems imply the control structure, which stabilize the system in the sense of Lyapunov and the controller design tasks is a solvable numerical problem. An additional benefit of the method is that controller uses minimum feedback information with respect to desired system output and the approach is enough flexible to allow the inclusion of additional design condition bounds.

ACKNOWLEDGEMENTS

The work presented in the paper was supported by VEGA, the Grant Agency of the Ministry of Education and the Academy of Science of Slovak Republic, under Grant No. 1/0348/14. This support is very gratefully acknowledged.

REFERENCES

- [1] Thau, F.E. (1973) "Observing the state of nonlinear dynamical systems", *International Journal of Control*, Vol. 17, No. 5, pp. 471-479.
- [2] Koshkouei, A.J. & Zinober, A.S.I. (1999) "Partial Lipschitz nonlinear sliding mode observers", *Proceedings of the 7th Mediterranean Conference on Control and Automation MED99*, Haifa, Israel, pp. 2350-2359.
- [3] Takagi, T. & Sugeno, M. (1985) "Fuzzy identification of systems and its applications to modeling and control", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 15, No. 1, pp. 116-132.
- [4] Ichalal, D., Marx, B., Ragot, J. & Maquin, D. (2007) "Design of observers for Takagi-Sugeno discrete-time systems with unmeasurable premise variables", *Proceedings of the 5th Workshop on Advanced Control and Diagnosis ACD 2007*, Grenoble, France.
- [5] Gao, Z., Shi, X. & Ding, S.X. (2008) "Fuzzy state/disturbance observer design for T-S fuzzy systems with application to sensor fault estimation", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 38, No. 3, pp. 875-880.
- [6] Tanaka, K. & Wang, H.O. (2001) *Fuzzy Control Systems Design and Analysis: A Linear Matrix Inequality Approach*. New York, John Wiley & Sons.
- [7] Lu, G., Ho, D.W.C. & Zheng, Y. (2004) "Observers for a class of descriptor systems with Lipschitz constraint", *Proceeding of the 2004 American Control Conference*, Boston, MA, USA, pp. 3474-3479.
- [8] Fridman, E. & Shaked, U. (2002) "A descriptor system approach to H_∞ control of linear time-delay systems", *IEEE Transactions on Automatic Control*, Vol. 47, No. 2, pp. 253-270.
- [9] Ilhem, K., Dalel, J., Saloua, B.H.A. & Naceur, A.M. (2012) "Observer design for Takagi-Sugeno descriptor system with Lipschitz constraints", *International Journal of Instrumentation and Control Systems*. Vol. 2, No. 2, pp. 13-25.
- [10] Tong, S., Yang, G. & Zhang, W. (2011) "Observer-based fault-tolerant control against sensor failures for fuzzy systems with time delays", *International Journal of Applied Mathematics and Computer Science*. Vol. 21, No. 4, pp. 617-627.
- [11] Filasova, A. & Krokavec, D. (2013) "On the Takagi-Sugeno model-based state estimation for one class of bilinear systems", *Proceedings of the 14th International Carpathian Control Conference ICC'13*, Rytro, Poland, pp. 83-87.
- [12] Krokavec, D. & Filasova, A. (2012) "Optimal fuzzy control for a class of nonlinear systems", *Mathematical Problems in Engineering*, Vol. 2012, ID 481942, 29 p.
- [13] Haddad, W.M. & Chellaboina, V. (2008) *Nonlinear Dynamical Systems and Control: A Lyapunov - Based Approach*, Princeton, NJ, USA, Princeton University Press.
- [14] Peaucelle, D., Henrion, D., Labit, Y. & Taitz, K. (2002) *User's Guide for SeDuMi Interface 1.04*, Toulouse, France, LAAS-CNRS.

AUTHORS

Anna Filasova graduated in technical cybernetics and received M.Sc. degree in 1975, and Ph.D. degree in 1993 both from the Faculty of *Electrical* Engineering and Informatics, Technical University of Kosice, Slovakia. In 1999 she was appointed Associated Professor from the Technical University in Kosice in technical cybernetics. She is with the Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, and she has been working with as an Assistant Professor from 1975 to 1999. Her main research interests are in robust and predictive control, decentralized control, large-scale system optimization, and control reconfiguration.

Dusan Krokavec received M.Sc. degree in automatic control in 1967 and Ph.D. degree in technical cybernetics in 1982 from the Faculty of Electrical Engineering, Slovak University of Technology in Bratislava, Slovakia. In 1984 he was promoted Associated Professor from the Technical University in Kosice, Slovakia, and in 1999 he was appointed Full Professor in automation and control. He is with the Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Kosice. In the long term, he specializes in stochastic processes in dynamic systems, digital control systems and digital signal processing, and in dynamic system fault diagnosis.

COOPERATING ADAPTIVE DEVICES APPLIED IN GENERAL GAME PLAYING

Jose Maria Novaes dos Santos and Joao Jose Neto

Escola Politecnica – Universidade de Sao Paulo (USP), Sao Paulo, Brazil

ABSTRACT

There are formalisms in literature where behavior can be described by a finite set of rules that maps the current device's state into a new one like the finite state machines, statecharts and petri nets. Those formalisms are named as rule-driven devices. A formal device is said to be adaptive if its behavior changes dynamically in response just to its input stimuli and its current state without any external help. Adaptive rule-driven devices can be used for modeling complex problems in artificial intelligence, natural language, reactive systems, synchronous system and others applications. General game playing (GGP) is a research subfield of Artificial Intelligence which aims at developing systems able to play a variety of games, knowing and understanding the rules only in execution time. Most of the proposed GGP systems are based on statistic methods. This paper aims at a new approach in GGP systems, proposing a solution based on adaptive technology modeling. In brief, a GGP system has been developed in a way that it changes its behavior dynamically in response to rules information and its history of games played. This proposal is presented here as well as results of some games played comparing their strategies approach.

KEYWORDS

Modelling, Reactive System, Cooperating Adaptive Devices, Rule-Driven Formalisms, Self-Modifying Machine, Adaptive Automata, General Game Playing, Adaptive Device

1. INTRODUCTION

Many traditional formalisms can be described as a finite set of rules which maps its current state into a new one in response to some stimulus or event. Some states are defined as final states and indicate successfully finished operations. Such formalisms are named rule-driven devices.

A formal device is said to be adaptive if its behavior changes dynamically without any external help. The behavior changes in response to stimuli occurred in the environment. Such device property is called self-modification or adaptivity and has been formally described in [1].

In some situations, modeling of the problem requires simultaneous usage of devices of different nature types.

General Game Playing (GGP) research is in the field of Artificial Intelligence and its purpose is to design and implement systems with the ability to understand the rules of new games and be able to play any game described in a specific language called Game Description Language (GDL). So, the most meaningful characteristic of the GGP is that players do not know the games rules before beginning to play. Hence, this field encourages the development of learning

algorithms and search mechanisms which could be applied in a wide range of games, similar to human learning in playing games.

Concerning the feature of the adaptive technology and the general game playing, this work presents a GGP system where the behavior is modeled based on adaptive technology.

In this work is presented an approach based in adaptive technology and learning process which uses the history of games, thereby as more games are played, the system will provide more and more options of new moves.

In section 2, we briefly describe the static rule-driven devices. Afterwards, we introduce the concept of adaptivity in section 3. In section 4, we present the concept of Cooperating Adaptive Devices. In section 5, we introduce the main ideas of general game playing. In section 6, we present our system proposal and the results of the tests. Section 7 brings the conclusion and some remarks.

2. RULE-DRIVEN DEVICES

Many formalisms change their state in response to an environment stimulus according to their set of rules. The state of the device comprehends the contents of the whole set of elements that hold information along with the current status of the device.

Some of these traditional formalisms include finite state machines, statecharts, natural language grammars and parsers, ontologies, decision trees, decision tables and petri nets. These devices have been used to model and represent complex problems in many fields such as Artificial Intelligence and Natural Language Processing. One of the most popular among these formulations is the finite state machine, which is widely used for describing the behavior of real time systems, communication protocols, software design and language parsing.

A rule-driven device is characterized by a set of rules that determines its reaction to a given set of conditions. Each rule represents a change in its state in response to a set of conditions and events. The device starts its operation at some known initial state and modifies its state by applying the best suited of its rules.

The device is said to be deterministic if and only if, for any given state and for any input stimuli, there is just a single next state defined by its set of rules. Otherwise, the device is said to be non-deterministic. In other words, a non-deterministic device has a set of rules that maps at least one possible state into more than one next state. We usually achieve better efficiency from deterministic devices than from non-deterministic equivalent ones. Unfortunately, for some problems, it may be very difficult or even impossible to obtain solutions based on deterministic devices.

3. ADAPTIVE RULE-DRIVEN DEVICES

A formal rule-driven device is said to be adaptive if its behavior may change dynamically. The concept of adaptivity applies to any device that is able to change its own behavior. In particular, when the behavior is determined by a set of rules, adaptivity is easily achieved by changing the set of rules that define the device's behavior.

The general formulation for rule driven adaptive devices, can be thought as an adaptive layer placed around the original subjacent non-adaptive device (standard rule-driven device).

By building such devices, one can conceptually identify two major components: an equivalent underlying device similar to those described in the previous section and an adaptive mechanism responsible for the adaptivity. One notation elaborated for representing adaptive formalisms in a way as similar as possible to its original non-adaptive underlying formulation was presented in detail in [1].

Historically, adaptive devices emerged from automata theory and most of the early applications were in the fields of formal languages, and later, in computer languages. Some of those early works are described in [2], [3] and [4].

In [5], we have the formal formulation of the Adaptive Automaton. Hereafter, some works have been developed applying the adaptive technology using classical formalisms like Finite Automaton [6], Statechart [7], Markov Chain [8], Grammar [9] and Decision Table [10]. In [11], we have the conceptual proof that Adaptive Automaton and the Turing Machine have the same power of expression.

In [12], the use of adaptive technology is presented as an alternative approach for modeling biological species while in [13] we have the adaptive version of the GARP (Genetic Algorithm for Rule-set Production) algorithm for mapping the environmental distribution of the “Penonapis” and “Cucurbita”.

The adaptive technology has been applied also in other fields like robotic [14], programming language [15], [16] and [17] and reactive system modeling [18], [19] and [20].

4. COOPERATING ADAPTIVE DEVICES

We have briefly mentioned the adaptive rule-driven device’s main idea in the previous section. In this section, we present the Cooperating Adaptive Devices. Concisely, this formalism is formed by a group of adaptive devices with the feature of communication between them, increasing the number of real complex problems that can be modelled. One device can send messages to any other devices, that in response, can modify its state. Based on this behavior, one device can cooperate with another one, sending messages to communicate a condition in the environment that may cause a state changing in the second device.

Cooperating Adaptive Devices (CAD) are composed of a finite group of rule-driven devices of potentially heterogeneous types with a common communication mechanism (CCM). A further device or mechanism is employed for managing and coordinating message communication (MCM), resulting from interactions between a pair of distinct devices. The main consequence of such a decision is that devices have now the capability of sending a message that may cause the behavior changing of other devices.

The concept of the CAD can be summarized as the property in which, based on its behavior or on the basis of its current situation and input stimuli, any member of the group of heterogeneous devices can send communication instructions resulting in changing device rules.

Any action for changing rules belonging to any other device is initiated by a message communication from a source device to a target device, using a common communication mechanism as an intermediary agent for such interactions. We call this intermediary agent the common communicate mechanism (CCM).

Figure 1 illustrates an adaptive interaction among devices. Device 4 sends a message to device 1 through the common communicate mechanism (CCM). Such interaction between devices is represented by the red arrow in figure 1.

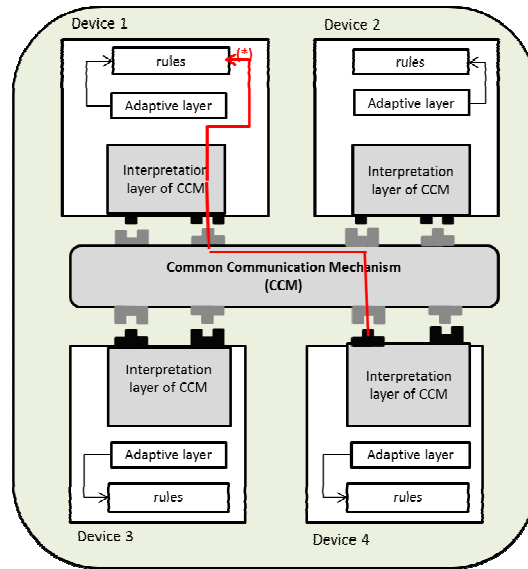


Figure 1. Illustration of cooperating adaptive devices

Technically, we can say that each adaptive device itself changes its own rules, however, any device belonging to the group of CAD has the capability to send messages to any other device of the group, which may cause rules modifications. This capability is represented in figure 1 by the layer of communication and interpretation.

Communications between devices, managed and controlled by CCM mechanism is performed using a communication protocol, which is named CP protocol. The communication between devices is formed of a sequence of standard messages from CP, like “insert rule”, “delete rule”, “trigger event”, etc.

Definition

Similar to the behavior of adaptive devices, a built-in counter T2 is defined for cooperating adaptive devices with initial value 0 and automatic increments by 1 whenever a non-null communication message is executed. Thus, each name of a device during a step tk ($tk \geq 0$) is identified for each value assumed by T2.

Thus, the cooperative adaptive device can be described as:

$$CAD_{tk} = (\{ AD_{k1,tk}^1, AD_{k2,tk}^2, \dots, AD_{km,tk}^m \}, CCM, MCM) \quad (1)$$

In this formulation $AD_{kr,tk}^r$ represents an adaptive device. CAD_{tk} is said cooperative adaptive device when for all operations, in each step tk ($tk \geq 0$), any element of the set of devices follows the behavior of the corresponding element until the execution of some non-null message in the common communicate mechanism (CCM), when the current step tk terminates and the next one (tk+1) starts.

Similar to adaptive actions, each step increment is composed of two groups of messages. The first one is performed before the rule execution and the second one after the execution. For a non-null message, at least one of the components must be non-null. A message is formed by elementary messages and its execution may cause multiple additions and/or multiple deletions in some other device's rules.

The cooperative adaptive device starts its operation at some known initial shape for all m ($m > 1$) devices of the CAD, $(AD^1_0, AD^2_0, \dots, AD^m_0)$ from the perspective of the message communications. Remembering that each device AD^r can change its rules through conventional adaptivity, we can express CAD as $(AD^1_{k1,0}, AD^2_{k2,0}, \dots, AD^m_{km,0})$ whereas each $AD^r_{kr,0}$ (for $r=1, \dots, m$ and $kr=k1, \dots, km$; all $kr \geq 0$) means a device at some internal step of the adaptivity and at the initial step of the message communications. So, in this state, no message communication has occurred. However, each device can have changed its own rules through conventional adaptivity.

At step tk ($tk \geq 0$), an input stimulus changes the cooperative adaptive device CAD to the next step ($tk+1$) if, and only if, any non-null message communication is performed. Then, in any combination of step kr for all m devices ($kr=k1, k2, \dots, km$) and the step tk , every device can be represented in the form $AD^r_{kr,tk}$. In this formulation, kr indicates the step of its adaptivity while tk indicates the step of the message communications for all devices.

So, we have:

$$CAD = \{(AD^r)_{tk}\} \quad (2)$$

$$(AD^r)_{tk} = AD^r_{kr,tk} \quad (3)$$

$$AD^r_{kr,tk} = (C^r_{kr,tk}, IAR^r_{kr,tk}, S^r, c^r_{kr,tk}, A^r, NA^r, BA^r, AA^r, IBA, IAA) \quad (4)$$

Where $r=1, \dots, m$; $kr=k1, \dots, km$ and $tk \geq 0$.

- In this formulation, we have:
- $CAD = (\{ AD^1_0, AD^2_0, \dots, AD^m_0 \}, CCM, MCM)$. The cooperative adaptive device consists of an initial set of m adaptive devices, a common communicate mechanism (CCM), managed by a mechanism in order to ensure only one concurrent communication (MCM).
- $(AD^r)_{tk}$, for $r = 1, \dots, m$ represents the adaptive state of each device of the cooperative adaptive device, $(AD^r)_0$ is its initial state ($tk=0$) and it is defined by its set of rules $IAR^r_{kr,0}$ for kl step of the adaptivity ($kr \geq 0$).
- $GADM_{tk} = \{AD^1_{k1,tk}, AD^2_{k2,tk}, \dots, AD^m_{km,tk}\}$ represents the devices at step tk of message communication, being $AD^r_{kr,0}$ its initial state. Each device $(AD^r_{kl})_{tk}$ is the mirror of adaptive device AD^r at step tk , each one within its own step kl ($kr \geq 0$ and $m > 1$; for $kr = k1 \dots, km$) of adaptivity.
- $C^r_{kr,tk}$ is the set of all possible states for $AD^r_{kr,tk}$ for tk and kr steps ($tk \geq 0$ e $kr \geq 0$, for $r=1 \dots m$).

- IBA e IAA (for $r=1, \dots, m$) are sets of message communication, both containing the null action ε ($\varepsilon \in IBA \cap IAA$).
- S^r (for $r=1, \dots, m$) is a finite set of all possible events considered valid input stimuli for AD^r , containing the null event ($\varepsilon \in S^r$).
- The input stimulus w^r is:
- $w^r = w_1^r w_2^r w_3^r w_4^r \dots w_n^r$ ($w^r \in S^r$) (for $r=1, \dots, m$ and $n_r \geq 0$).
- $c_{kr,0}^r$ belongs to C^r and is the initial device configuration ($c_{kr,tk}^r \in C_{kr,tk}^r$), for $r = 1 \dots, m$; $kr = k_1 \dots, k_m$ and $tk \geq 0$. Before the occurrence of the first adaptive action ($kr = 0$) and the first communication message ($tk = 0$), $c_{0,0}^r$ is the initial configuration of the device “r”.
- A^r is the subset of its accepting states (acceptance) of device r, $A^r \subseteq C^r$ (for $r=1, \dots, m$)
- NA^r is a finite set of output symbols of device r (for $r = 1, \dots, m$).
- IAR_{tk}^r is the finite set of all possible CAD rules, given by a relation $IAR_{tk}^r \subseteq IBA \times BA^r \times C^r \times S^r \times C^r \times NA^r \times AA^r \times IAA$. The rules of IAR_0^r (for $r=1, \dots, m$) define the initial performance of CAD devices. Device rule containing message communication changes another device rules set by adding and/or deleting rules. The rules HAR_{tk}^r ($r=1, \dots, m$) have the form $har^r = (iba, ba^r, c_i^r, s^r, c_j^r, z^r, aa^r, iaa)$, meaning that, in response to some stimulus $s^r \in S^r$, har^r initially performs the first group of message iba (group before), then, the adaptive rule $ar^r = (ba^r, c_i^r, s^r, c_j^r, z^r, aa^r)$ and finally performs the second group of message iaa (group after). The adaptive rule action, ar^r , is performed as described in [1].

5. GENERAL GAME PLAYING

General game playing is associated to the system's main goal which is to play games without previous knowledge with respect to its rules. The challenge is to construct a system able to understand the games rules, which are described in GDL (General Definition Language) [21], similar to the Prolog language.

Many works have been developed in Artificial Intelligence related to playing specific games like chess. In these systems, the programmer designer specifies the strategies of the programs to play just one game based on a set of known rules. The strategies are specific to a game play and it is difficult to use the same programs to play another game. Such programs have in common that they depend heavily on elaborated game-dependent knowledge provided by the developers.

In general game playing, on the contrary, the aim is to create programs capable of playing a wide range of different games, even those that may have never been played before. Since 2005, the Stanford Logic Group organizes yearly a competition in GGP games, improving studies concerning artificial intelligence, search techniques and knowledge representation, machine learning, knowledge discovery and online optimization.

The GGP games are synchronous and finite and can be modeled using the finite state machine formalism. In this representation, in every step of the game, we must have moves of all players. After the moves of all players, the environment is updated as a response without any external interference according to the set of rules. By definition, every GGP has at least one terminal state that can be reached after a finite number of rounds.

GGP is similar to the traditional game theory, however, there are some exceptions like the movement and modeling. GGP is modeled as a state machine while the traditional games are modeled as a tree. The GGP movements are synchronous.

The challenge in GGP systems is developing general method learning strategies concerns only the definitions rules. For example, a simple game has around many possible states, so identifying all possible moves leading to a final state is not feasible.

Concerning the restriction described, there are two main questions to be addressed in developing GGP systems: search and evaluation. Search implies in the ability of the system to think forward while evaluation intends to provide a mechanism that associates the merits of the current position. The merit value must be as greater as possible, meaning the maximum value is associated to a winning state. And the main challenge is to discover the relevant information that must be considered to build the mechanism during the game.

In GGP, the current state-of-the-art approach to search the game tree is the Upper Confidence Bounds Applied to Tree (UCT), where the goal is to provide balance between exploration and exploitation. In earlier competitions, one of the most applied approach was the Monte-Carlo Tree Search (MCTS) [22].

The basic idea of the MCTS simulation is to play in a randomly way until reaching a terminal state. In this state, a goal value is assigned to all states reached in the path. Each state value assigned is estimated by the average result of all simulations which visited this state. In short, the MCTS approach is comprised of four phases: selection, expansion, simulation and back propagation of results.

The work presented in [23], is a related work to the proposal presented in this paper. See more details concerning general game playing in works [24], [25], [26] and [27].

6. GENERAL GAME PLAYING APPLICATION

A problem related to developing a general game playing is to design a system which plays any game described in GDL language with the capability of learning play strategies dynamically.

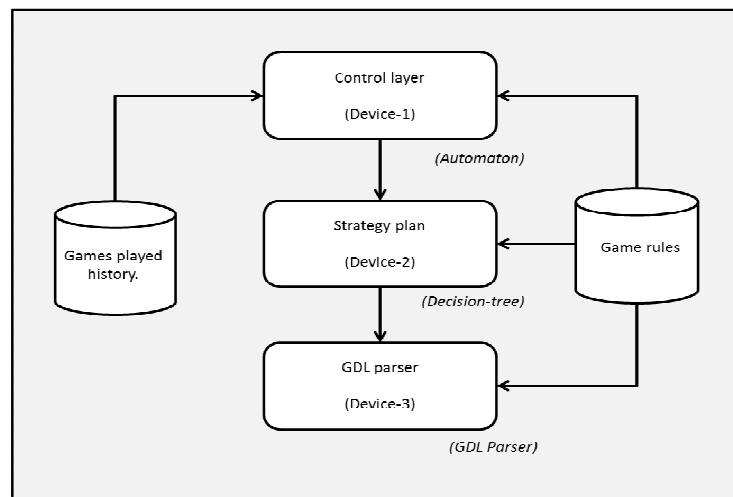


Figure 2. Three devices: an automaton, a decision-tree and a GDL parser

A problem related to developing a general game playing is to design a system which plays any game described in GDL language with the capability of learning play strategies dynamically. This problem can be modeled by the Cooperating Adaptive Devices, which has the feature of representing dynamic and autonomous behavior.

The modeling is based on three distinct type rule-driven devices. The first is an automaton, called “Control layer”, which is responsible to analyses the history of the games already played and the game rules to learn new strategies to play the game. Based on the learned strategy information, this device causes the inclusion of new rules in the second device, the decision tree, which is responsible for deciding what move is the best choice to apply in each round play of the game with respect to its rules and the strategy information. The third device, the GDL parser, is responsible to execute the rules, changing the current state of the game. Figure 2 illustrates these Cooperating Adaptive Devices.

6.1. Strategies Approach

We developed a GGP system where strategies are based on learning from historical information about the games played. In our approach, we have four different strategies based on history, one based on random strategy and one based on statistics. The statistic strategy is based on MCTS approach mentioned in section 5. The random strategy was used to compare the efficiency of each different strategy.

Our strategies are based on two simple ideas. The first is to follow some path of a successful game, comparing the state of the current game with the historic games. Based on this idea, we have come up with strategies A and B. The second idea is to follow a final state of a successful game. Based on this idea, we have come up with strategy C. The strategy D is based on a final state, but it considers only relevant information that leads to a final state.

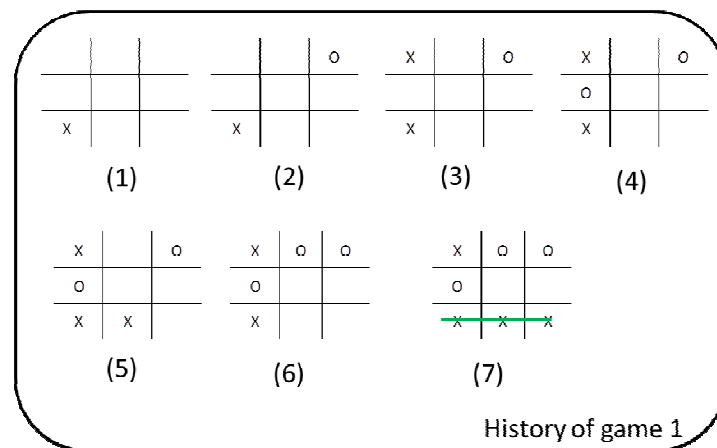


Figure 3. Example of “noughts and crosses” game stored in the historic database

To illustrate our ideas, let’s suppose our historic database formed only by one game played as showed in figure 3 (history of game 1). In this situation and in another “noughts and crosses” game, with current configuration as showed in figure 4, a role player “X” has only one option to move based on strategies A or B, the cell(1,1), like the third round of the game 1. On the order hand, based on strategy C, we have three options considering the final state of game 1. Strategy D considers only two positions, the last row positions considering the game 1. In every round, we consider all possible moves. After, we analyse all next configurations, concerning the strategy.

Concisely, our strategies are:

- Strategy R.
In every step of the game, there is a random choice of all valid moves.
- Strategy E.
The statistic strategy verifies the state with the highest average of the visited number and the number of victories.
- Strategy A.
It looks for a game state that is the same as the current game state. If there is more than one choice, the most recent one is chosen.
- Strategy B.
It looks for a game state that is most similar to the current game state. If there is more than one choice, the most recent one is chosen.
- Strategy C.
It looks for a final state of a successful game that is most similar to the current game state.
- Strategy D.
It looks for a final state of a successful game that is most similar to the current game state, considering only the relevant information in the final state.

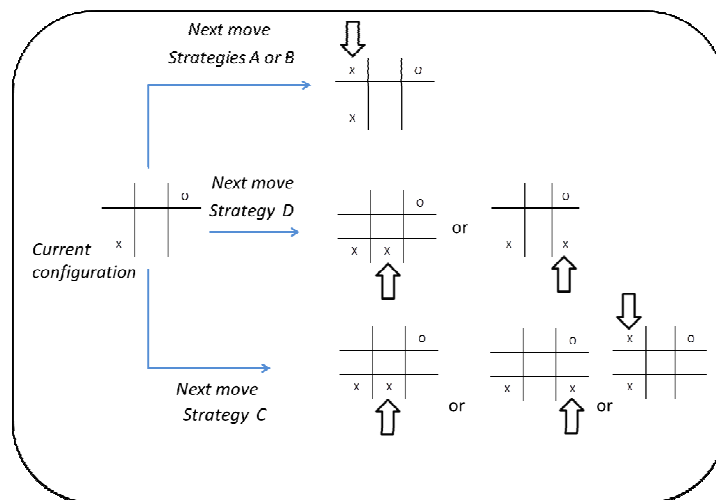


Figure 4. Example of possible moves, according strategy

6.2. Result of Games Played

In our tests, we played the “noughts and crosses” and the “cross-block” games. The “noughts and crosses” game is the classical game for two players who take turns marking the spaces in a 3×3 grid. The player who succeeds in placing three respective marks in a horizontal, vertical, or diagonal row wins the game. The “cross-block” game is for two players who take turns marking

the spaces in a 4×4 grid. The cross role wins if its marks link the left side to the right side or the upper side to the bottom side. The block role wins if the cross loses.

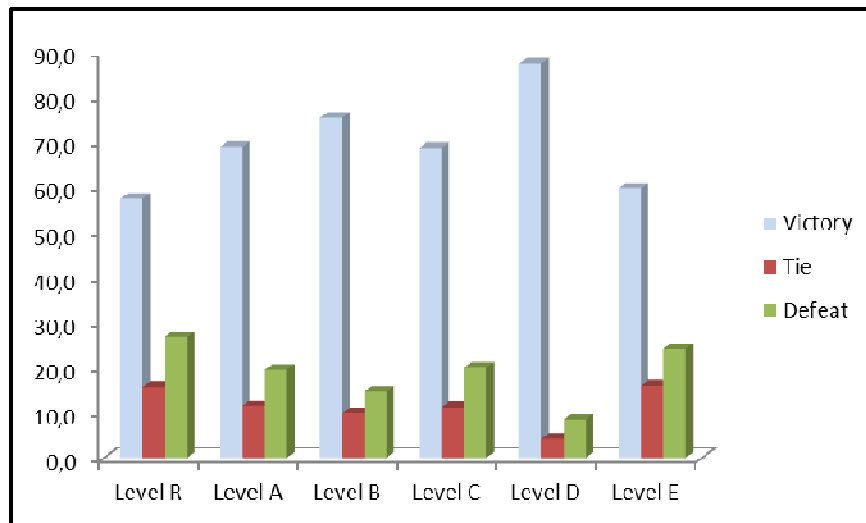


Figure 5. Results of “noughts and crosses” games played

We played several games combining all possible strategies for the “x” role player in the “noughts and crosses” games and maintaining the same strategy for the opponent, the random one. For the “cross-block” game, we combined all possible strategies in the cross role, maintaining the same strategy (random) for the block role. Each combination was played 500 times. In figures 5 and 6, we have the graphs of the results of games played

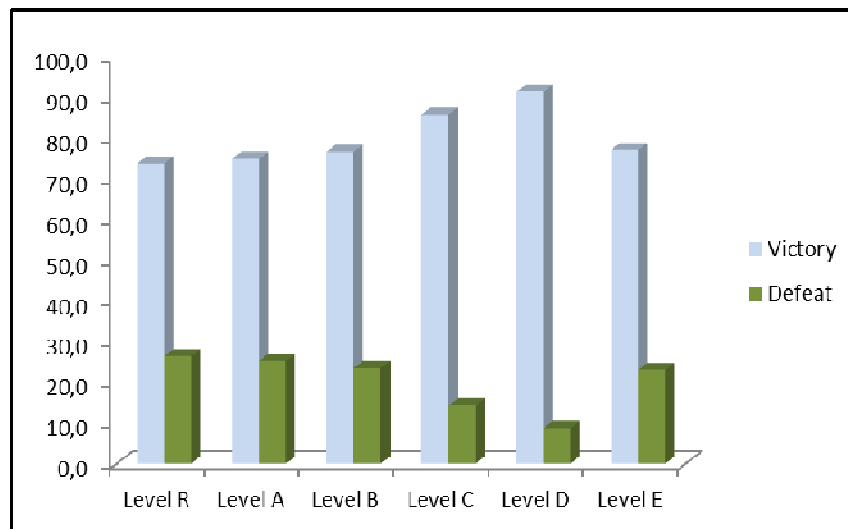


Figure 6. Results of “cross-block” games played.

Based on result analysis, we conclude that the best strategy is “D”, because it only considers the information that is essential in the final conditions, having fewer no important moves.

It is important to highlight that all strategies have improvements comparing to the random strategy. These results stimulate new studies in learning algorithm, knowledge representation and adaptive technology applied to GGP.

7. CONCLUSIONS

The GGP system behavior depends which game will be played, represented in the rules of the game. Thus, it is a typical application that can be modeled using adaptive devices.

This work's main result is the use of the adaptive technology to model and develop a general game playing system. The results have showed that this proposal encourages new approaches concerning techniques in knowledge representation and learning reasoning. We have many GGP application based on statistic approaches and this work presents a new perspective of dealing with unknown environment. The preliminary results have encouraged new challenges and we intend to test the system under even more complex games thus improving the techniques in learning strategies.

We hope this work will assist in the conception and building of Cooperating Adaptive Devices for a new and cleaner perspective in activities involving complex problem solving with adaptive technology.

REFERENCES

- [1] Neto, João José, (2002) "Adaptive rule-driven devices-general formulation and case study." Implementation and Application of Automata. Springer Berlin Heidelberg, pp. 234-250.
- [2] Cabasino, Simone & Pier S. Paolucci & Gian Marco Todesco, (1992) "Dynamic parsers and evolving grammars." ACM Sigplan notices 27.11, pp. 39-48.
- [3] Burshteyn, Boris, (1990) "Generation and recognition of formal languages by modifiable grammars." ACM SIGPLAN Notices 25.12, pp. 45-53.
- [4] Rubinstein, Roy S. & John N. Shutt, (1995) "Self-modifying finite automata: An introduction." Information processing letters 56.4, pp. 185-190.
- [5] Neto, João José, (1994) "Adaptive automata for context-dependent languages." ACM Sigplan Notices 29.9, pp.115-124.
- [6] Pistori, Hemerson & Priscila S. Martins & Amaury Antonio de Castro-Jr., (2005) "Adaptive finite state automata and genetic algorithms: Merging individual adaptation and population evolution". Springer Vienna.
- [7] Almeida-Jr, Jorge Rady (1995) "STAD: Uma ferramenta para representação e simulação de sistemas através de statecharts adaptativos". Tese de Doutorado, Escola Politécnica, Universidade de São Paulo, São Paulo (in portuguese).
- [8] Neto, João José & Bruno Arantes Basseto, (1999) "A Stochastic Musical Composer Based on Adaptive Algorithms." Proceedings of the 6th Brazilian Symposium on Computer Music-SBC&M99, Rio de Janeiro.
- [9] Iwai, Margarete Keiko, (2000) "Um formalismo gramatical adaptativo para linguagens dependentes de contexto." Departamento de Computação e Sistemas Digitais (PCS)-Escola Politécnica, Tese de Doutorado, Escola Politécnica, Universidade de São Paulo (USP), São Paulo, SP (in portuguese) .
- [10] Pedrazzi, Thiago Carvalho & Angela Hum Tchemra & Ricardo L. Azevedo Rocha, (2005) "Adaptive Decision Tables A Case Study of their Application to Decision-Taking Problems". Springer Vienna, pp. 341-344.
- [11] Rocha, Ricardo L. Azevedo & João José Neto, (2001) "Autômato Adaptativo, Limites e Complexidade em Comparação com Máquina de Turing". In: Proceedings of the second Congress of Logic Applied to Technology – LAPTEC'2000. São Paulo: Faculdade SENAC de Ciências Exatas e Tecnologia, pp. 33-48 (in portuguese).

- [12] Rodrigues, Elisângela Silva da Cunha & Fabricio Augusto Rodrigues & Ricardo Luis Azevedo Rocha, & Pedro Luiz Pizzigatti Correa, (2011) "Adaptive Approach for a Maximum Entropy Algorithm in Ecological Niche Modeling." *Latin America Transactions, IEEE (Revista IEEE America Latina)* 9.3, pp. 331-338.
- [13] Stange, Renata Luiza & Teresa Cristina Giannini & Fabiana Soares Santana & João José Neto & Antonio Mauro Saraiva (2011) "Evaluation of Adaptive Genetic Algorithm to Environmental Modeling of Peponapis and Cucurbita." *Latin America Transactions, IEEE (Revista IEEE America Latina)* 9.2, pp.171-177.
- [14] Hirakawa, Andre Riyuiti & Antonio Mauro Saraiva & Carlos Eduardo Cugnasca, (2007) "WTA 2007-III. 1-Adaptive Automata Applied on Automation and Robotics (A4R)." *Latin America Transactions, IEEE (Revista IEEE America Latina)* 5.7, pp. 539-543.
- [15] Pelegrini, Eder José & João José Neto, (2008) "A dynamically variable code execution model, based on adaptive automata." *Latin America Transactions, IEEE (Revista IEEE America Latina)* 6.5, pp. 424-435.
- [16] Silva, Salvador Ramos Bernardino da & João José Neto, (2011) "Proposal of a High-Level Language for Writing Self Modifying Programs." *Latin America Transactions, IEEE (Revista IEEE America Latina)* 9.2, pp. 192-198.
- [17] Sabaliauskas, Jorge Augusto & Ricardo Luis Azevedo da Rocha, (2011) "Project and Implementation for a Programming Language Suitable to Express Adaptive Algorithms." *Latin America Transactions, IEEE (Revista IEEE America Latina)* 9.6, pp. 969-973.
- [18] Almeida-Jr, Jorge Rady & João José Neto, (1999) "Using Adaptive Models for System Description." In: *IASTED International Conference Applied Modelling and Simulation, 1999, Cairns. Applied Modelling and Simulation*, p. 452-457.
- [19] Neto, João José & Jorge Rady de Almeida-Jr, (1999) "Modeling Adaptive Reactive Systems." *International Conference on Applied Modelling and Simulation. Cairns, Australia:[sn]*.
- [20] Neto, João José & Jorge Rady Almeida-Jr & José Maria Novaes dos Santos, (1998) "Synchronized statecharts for reactive systems." In: *Proceedings of the IASTED International Conference on Applied Modelling and Simulation. Honolulu, Hawaii*, pp. 246-251.
- [21] Genesereth, Michael & Nathaniel Love & Barney Pell, (2005) "General game playing: Overview of the AAI competition." *AI magazine* 26.2: 62.
- [22] Cazenave, Tristan, (2009) "Nested Monte-Carlo Search." *IJCAI. Vol. 9. 2009*.
- [23] Swiechowski, Maciej & Jacek Mandziuk, (2013) "Self-Adaptation of Playing Strategies in General Game Playing." *IEEE Trans. Comput. Intell. AI Games*.
- [24] Méhat, Jean & Tristan Cazenave, (2010) "Combining uct and nested monte carlo search for single-player general game playing." *Computational Intelligence and AI in Games, IEEE Transactions on* 2.4, pp. 271-277.
- [25] Bjornsson, Yngvi & Hilmar Finnsson, (2009) "Cadiaplayer: A simulation-based general game player." *Computational Intelligence and AI in Games, IEEE Transactions on* 1.1, pp. 4-15.
- [26] Schiffel, Stephan & Michael Thielscher, (2007) "Fluxplayer: A successful general game player." *Proceedings of the National Conference on Artificial Intelligence. Vol. 22. No. 2. Menlo Park, CA; Cambridge, MA; London; AAI Press; MIT Press*.
- [27] CLUNE, Jame, (2007) "Heuristic evaluation functions for general game playing." In: *AAAI, vol. 7*, pp. 1134-1139.

AUTHORS

José Maria Novaes dos Santos, graduated in Applied Mathematic in São Paulo University (USP), Brazil (1989) and MSc in Computer Engineering in the Polytechnical School (EPUSP) of Sao Paulo University (USP), Brazil (1997). He is doctoral student in Computer Engineering in the Polytechnical School (EPUSP) of Sao Paulo University (USP), Brazil. His main interests are adaptive devices, adaptive technology, machine learning, system modeling, computer learning, ontology and natural language processing.



João José Neto graduated in Electrical Engineering (1971), MSc in Electrical Engineering (1975) and doctor in Electrical Engineering (1980), and "livre docente" associate professor (1993) in the Polytechnical School of Sao Paulo (EPUSP) University. Nowadays he is the head of LTA - Adaptive Technology Laboratory at the Department of Computer Engineering and Digital Systems at the EPUSP. His main experience is in the Computer Science area, with emphasis on the foundation of computer engineering and adaptivity. His main activities include adaptive devices, adaptive technology, adaptive automata and their applications to computer engineering and other areas, especially in adaptive decision making systems, natural language processing, compiler construction, robotics, computer education, intelligent system modeling, computer learning, pattern matching, inference and other applications founded on adaptivity and adaptive devices.



INTENTIONAL BLANK

PERFORMANCE EVALUATION OF VANETS ROUTING PROTOCOLS

Abduladhim Ashtaiwi¹, Abdusadik Saoud² and Ibrahim Almerhag¹

¹College of Information Technology, University of Tripoli,

²Electrical and Computer Engineering, Libyan Academy

¹a.ashtaiwi@it.uot.ly.edu.ly,

¹i.almerhag@it.uot.ly.edu.ly, ²elsadik.bensaoud@yahoo.com

ABSTRACT

Lately, the concept of VANETs (Vehicular Ad hoc Networks) has gotten a huge attention as more wireless communication technologies becoming available. Such networks are expected to be one of the most valuable technologies for improving efficiency and safety of the future transportation. Vehicular networks are characterized by high mobility nodes which pose many communication challenging problems. In vehicular networks, routing Collision Avoidance Messages (CAMs) among vehicles is a key communication problem. Failure in routing CAMs to their intended destination within the time constraint can render these messages useless. Many routing protocols have been adapted for VANETs, such as DSDV (Destination Sequenced Distance Vector), AODV (Ad-hoc On demand Distance Vector), and DSR (Dynamic Source Routing). This work compares the performance of those routing protocols at different driving environments and scenarios created by using the mobility generator (VanetMobiSim) and network simulator (NS2). The obtained results at different vehicular densities, speeds, road obstacles, lanes, traffic lights, and transmission ranges showed that on average AODV protocol outperforms DSR and DSDV protocols in packet delivery ratio and end-to-end delay. However, at certain circumstances (e.g., at shorter transmission ranges) DSR tends to have better performance than AODV and DSDV protocols.

KEYWORDS

VANETs, MANETs, Routing protocols, ITS

1. INTRODUCTION

The World Health Organization (WHO) presented a report on road safety covers 182 countries which account for almost 99 % of the world's population. The report indicated that worldwide the total number of road traffic deaths remains unacceptably high at 1.24 million per year [?]. The advances in many technologies have helped many countries around the world implement plans to reduce the road traffic fatalities. Vehicular Ad hoc Networks (VANETs) is very promising that plays an important role in Intelligent Transportation System (ITS). VANETs assist vehicle drivers to communicate (through enabling Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communications) to avoid many critical driving situations. VANETs supports variety of safety applications such as co-operative traffic monitoring, control of traffic flows, blind crossing, prevention of collisions, nearby information services, and real-time detour routes computation. VANETs consist of two entities: access points, called Road Side Units (RSUs), and vehicles,

called On Board Unit(OBUs). RSUs are fixed and can act as a distribution point for vehicle networks. Figure ?? shows VANETs communication model of VANETs system.

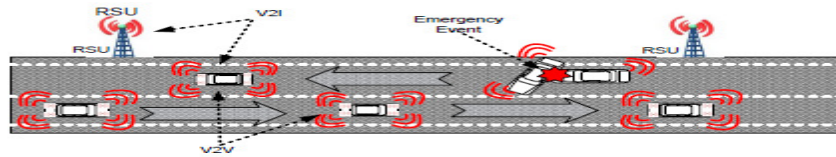


Fig. 1: VANETs Connection topology

In addition to road safety applications, VANETs provide Internet connectivity to vehicles while on the move, so passengers can download music, send emails, book restaurants and/or play games.

Due to vehicle's high speed, vehicular networks are characterized by rapid topology changes. The latter makes designing an efficient routing protocols for vehicular environment very difficult. Designing an adaptive routing protocols to such rapidly changing network typologies is very critical to many vehicular safety applications as failing to route collision avoidance messages to their intended vehicles can render these messages to be useless.

Considerable effort has been spent in performance comparison of VANETs routing protocols: the authors in [?] evaluated the performance of Ad hoc On demand Distance Vector (AODV) and Optimized Link State Routing Protocol (OLSR) for VANETs in city environment, in their study all the characteristics are handled through the vehicle mobility model. Their study showed that OLSR has better performance than AODV in the VANETs, as the AODV protocol has higher routing overhead compared to OLSR. The performance analyses of traditional ad-hoc routing protocols like AODV, Destination Sequenced Distance Vector (DSDV), and Dynamic Source Routing (DSR) for some highway scenarios have been presented in [?]. The authors argued that these routing protocols are not suitable for VANETs. Their simulation results showed that these conventional routing protocols have higher routing overhead which cause less packet delivery ratio. The work in [?] compared AODV and DSR with SWARMIntelligence based routing protocol by varying mobility, load, and size of the network. Their results showed that AODV and DSR have less performance than swarm intelligence routing algorithm in VANET environments. The authors in [?], [?] compared the performance of the routing protocols: AODV, DSR, Fisheye State Routing (FSR) and Temporally-Ordered Routing Algorithm (TORA), in city traffic scenarios. Their results showed that both protocols AODV and DSR have the lowest routing overheads and deliver packets quite fast.

Most previous studies on VANETs routing protocols focused on single driving environment. Therefore, in our study we focus on evaluating these protocol at different environments, i.e., downtown, residential, and suburban. Moreover, the performance of different routing protocols have not been well measured since each researcher used different simulator and performance metrics for performance evaluation. Due to aforementioned problems, there is continuous need to study various ad hoc routing protocols in order to select appropriate routing protocols for different driving environments of VANETs. In this work we evaluate the performance of DSDV, AODV, and DSR in different driving scenarios using the mobility generator (VanetMobiSim) and network simulator (NS2) to model all the driving environment and networking details of the vehicular ad hoc networks.

The remaining of this work is organized as follows: in Section ?? we give background about the objectives behind VANETs and how the standard work is produced. Section ?? classifies the routing protocols and shows their scope and structure. Mobility generator and network simulator tools which used to create different driving environments and scenarios are explained in Section ??. Section ?? defines the scope and structure of the simulation model framework. The obtained results are presented and analyzed in Section ??. Finally our concluding remarks are presented in our conclusion in Section ??.

2. VEHICLE ADHOC NETWORKS (VANETS)

Vehicles independently produce and analyze large amount of data such as time, heading angle, speed, acceleration, position, brake status, steering angle, headlight status, turn signal status, vehicle length, vehicle width, vehicle mass, and even the number of occupants in the vehicle. This data is selfcontained within a single vehicle. VANETs enable vehicles to share this data among themselves and with the road infrastructure. This shared driving information can then be used to implement many road safety applications that help to avoid many critical driving situations such as road side accidents, traffic jams, speed control, free passage of emergency vehicles and unseen obstacles and etc. In October 1999, the United States Federal Communications Commission (FCC) allocated 75 MHz of spectrum in the 5.9 GHz band to Dedicated Short Range Communications (DSRC). As shown in Figure ?? two standards are primarily involved: the IEEE 1609 standards (which defines the communications services and also known as Wireless Access in Vehicular Environment (WAVE)) and IEEE 802.11 p (which defines the physical and medium Application Layer access layer details). The IEEE 1609 standard breaks down into the following components: 1609.1 (WAVE resource manager), 1609.2 (WAVE security services for applications and management messages), 1609.3 (WAVE networking services), and 1609.4 (WAVE multi-channel operations). The IEEE task group "P" has approved the IEEE 802.11p amendment of IEEE 802.11 standard to support VANETs applications. The main enhancements include short latency and higher ranges, up to 1000 meters.

Application Layer	IEEE P1609 WAVE
Presentation Layer	
Session Layer	
Transportation Layer	
Network Layer	IEEE 802.11p DSRC
Data Link Layer	
Physical Layer	

Fig. 2: protocol Stack

3. ROUTING PROTOCOLS FOR VANETS

The main goal of routing protocols is to provide optimal paths between network nodes at minimum overhead possible. Figure ?? classifies the routing protocols into: topologybased and position-based routing protocols. In topology-based routing, each node should be aware of the

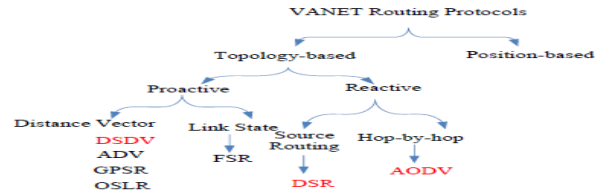


Fig. 3: Routing Protocols hierarchy

Network layout also be able to forward packets using information about available nodes and links in the networks. Topology-based routing protocols use link's information which stored in the routing table as a basis to forward packets from source to destination node; they are commonly classified into two categories (based on their underlying architecture): Proactive (periodic) and Reactive (on-demand) routing protocols.

Proactive protocols (also called table driven protocols) allow a network node to use the routing table to store routes information for all other nodes in the network, each entry in the table contains the next hop used in the path to the destination, regardless of whether the route is actually needed or not. To reflect the network topology changes, the proactive protocols frequently update their routing table. The topology changes are broadcasted periodically to all neighbors. In proactive routing protocols routes to destinations are always available when needed. Proactive protocols usually depend on shortest path algorithms to determine which route is chosen. They generally use two routing strategies: Link State (LS) strategy and Distance Vector (DV) strategy. The most representative are DSDV [?], ADV (Adaptive Distance Vector) [?], GPSR (Greedy Perimeter Stateless Routing) [?] and OLSR [?].

Reactive routing protocols (also called on-demand) reduce the network overhead by maintaining routes only when needed. The source node starts a route discovery process if it needs a non-existing route to a destination. It does this process by flooding the network by a route request message. After the message reaches the destination node or to the node that has a route to the destination, the receiving node sends a route reply message back to the source node using unicast communication. Depending on how the routing method is implemented, reactive routing protocols can be divided into source routing protocols and hop-by-hop or point-to-point protocols.

In source routing protocols every data packet carries the whole path information in its header. Before a source node sends data packets, it must know the total path to the destination, that is, all addresses of intermediate nodes which compose the path from source to destination. There is no need that intermediate nodes update their routing tables, since they only forward data packets according to the header information. The most representative source routing protocol is DSR [?].

On the other hand, hop-by-hop routing protocols try to improve the performance by keeping the routing information in each node. Every data packet does not include the whole path information any more. They only include the address of the following node where data packet must be forwarded to get the destination as well as the destination address. Every intermediate node, along the path, must look up its own routing table to forward the data packets to the intended destination, so that the route is calculated hop-by-hop. The most representative hop-by-hop routing protocol is AODV [?].

4. VEHICULAR SCENARIOS AND ENVIRONMENTS MODEL

In this work we used VanetMobiSim simulator which models all the VANETs traffic and environment details. The Vanet- MobiSim framework includes a number of mobility modules, parsers for geographic data sources in various formats, as well as visualization module. The framework is based on the concept of pluggable modules so that it is easily extend the model to cover many traffic and environment details. The chosen scenario is based on random street configurations. While vehicles are distributed randomly and move in a random direction. Roads in different driving environments are configured with single-lane as well as multi-lane.

The simulated scenario also includes different traffic flows and traffic lights located in different places. Several speeds have been selected as well as level of congestion. The selected mobility pattern is Random Waypoint mobility (RWP) with obstacles avoidance [?], in which vehicles move randomly and freely without restrictions. In addition, vehicles motion is enhanced with IntelligentDriving Model (IDM) which incorporates intersection management and lane changing mechanism [?]. IDM with Intersection Management (IDM-IM) module describes perfectly vehicle-to-vehicle and intersection managements. This module allows vehicles to adjust their speed based on the movements of neighboring vehicles (e.g., if a vehicle in frontbrakes, the succeeding vehicles also slow down and stop at intersections, or act according to traffic lights). IDM with Lane Changing (IDM-LC) is an overtaking module which interacts with IDM-IM to manage lane changes, vehicle accelerations, and deceleration. Based on IDM-LC module, vehicles are able to change lane and perform over takings in presence of multilane roads. Continuous bit rate (CBR) traffic sources are used in vehicles. The source-destination pairs are spread randomly over the simulation area. The number of source-destination pairs and the packet sending rate in each pair is varied to change the offered load in the network. 32-byte data packets are used. Table ?? lists the mobility model parameters and their values used in the simulation.

TABLE I: Mobility modele parameters

Parameter	Value
Number of roads managed by traffic lights	6 roads
Maximum number of roads with multi-lane	4 roads
Number of lanes in multi-lane roads	2 lanes
Time interval between traffic light changes	10000 <i>ms</i>
Minimum initial stay duration of the vehicle	5 Seconds
Maximum stay duration of the vehicle	30 Seconds
Step for recalculating movement of the vehicle	0.1 Seconds
The length of vehicle	5 Meter
Vehicle acceleration	0.6 (<i>m/s²</i>)
Vehicle deceleration	0.5 (<i>m/s²</i>)
Minimal distance to a standing vehicle (i.e., Jam distance)	2 Meters
Safe time headway	1.5 Seconds

To model the vehicular driving environment as close as possible to the real world environment, in our created model we divide the simulation area into three clusters: downtown, residential, and suburban. Clustering in VanetMobiSim tool are used to create different simulation areas with different driving environment and obstacles. Figure ??depict the different simulation areas: downtown, residential, and suburban. The clustering density parameter describes how many clusters per squared area (i.e., clusters/m²). In this model we set the cluster density to be 4 clusters per 1000000 m² which is 250000 m² per cluster area. Based on the simulated area which is

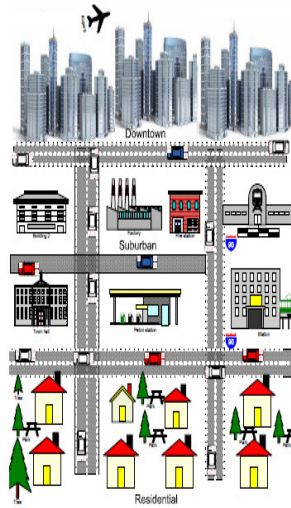


Fig. 4: Different driving environments

TABLE II: simulation area configurations

Cluster description	Number clusters	obstacles
Downtown	1	50
Residential	5	62
Suburban	16	15
Total	12	127

3000000 m² we have 12 clusters in our simulation model. Table ??lists the number clusters and obstacles configured in each simulation area.

5. SIMULATION MODEL FRAMEWORK

Figure ?? depicts the simulation process framework which is divided into three stages: stage 1, 2, and 3. In stage 1, we first define the scenario by writing all the vehicle mobility and environment details using XML file. Next we run the VanetMobiSim simulator to generate vehicular traffic trace file which contain all the details related to vehicular network including environment details such as node identifier, time, position, speed and etc. Figure ?? depicts a snapshot of the created model using VanetMobiSim at one instance time. The generated trace file is going to be the input to NS2 simulator in stage 2. In stage 2 we writing the details related to communications and network configuration using script Tcl programming language. The scripting files from stage 1 and 2 are used to run NS2 simulator. At stage 3, after running the NS2 simulator, the NS2 tool generates two files: Network Animator (NAM) file (*.nam) and a trace file (*.tr) as the outputs. The NAM file records all the positioning and graphical information performed during the simulation time. The trace file (.tr) (generated by NS2) contains all of the information about the simulation, e.g. packets sent, received, dropped, attached sequence number, protocol type, packet sizes, and etc. The trace file is simply available in a text format and could be called as a log file of the simulation.

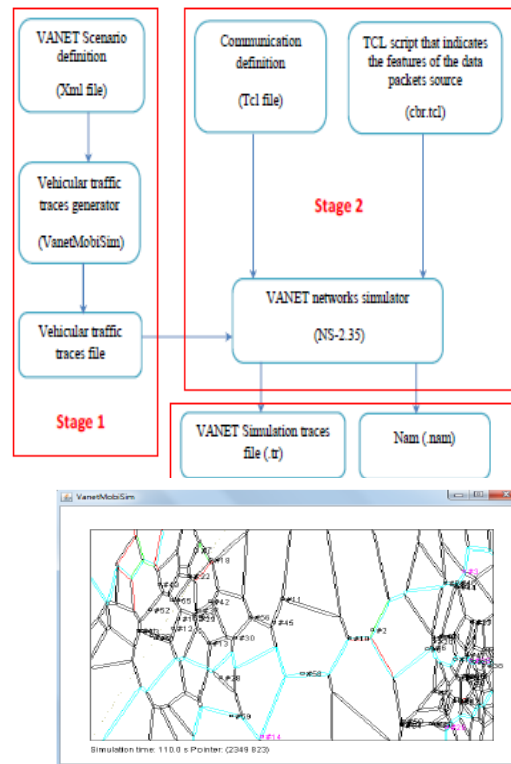


Fig. 6: Simulated vehicular environment snapshot

To extract the statistics (such as transmitted/received bytes and packet loss) from the generated file of NS2, we utilize AWK tool.

6. PERFORMANCE EVALUATION

The results obtained from modeled traffic and environment of vehicular networks using VanetMobisim and NS2 simulators are presented and analyzed in this section. We evaluate the scenarios using global metrics such as packet delivery ratio and end-to-end delay. The results are studied at different driving environment parameters such as node's density, speed and/or transmission range. We used different node densities (i.e., 20 to 45 nodes/km²) while the speed of the nodes are configured randomly between 30 to 50 km/h.

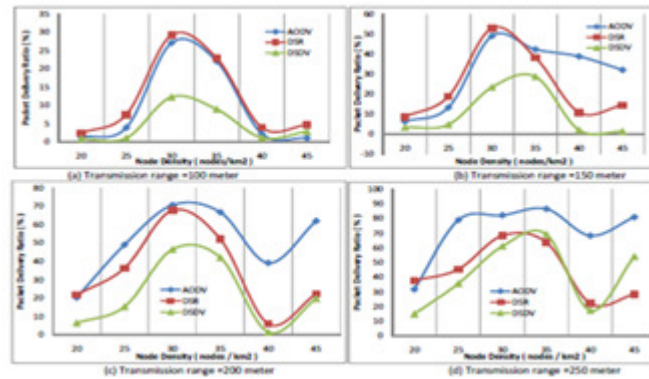


Fig. 7: Average packet delivery ratio vs node densities

Figure 7 shows the average packet delivery ratio versus node densities at different transmission ranges. The adaptability of reactive protocols to the rapid network topology changes of VANETs networks, make AODV and DSR outperform DSDV protocol, the latter protocol is belong to proactive protocols category. We notice that at different node density levels, the performance of three protocols produce the bell shape during the middle values of network node densities. That is because for small node densities (i.e., at sparse networks) only few nodes are available for routing functions. This shortage in forwarding nodes reduces the probability of finding multiple paths between the sources and the destinations. As the vehicular network density increases (e.g., from 30 to 35 nodes/km²), the probability of finding multiple paths through multiple intermediate nodes increases which can result in higher packet delivery ratio. However, as vehicular network density increases (i.e., beyond the optimum range) the performance decreases. That is because higher node densities make more intermediate nodes available for routing, this can produce higher number of paths characterized by higher number of hops compared with small density nodes. Higher number of hop count increases the probability of packet collision and loss as for each hop there is a chance of packet loss introduced by medium access or any other channel related parameters such as fading. Figure 7 shows that at different transmission ranges, the evaluated routing protocols tend to have slightly different performance. For example, when we increase the transmission range (as in Figure 7(c) and 7(d)), the AODV protocol achieve higher packet delivery ratio compared to DSDV and DSR protocols. That is because in AODV protocol all the intermediate nodes share the routing load, i.e., every node along the path uses fresh and updated routing information to forward the packets. However, DSDV and DSR protocols do not seem to gain a substantial improvement at higher transmission ranges. The poor performance of DSDV protocol is resulted from the protocol trying to maintain network connectivity by flooding the network for any topology change. Higher number of paths can cause DSDV protocol to produce higher network overload which turn to overall poor performance. Being source routing protocol, packets in DSR protocol carries the whole path routing information. Intermediate nodes only forward the packets based on the loaded information in the header. For any path break the source node has to use another path stored for the same destination or flood route request if none is available. Higher number of paths and/or hops per each path can also result in network performance degradation in DSR protocol.

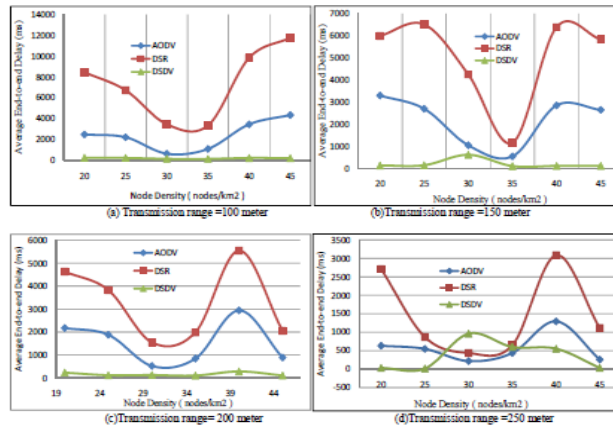


Fig. 8: End-to-End delay vs node density

Figure ?? shows the average end-to-end delay at different node densities and transmission ranges. At all ranges and network densities used, DSDV protocol tends to have the lowest end-to-end delay compared to AODV and DSR protocols. Comparing these results to the packet delivery ratio in Figure ??, we observe that DSDV also has the lowest packet delivery ratio. As explained above, DSDV is proactive protocol where every node in the network stores routing path for every node in the network. Due to the fast topology change of VANETs environment, DSDV can not maintain valid paths to the other destinations specially paths with multiple hops. The stale routing table entry of DSDV protocol, make the latter protocol forwards packets toward broken links. Due to this, the successfully delivered packets are only those of few number of hops which then can result in small end-to-end delay. It is apparent from Figure ?? that AODV outperforms DSR in terms of end-to-end delay at all network densities and transmission ranges. That is because in AODV protocol all nodes along the path share the routing load, where in DSR protocol only the source node is responsible for maintaining the whole path information. Larger header size of DSR (because of routing information stored in the header) consumes higher network capacity which then can result in higher network delay compared to AODV. As shown in Figure ??, the three evaluated protocols tend to have higher packet delivery ratio at the density range of 30-35 nodes/km². They similarly tend to have lower end-to-end delay for the same density range as shown in Figure ?. That is because higher packet delivery ratio means more packets have made it cross the network to the destination, this, as result, indicates the optimum availability of number of routing paths and network congestion during these values of node densities. As we increase the transmission range, the end-to-end delay is decreasing for all the evaluated protocols.

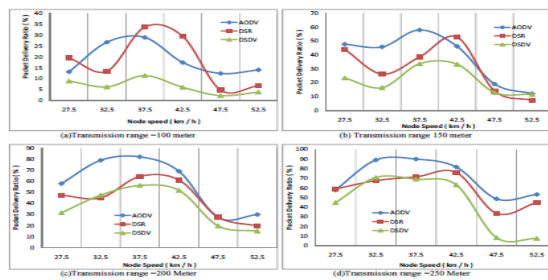


Fig. 9: Packet delivery ratio vs speed

Next we evaluate the performance the routing protocols at different vehicle speed ranges. In this scenario, the number of nodes is fixed to 100 nodes (i.e., 33.33 nodes/km²) and the average node speed increases from (25-30) km/h to (50- 55) km/h. Figure ??shows the average packet delivery ratio at different average node speed and transmission ranges (i.e., 100, 150, 200, and 250 meters).

It is apparent that, on average, all the evaluated protocols tend to perform better as the transmission range is increased. However, at different vehicle speeds the protocols start having different performance. For example, the performance of all protocols is degrading during high speed ranges. However, the adaptability of AODV protocol to higher network topology changes caused by high node speed, improve the performance of AODV compared to DSR and DSDV protocols in terms packet delivery ratio. Reactive source routing protocol (i.e., DSR) slightly perform better than AODV protocol at small transmission range and medium node speed. We think short transmission range can be the cause of higher link failures. Link failures trigger new route discoveries in AODV since it has at most one route per destination in its routing table. Thus, the frequency of route discoveries in AODV is directly proportional to the number of route breaks. The reaction of DSR to link failures in comparison is mild and causes route discovery less often. The reason is the abundance of cached routes at each node. Thus, the packet delivery ratio seems to be better for DSR during short transmission range and mild node speed.

7. CONCLUSION

VANETs is a mile stone enabling technology for ITS applications.VANET system supports many collision avoidance applications. Due to the high mobility of vehicular networks, many challenging problems are still open and requires more focus research. Routing collision avoidance messages in vehicular networks is a vital function for vehicular ad hoc networks. Failure to route these critical message to their desired destinations can make vehicles end up in road crashes. In this work we have studied three routing protocols: AODV, DSR, and DSDV. We used different clusters in VanetMobiSim simulator to create different vehicular driving environments: downtown,residential, and suburban areas. Each created area is characterized by different driving environment parameters: different road obstacles, road lanes, and/or traffic light. The obtained results showed that the routing protocols perform differently at different combination of transmission ranges, vehicular densities, and vehicle speeds. On average, the adaptability and the network load sharing of AODV protocol improved its performance compared to DSR and DSDV protocols. Although DSR protocol showed better performance, at certain values of simulation parameters, than AODV and DSDV protocols, these combination of parameters only represent vehicular environment for only certain cases of transmission range, vehicular density, and/or vehicle speeds.

REFERENCES

- [1] Xiong Wei, Li Qing-Quan, "Global status report on road safety," Decade of Action for Road Safety 2011-2020, Declared by the UN General Assembly, 2013.
- [2] Haerri, J., Filali, F., and Bonnet, C., "Performance Comparison of AODV and OLSR in VANETs Urban Environments under Realistic Mobility Patterns," Proceedings of the 5th IFIP Mediterranean Ad-Hoc Networking Workshop, pp. 14–17, June 2006.
- [3] Xiong Wei and Li Qing-Quan, "Performance Evaluation of Data Disseminations for Vehicular Ad hoc Networks in Highway Scenarios," The International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences, pp. 21–25, April 2008.

- [4] Manvi, S.S., Kakkasageri, M.S., and Mahapurush, C.V., "Performance Analysis of AODV, DSR, and Swarm Intelligence Routing Protocols In Vehicular Ad hoc Network Environment," In International Conference on Future Computer and Communication, pp. 21–25, April 2009.
- [5] Iwata A. et al., "Scalable Routing Strategies for Ad-hoc Wireless Networks," IEEE Journal on Selected Areas in Communications, pp.1369–79, Aug 1999.
- [6] Sven J., Marc B., and Lars W., "Evaluation of Routing Protocols for Vehicular Ad Hoc Networks in City Traffic Scenarios," in Proc of the 11th EUNICE Open European Summer School on Networked Applications, Colmenarejo, pp. 584–602, April 2005.
- [7] Perkins C.E. and Bhagwat P., "Highly Dynamic Destination-Sequenced Distance Vector Routing (DSDV) for Mobile Computers," SIGCOMM '94 Proceedings of the Conference on Communications Architectures, Protocols and Applications , pp. 234–244, Sep 1994.
- [8] Boppana R.V. and Konduru P., "An Adaptive Distance Vector Routing Algorithm for Mobile Ad Hoc Network," Proceedings of the Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies, pp. 1753–1762, Oct 2001.
- [9] Brad K. and Kung H. T., "Greedy Perimeter Stateless Routing for Wireless Networks," Proceedings of the 6th Annual International Conference on Mobile Computing and Networking , pp. 243–254, Oct 2000.
- [10] Clausen, T. and Jacquet, P., "Optimized Link State Routing Protocol (OLSR)," RFC Editor, vol. OLS:RFC3626, pp. 195–206, 2003.
- [11] D. Maltz D. Johnson, and Y. Hu., "The Dynamic Source Routing (DSR) Protocol for Mobile Ad hoc Networks for IPv4," URL <http://tools.ietf.org/html/rfc4728>, February 2007.
- [12] Perkins C., Belding-Royer E., and Das E., "Ad-hoc On-demand Distance Vector Routing," Mobile Computing Systems and Applications, 1999. Proceedings. WMCSA '99, pp. 90–100, Feb 1999.
- [13] Johanson P., Larsson T., Hedman N., Mielczarek B., and Degermark M., "Scenario-based Performance Analysis of Routing Protocols for Mobile Ad-hoc Networks," Proceedings of ACM/IEEE MOBICOM'99, pp. 195–206, Aug 1999.
- [14] Helbing, D., Hennecke, A., Shvetsov, V., and Treiber, M, "MicroandMacrosimulation of Freeway Traffic," Mathematical and Computer Modelling Journal, vol. 35, pp. 517–547, 2002.

INTENTIONAL BLANK

ANALOG SIGNAL PROCESSING SOLUTION FOR IMAGE ALIGNMENT

Nihar Athreya¹, Zhiguo Lai², Jai Gupta² and Dev Gupta²

¹Department of Electrical and Computer Engineering,
University of Massachusetts, Amherst, MA, USA,

²Newlans Inc., 43 Nagog Park, Acton, MA, USA

¹nathreya@umass.edu , ²zlai@newlans.com

ABSTRACT

Imaging and Image sensors is a field that is continuously evolving. There are new products coming into the market every day. Some of these have very severe Size, Weight and Power constraints whereas other devices have to handle very high computational loads. Some require both these conditions to be met simultaneously. Current imaging architectures and digital image processing solutions will not be able to meet these ever increasing demands. There is a need to develop novel imaging architectures and image processing solutions to address these requirements. In this work we propose analog signal processing as a solution to this problem. The analog processor is not suggested as a replacement to a digital processor but it will be used as an augmentation device which works in parallel with the digital processor, making the system faster and more efficient. In order to show the merits of analog processing the highly computational Normalized Cross Correlation algorithm is implemented. We propose two novel modifications to the algorithm and a new imaging architecture which, significantly reduces the computation time.

KEYWORDS

Analog Signal Processing, Parallel Architecture, Image Alignment, Stereo Correspondence & Normalized Cross Correlation.

1. INTRODUCTION

The imaging and image sensor industry is going through a huge wave of change. Very soon there is going to be a great demand in the market for wearables like smart watches, headbands, glasses etc. Cameras will be an integral part of these devices. However these devices have severe Size, Weight and Power (SWaP) constraints. On the other hand companies are also trying to develop multi-megapixel sensors and there have been talks of developing gigapixel sensors for use in defence, space and medical applications. Collecting such huge amounts of data and processing it is not an easy task. There are a lot of emerging applications in the field of computer vision, biometric analysis, bio-medical imaging etc. which require ultra-high speed computations. Another area that is gaining traction is the use of stereo cameras, camera arrays and light-field cameras to perform computational imaging tasks. Current imaging architectures and digital image processing solutions will not work in all of these situations because they will not be able to handle the high computational loads and meet the SWaP requirements simultaneously. Hence there is a need for novel ideas and solutions that can address these requirements.

In this paper we reintroduce the concept of analog image processing and present it as a solution to the above problems of reducing SWaP and high computational load. Generally the use of the term analog image processing has been restricted to film photography or optical processing. We are using neither of these approaches but we are performing analog signal processing by considering the image data to be a continuous stream of analog voltage values.

In order to show the advantages of analog processing we chose the problem of image alignment in stereo cameras. Image pairs captured from the stereo cameras can be used for a variety of purposes like constructing disparity and depth maps, refocusing, to simulate the effect of optical zoom etc. Image alignment can be used for stitching images to create panoramas, for video stabilization, scene summarization etc. Whatever the application, one of the most important steps in stereo image processing is to find correspondence between the points in the two images which represents the same 3D point in the scene. This has been an active area of research for many years now and there are a lot of stereo correspondence algorithms that have been developed. However some of these algorithms are either slow or have poor performance in the presence of noise or low light. The reason for these algorithms being slow is the very high computational requirement. In this work we pick one such stereo correspondence algorithm, Normalized Cross Correlation (NCC). NCC is very robust to noise and changes in the image intensity values but it is not preferred because of its high computational intensity. In this paper we propose two novel modifications to the algorithm which improves the computational speed without compromising the performance and also making it efficiently implementable in hardware. We also propose a new circuit architecture that can be used to implement the modified NCC algorithm in the analog domain. The analog domain implementation provides further speedup in computation and has lower power consumption than a digital implementation.

The organization of the paper is as follows. Section 2 gives a brief introduction to the NCC algorithm. Section 3 discusses the proposed modifications to the NCC algorithm. In section 4 a hardware circuit architecture for the implementation of the NCC algorithm is proposed. Experimental results are discussed in section 5. The work is concluded in section 6.

2. NORMALIZED CROSS CORRELATION

All stereo correspondence algorithms can be broadly classified into intensity based algorithms and feature based algorithms. In feature based algorithms features such as edges and contours are extracted from both stereo images and then a correspondence is established between them. In intensity based algorithms, blocks of pixels from one image are compared to blocks of pixels from other images and a similarity measure such as correlation or sum absolute difference is used to find the best matching block. Each of these algorithms have their own advantages and disadvantages. Both algorithms are used widely.

The Intensity based algorithms are very simple to implement, they are robust and they produce dense depth maps. They fail to perform well when the distance between the stereo cameras is too large or if there are rotations and shears in the stereo images. However the major drawback of the intensity based algorithms is that they are highly computational and hence it will be the algorithm of interest in this paper. In this study we address the issue of high computational load of the intensity-based algorithms through novel modifications to the algorithm and by the way of analog signal processing.

2.1. Review of Related Work

There has been a lot of research done on stereo image registration techniques as it relates to multiple fields like computer vision, medical imaging, photography etc. A variety of algorithms,

both feature based and intensity based have been developed. In [1], the author provides a survey of different image registration techniques used in various fields.

In this study we are mainly concerned with the implementation of an intensity based image alignment algorithm in hardware. There has been some work done in this regard but most of them are improvements to the old algorithms and some are digital hardware implementations of these algorithms.

In [2] Lewis proposes a fast normalized cross correlation algorithm, which reduces the computational complexity of the normalized cross correlation algorithm through the use of sum table methods to pre-compute the normalizing denominator coefficients. In [3] the authors take the fast normalized cross correlation algorithm one step further by using rectangular basis functions to approximate the template image. The number of computations in the numerator will then be directly proportional to the number of basis functions used to represent the template image. Using a smaller number of basis functions to represent the template image will certainly reduce the computation but it may give a bad approximation of the template image, which would result in poor image alignment. In [4] the author uses a pipelined FPGA architecture to perform the Normalized Cross Correlation operation. This increases the computation speed significantly.

There have been various other improvements and implementations of the NCC algorithm in literature however none of the implementations, to our knowledge, try to tackle the computational intensity problem of the normalized cross correlation algorithm from an analog signal processing perspective.

2.2. Reasons for Choosing Normalized Cross Correlation

There are a lot of intensity based stereo correspondence algorithms. We chose Normalized Cross Correlation (NCC) as the algorithm that we would implement because of the following reasons:

1. The images being aligned have translation in the X and Y direction but no rotation or shear. NCC algorithm performs well for such images.
2. NCC is less sensitive to variation in the intensity values of two images being aligned.
3. The NCC algorithm is computationally intensive. Hence it would be challenging to come up with methods to reduce the computation and make it implementable in real time or near real time and we believe analog signal processing would have a lot of value in such situations.

2.3. The General Algorithm

Template matching is one of the simplest methods used for image alignment. There are two images to be aligned. One image is called the template and the other image is called the reference. The template image is generally divided into blocks of smaller images. There is always a tradeoff between the depth accuracy that can be achieved and computation that can be handled in a NCC algorithm. Increasing the number of blocks by reducing the block size increases the accuracy to which depth can be estimated but it also increases the number of times the computations have to be performed. There is also a limit to which the block size can be reduced. If the block size is made too small then it might not have enough information to align with a matching block. Therefore choosing an optimum template block size is important.

Each template block is shifted on top of the reference image and at each point a correlation coefficient is calculated. This correlation coefficient will act as a similarity metric to identify the closest matching blocks.

The disadvantage of using cross correlation as a similarity measure is that it is an absolute value. Its value depends on the size of the template block. Also the cross correlation value of two exactly matching blocks may be less than the cross correlation value of a template block and a bright spot. The way around this problem is to normalize the cross correlation equation.

Equation (1) shows how the NCC algorithm is implemented [2].

$$C(u, v) = \frac{\sum_{x,y} [r(x,y) - \bar{r}_{u,v}] [t(x-u, y-v) - \bar{t}]}{\left\{ \sum_{x,y} [r(x,y) - \bar{r}_{u,v}]^2 \sum_{x,y} [t(x-u, y-v) - \bar{t}]^2 \right\}^{0.5}} \quad (1)$$

In the above equation \bar{t} represents the mean of the template image block and $\bar{r}_{u,v}$ is the mean value of the reference image present under the template image block. The two summation terms in the denominator of the above equation represent the variances of the zero-mean reference image and template image respectively. Due to this normalization, the correlation coefficient is independent of changes to image brightness and contrast.

The denominator of the NCC equation can be calculated efficiently through the use of sum tables as suggested in [2]. However the numerator of the NCC is still computationally intensive. A direct implementation of the numerator of NCC algorithm on a template image of size ($T_x \times T_y$) and a reference image of size ($R_x \times R_y$) would require ($T_x * T_y$) multiplications and additions for each shift (u,v). Reducing the template image size would reduce the number of computations per block but it will also increase the total number of blocks on which the NCC has to be performed.

3. MODIFICATIONS TO THE NCC ALGORITHM

In a general Normalized cross correlation algorithm the template image is divided into blocks and each block is shifted on top of the reference image. At each shift a normalized correlation coefficient is calculated. All the pixels in the block are used to perform this calculation as shown in equation (1). Once this is done for all shifts, a best matching block is picked and all the pixels in the template block are assigned the same shift/disparity value.

In the worst case scenario where there is no information available about the camera system or the scene, a brute force approach has to be used where the template image blocks have to be shifted all over the reference image. The computational complexity in this case would be very high. When some information is available about the camera system the maximum disparity that will be observed can be calculated and hence the number of shifts can be restricted. However this does not address the fact that the number of computations that have to be performed per block for each shift is still high.

A pre-processing step that is generally used in most stereo correspondence algorithms is image rectification. Image rectification projects stereo images onto a common reference plane so that the correspondence points have the same row coordinates. This essentially transforms the 2D stereo correspondence problem to 1D. However the rectification process itself will add to the computational complexity of the algorithm.

In order to further reduce the computation and address the above mentioned issues we decided to use only the diagonal elements of the template image block and the reference image blocks to compute the correlation coefficient. All the other steps in the algorithm are followed as given by equation (1). The thought behind this approach is that the diagonal elements of a block have enough information to calculate the disparity. By introducing this modification we have effectively converted the problem of 2D NCC operation to a 1D NCC operation. This is very similar to the image rectification operation but since we are choosing only the diagonal elements we introduce two advantages both of which contribute to the reduction in computation.

1. We are not using an algorithm to reduce the NCC operation from 2D to 1D, it is a natural result of the data selection process and hence it does not involve any additional computations.
2. Since we are choosing only the diagonal elements, the number of computations per block is reduced to a great extent i.e. if we have a template image of size $T_x \times T_y$ the total computations per block in the numerator now reduces from $(T_x \times T_y)$ additions and multiplications to only $T_x(=T_y)$ additions and multiplications per shift. This reduction in the computation is even more significant when the template and reference image sizes are very large.

Another advantage of the modified NCC algorithm is in situations where the information of the camera systems capturing the images is not known and brute force shifts have to be applied. This modified algorithm will be a good solution for such cases.

In some pathological cases the images or particular template blocks might not have a lot of features like edges and contours or the features may all be located on the upper or lower triangular side. In such situations just using the diagonal elements to perform the NCC algorithm might not work. The solution to this problem may be as simple as using the off-diagonal elements instead of the diagonal elements. However in most practical applications we always have images with some features on the diagonal element so using all the pixels in a template image block is unnecessary as it will only add to the computation without producing any significant improvements in the results.

4. HARDWARE ARCHITECTURE

In a standard CMOS image sensor there are photodiodes that produce electrons proportional to the amount of light intensity that strikes them. This is then converted into voltage levels which are read out by the readout circuitry. In order to remove noise, a process called correlated double sampling (CDS) is used. After this the signals are amplified. All this happens in the analog domain. These signals are then digitized using analog-to-digital converters (ADCs) and stored in memory or are sent to digital processors for further processing. So the signal for the most part is in the analog domain and we can utilize this to our advantage to perform analog processing.

With this in mind we have come up with a new imaging architecture which would best utilize the features of both analog and digital domains. Figure 1 shows a top level block diagram of the proposed architecture. In this architecture we have the digital system accessing the sensor and it is pre-processing, digitizing and storing the images in memory as before. However we now have an analog system that is accessing the analog data on the sensor, processing it and then feeding it into a digital machine for any further computations. By doing this we have separated the process of image acquisition which is being done by a digital system and image processing which is being done by an analog system. The biggest advantage of such an architecture is that they are operating in parallel i.e. the image acquisition is independent of the processing. This is not true in the case of completely digital systems. In a fully digital system the image processing operation cannot

start until the images have been completely acquired and stored in memory. In this hybrid system the analog block is performing the computationally intensive task of running the NCC algorithm as and when the image data is being read off the sensor. This means that by the time the digital system has acquired the images the analog processor would have finished its computation and the outputs will be ready to be used by the digital system.

Another important point to be noted is that the analog processing block is not in the signal plane but in the control plane. One of the biggest disadvantages of an analog system is the amount of noise added by it. However in this kind of an architecture the analog block is not responsible for signal acquisition and hence the problem of signal being corrupted by noise vanishes immediately.

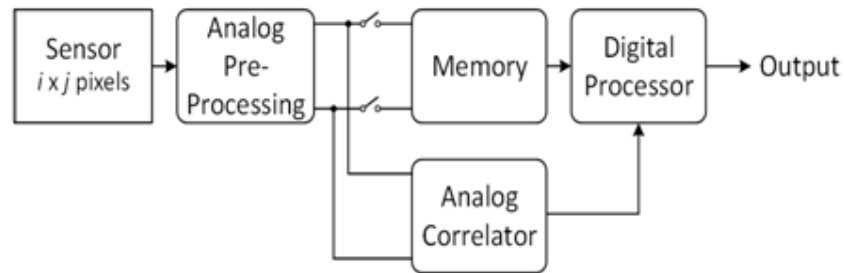


Figure 1: Proposed Architecture for the new Imaging systems

4.1. Implementing the modified NCC algorithm

Figure 2 shows the implementation of the modified NCC algorithm in the new imaging architecture. We have a digital system (shown on the left) which is accessing the sensor data, pre-processing, digitizing and storing it in memory. We have N analog channels which read the analog data from the sensor directly.

These N analog channels can be grouped into pairs in which one (odd numbered) channel is used to read the reference image data and the other (even numbered) channel is used to read template image data. The CMOS image sensors are capable of accessing individual pixel data. The readout circuitry is used to selectively read the diagonal elements of template and reference image blocks. Once the analog data has been read from the sensor we have it available to perform the NCC algorithm. According to eq. (1) in order to perform the NCC algorithm we first need a zero mean template image and zero mean reference image. In the case of digital systems it is not hard to compute the zero mean images from stored information. However in the new architecture we are directly accessing the analog data from the sensor and we would have to wait for an entire block of data to be read out in order to compute the mean and subtract it from the original signal. This would be a waste of processing time since this has to be done multiple times *i.e.* for each image block. In order to get around this problem we propose a second modification to the NCC algorithm. Here we use moving averages instead of regular averages and the moving average is subtracted from the original signal. The moving average circuit can be implemented as a low pass filter. The analog data from the sensor is passed through the moving average filter, the output of which is subtracted from the original signal. This is then fed to the multiplier and integrator which together perform the correlation operation. This calculates the numerator of the NCC algorithm. For calculating the denominator we plan to use the sum table method. This can be done efficiently by a digital system. So once the numerator calculation is done the analog signal is sampled and then fed into a DSP which normalizes the numerator and performs the decision making. The output will be disparity values in the X and Y direction. We also ran MATLAB simulations to

ensure that a change from an actual zero-mean image to one with moving averages does not affect the performance of the NCC algorithm.

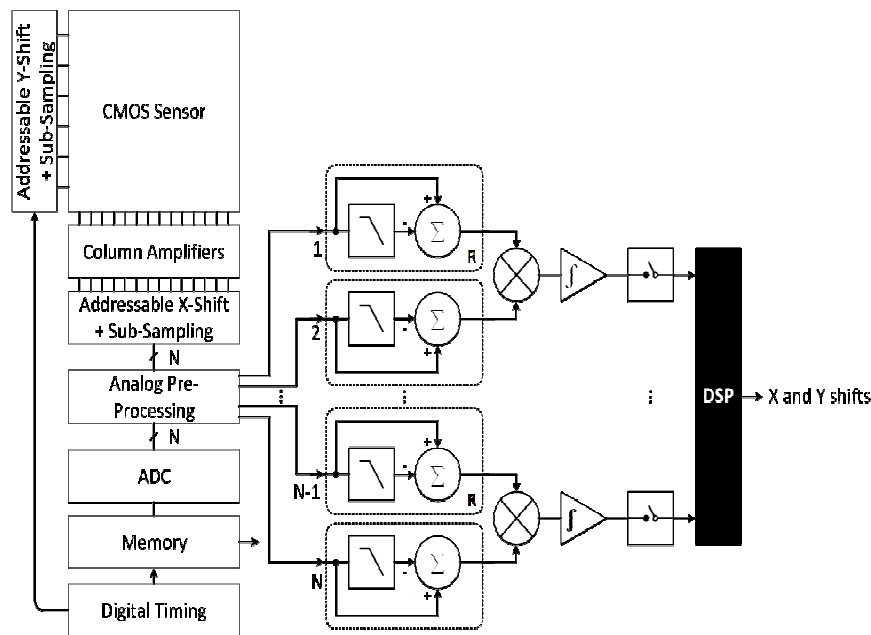


Figure 2: Implementation of NCC algorithm in the new Imaging architecture

5. EXPERIMENTAL RESULTS

In this section we compare the performance of the modified NCC algorithm to the original algorithm to show that the modified algorithm is faster and has a performance similar to the original algorithm. Since the modified algorithm has been developed to be implemented in analog hardware various other simulations are run that measures the performance of the modified algorithm.

We have run simulations on 15 sets of unrectified, grayscale stereo image pairs. These images have been captured under different illumination conditions which include incandescent light (In_Incd), outside bright light (Out_brt), outside low light (Out_clds), outside mixed shade lighting (Out_mxdsd). This allows us to test the performance of the algorithm in a more robust manner. All simulations have been done in MATLAB.

5.1. Cropping the template image

The stereo cameras have overlapping fields of view but the amount of overlap depends on the distance between the centres of the two cameras. In this work we have considered cameras whose centres are 6.5mm apart. Their hyperfocal distances are 70cm which means that all objects which are 35cm and beyond are in sharp focus. Since the distance between the centres are 6.5mm there will be some points in the scene that will be present in one of the images but not in the other. These points cannot be used for alignment and hence one of the images (template image) is cropped around the edges.

5.2. Evaluating the performance of the algorithms

The performance of the image alignment algorithms can be evaluated in a variety of different ways. Here the correlation coefficient is used as a performance measure. Once the final disparity values for all the template blocks are obtained, each template block is shifted by the disparity values obtained for that block. In order to get a uniform disparity variation across the entire image an interpolation technique is used. At the end of this process the two stereo images have been aligned. The correlation coefficient is calculated between the two aligned images and it is used as an indicator of the performance of the algorithm.

Figure 3 shows a comparison of the correlation coefficients obtained for 15 stereo image pairs by using the original NCC algorithm and the modified NCC algorithm. Each stereo image pair has a size of 1080x1920 pixels. The cropping of the template image around the edges is not more than 10% of the entire image size. The template block size chosen here is 128x128 pixels. As can be seen from the figure the performance of the modified NCC algorithm is very close to the original NCC algorithm. The performance was also tested for various other template block sizes and uniform performance was obtained for all. A speedup of 2x in MATLAB run time was observed for the modified NCC algorithm over the original algorithm.

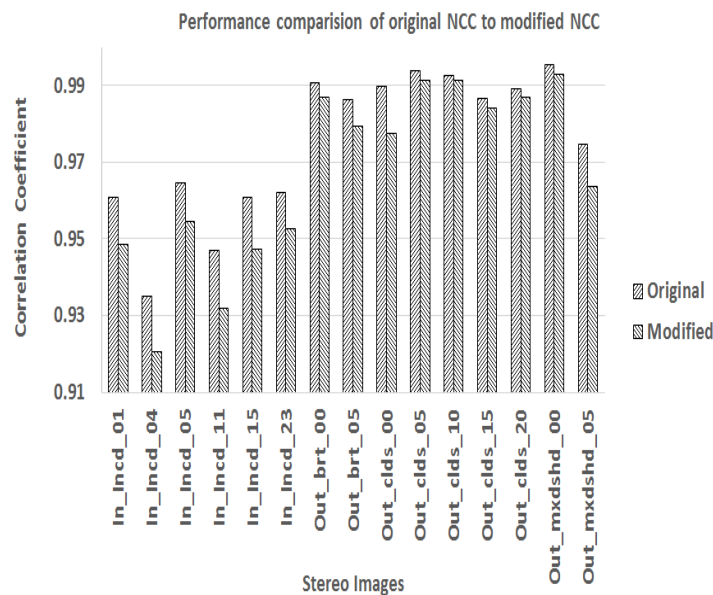


Figure 3: Performance comparison of the original NCC algorithm to modified NCC algorithm

It has to be noted that the objective of the above simulation is to compare the performance of the algorithms and not the digital and analog implementation of the algorithms.

Another performance measure that was used was the percentage improvement in the correlation coefficient values of the stereo images before and after alignment. Figure 4 shows these results. As it can be seen there is significant improvement in the correlation coefficients for all the images. It must be noted that for some images we see a very low percentage improvement. This is not because of the performance of the algorithm but because the input images were themselves almost aligned before the algorithm was applied.

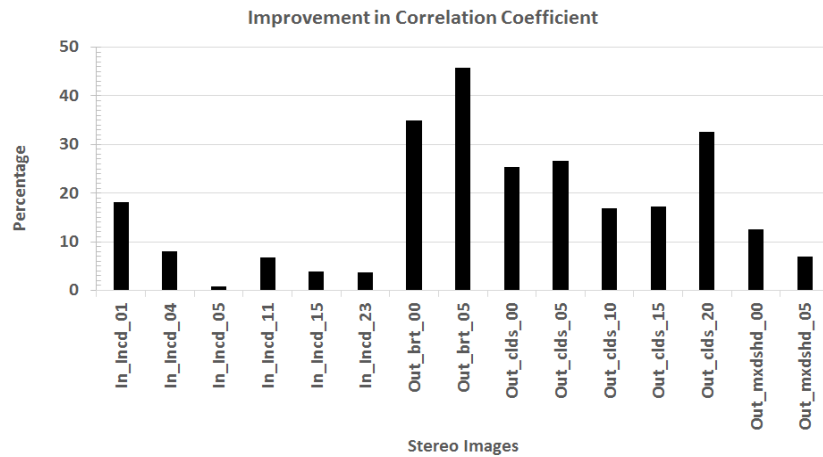


Figure 4: Percentage improvement in the correlation coefficient after alignment using modified NCC algorithm

As an example of the performance the modified NCC algorithm two figures 5 and 6 are shown. Figure 5 shows an overlap of a pair of stereo images before alignment. The areas of magenta and green show the areas of misalignment between the two images. The correlation coefficient measured for these two images before alignment is 0.7247. Figure 6 shows the overlap of two images after aligning them using the modified NCC algorithm. As can be seen from the figure there is hardly any misalignment between the two images. The correlation coefficient observed in this case is 0.9923.

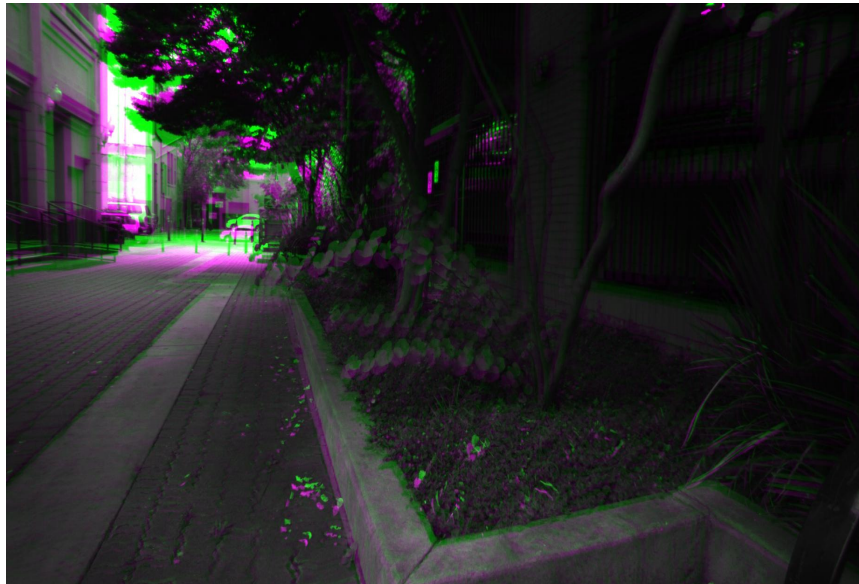


Figure 5: Overlap of two stereo images before alignment



Figure 6: Overlap of the two stereo images after alignment

Figure 7 and Figure 8 shows the disparity variation for the image shown in figure 5 and 6 in X and Y direction respectively. These disparity values have been obtained through the modified NCC algorithm. The disparity values have been colour coded and the colour bar indicates the different disparity values. The disparity values vary from 18 to 33 for Figure 7 in the X (horizontal) direction. The disparity values vary from 43 to 53 for Figure 8 in the Y (vertical) direction.

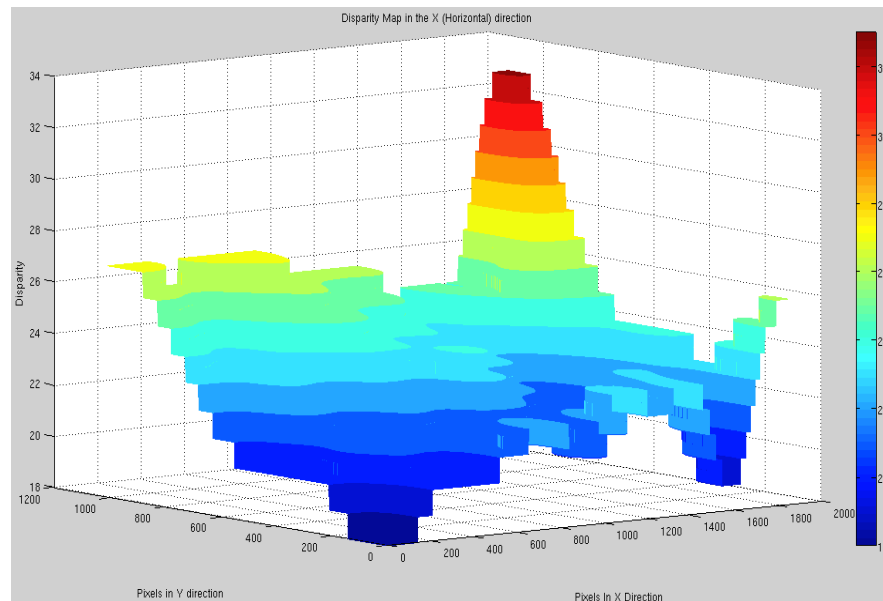


Figure 7: Disparity variation in the X(Horizontal) direction

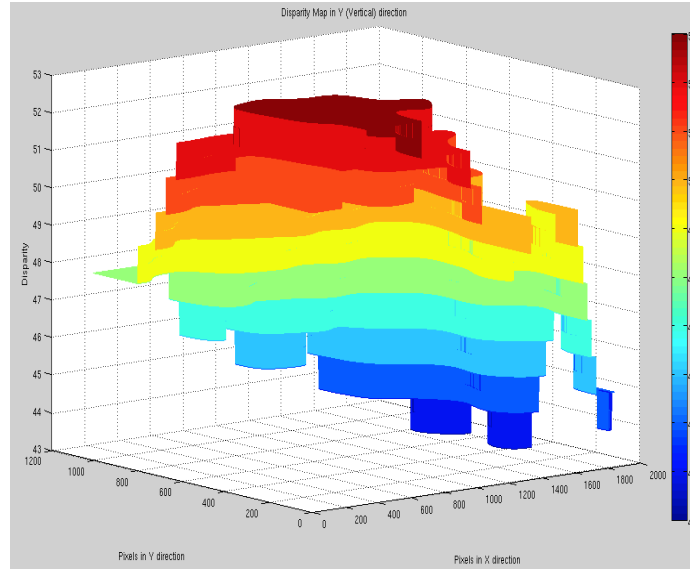


Figure 8: Disparity variation in the Y(Vertical) direction

5.3. Measuring the robustness of the modified NCC algorithm

We know that the NCC algorithm is robust to changes in the intensity values of the images. Here we try to measure the robustness of the modified NCC algorithm to changes in intensity values by changing the intensity values of one of the two stereo images. In the first case we reduced the intensity values of the template image by 90% uniformly across the image and used the modified NCC algorithm to align the images. This analysis addressed the fact that the illumination of a scene might change between the capture of two stereo images and the change in illumination was assumed to be uniform. However there might be rare situations where illumination on parts of the scene varies between captures of two stereo images. To address this issue we randomly varied the intensity values of the template image and analysed the performance of the modified NCC algorithm under these conditions as well. Figure 9 shows the results of these analyses. As it can be seen the modified NCC algorithm is very robust to changes in the image intensity values.

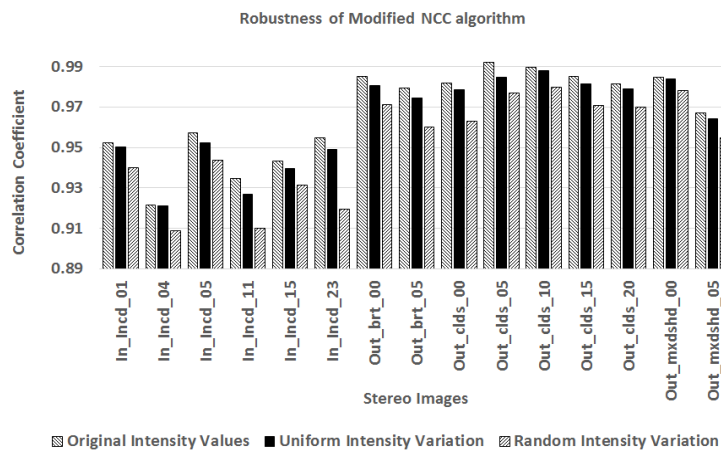


Figure 9: Robustness of the modified NCC algorithm to changes in image intensity values

Since the algorithm has been developed to be implemented in analog hardware it is very important to characterize the performance of the algorithm in the presence of noise. The two analog circuits that have been considered to be the primary contributors of noise are the multiplier and the integrator. In order to simulate the addition of noise by analog circuitry we first find the RMS value of the image intensity values. We multiply this RMS value by a number which indicates the percentage of noise being added by the circuit. This value is then multiplied by a random number picked from a Gaussian distribution. The outcome of this process is a noise value which is then added to the original image intensity. In our simulations it was found that the algorithm is more sensitive to the noise added by the multiplier than that by the integrator. Hence we maintain the noise added by the integrator at 20% and vary the amount of noise added by the multiplier. Figure 10 shows this performance variation. We have shown the performance for 3 different noise values added by the multiplier, 1%, 10% and 20%. As it can be seen the performance of the modified NCC algorithm is still good in the presence of noise added by the analog circuitry.

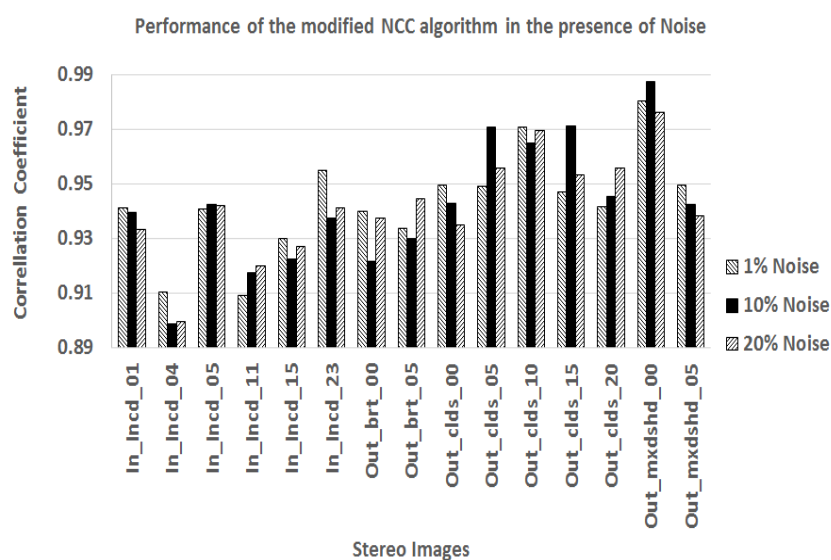


Figure 10: Performance of the modified NCC algorithm in the presence of Noise

5.4. Dynamic Range requirements and Power analysis

The amount of noise that can be tolerated by a circuit determines the dynamic range requirements for that circuit. The noise analyses done above will give us an idea of the dynamic range requirements for the analog circuits to perform the NCC operation. Analog circuits can easily achieve dynamic ranges of 40dB. This corresponds to a noise level of 1%. From Figure 10 we see that the modified NCC algorithm has excellent performance for a noise level of 1%.

The dynamic range requirements also dictates the power consumptions for analog circuits. The 4 major analog circuits required to implement the modified NCC operation are the low pass filter, analog summer, multiplier and integrator. We have performed initial circuit simulations for these components. Table 1 shown below gives approximate values of the power consumption for these components calculated for a dynamic range of 40dB based on these simulations. Here we assume that we have 64 analog channels in the architecture shown in Figure 2 and the number of components required are calculated based on that. The actual power consumptions can only be obtained once the components are realized and integrated.

Table 1: Approximate power consumption values for the analog circuits

<i>Component</i>	<i>Quantity</i>	<i>Power Consumption</i>
LPF	64	$2.8\text{mW}/\text{LPF} * 64 = 179.2\text{mW}$
Summer	64	$0.549\text{mW}/\text{sum} * 64 = 35.13\text{mW}$
Multiplier	32	$1.83\mu\text{W}/\text{mul} * 32 = 0.058\text{mW}$
Integrators	32	$0.024\text{mW}/\text{int} * 32 = 0.768\text{mW}$

The total power consumption by the analog circuitry is 215.15mW. Based on some of the digital implementations such as [11] and [12], we see that the power consumption for an analog implementation will be very low compared to that of a digital implementation. This shows that we have significant power savings as well.

5.5. A Note on Computation Time

The modifications proposed to the NCC algorithm contribute to a significant reduction in the computation of the algorithm. Simulation results show a 50% reduction in computation time for the modified NCC algorithm over the original algorithm. The other factors which add to the reduction of computation time are the novel imaging architecture and analog processing. In the new imaging architecture the analog processor works in parallel with the digital acquisition system and hence it does not have to wait for the entire image to be acquired before the processing starts. By the time the acquisition is done the analog processor would have finished its computation. So the image acquisition time can also be added towards the reduction in computation time. The implementation of the NCC algorithm is being done in analog hardware. The analog processor is not limited by the data converters (ADCs) or logic delays. The settling times of well-designed analog circuits are very small. Hence an analog implementation of the NCC algorithm would be faster than a digital implementation and would contribute towards a further reduction in computation time.

6. CONCLUSIONS

In this work we propose analog signal processing as a solution for handling the high computational load of some of the image processing algorithms while simultaneously meeting the reduced SWaP requirements. The analog processor will be used to augment the digital processor and work in parallel with it to perform key computations, making the system faster and more efficient. We implement a highly computational stereo correspondence algorithm to align stereo image pairs. Two novel modifications were proposed to the NCC algorithm which reduced the computation and made the algorithm efficiently implementable in analog hardware. The modified algorithm has a 50% reduction in MATLAB computation time over the original algorithm. The actual analog hardware implementation of the algorithm and the new imaging architecture will contribute to a further reduction in computation time as compared to a digital implementation. An approximate power consumption of 215.15mW for obtained for the analog correlation block. Various other simulations were also run to check the robustness and performance of the algorithm. The experimental results obtained are very promising and we believe analog processing will be a viable solution to these problems. As a part of the future work and as a proof-of-concept the analog image correlator circuit will be built from commercially available off the shelf components. A test plan will be setup for this circuit. Once the required results are obtained, the next step will be to build the architecture in silicon.

REFERENCES

- [1] Brown, L, (1992) "A survey of image registration techniques", ACM Journal (CSUR), Vol. 24, Issue 4, pp325-376.
- [2] Lewis, J.P, (1995) "Fast Normalized Cross Correlation", Industrial Light & Magic.
- [3] Briechele, Kai, & Uwe D. Hanebeck, (2001) "Template matching using fast normalized cross correlation", Aerospace/Defense Sensing, Simulation, and Controls. International Society for Optics and Photonics.
- [4] Radhamani & R, Keshaveni, (2012) "FPGA implementation of efficient and High speed template matching module", IJRTE, Vol. 2, Issue- 2.
- [5] Khaleghi, B et al. (2008) "A new Miniaturized Embedded Stereo-Vision System (MESVS-I)", CRV, pp26-33.
- [6] Gupta, N, (2007) "A VLSI architecture for Image Registration in Real Time", IEEE Transactions on VLSI systems, Vol. 15, Issue 9, pp981-989.
- [7] Roma, N et al. (2002) "A comparative analysis of cross-correlation matching algorithms using a pyramidal resolution approach", INESC Portugal.
- [8] Adhikari, G. et al. (2012) "Fast normalized cross correlation with early elimination condition", IEEE Transaction, ICRTIT 2012, pp136-140.
- [9] Tsai, Du-Ming & Chien-Ta Lin, (2003) "Fast normalized cross correlation for defect detection", Pattern Recognition Letters 24.15, pp2625-2631.
- [10] Szeliski, R, (2004) "Image Alignment and Stitching: A tutorial", Technical Report, Microsoft Research, Microsoft Corporation, MSR-TR-2004-92.
- [11] Ttofis, C., & Theocharides, T. (2012) "Towards accurate hardware stereo correspondence: A real-time FPGA implementation of a segmentation-based adaptive support weight algorithm", In Proceedings of the Conference on Design, Automation and Test in Europe, pp703-708.
- [12] Xiaotao Wang & Xingbo Wang, (2009) "FPGA Based Parallel Architectures for Normalized Cross-Correlation", 1st International Conference Information Science and Engineering (ICISE), pp225-229.
- [13] San Yong, Y., & Hon, H. W. (2008) "Disparity estimation for objects of interest", World Academy of Science, Engineering and Technology, 43.

AUTHORS

Nihar Athreyas received his B.E. degree in Electronics and Communication Engineering from VTU Belgaum, India and M.S. degree in Electrical and Computer Engineering from University of Massachusetts, Amherst, MA, in 2010 and 2013 respectively. He is currently pursuing his doctoral degree under the supervision of Dr. Dev Gupta at University of Massachusetts, Amherst, MA. He joined Newlans, Inc., Acton, MA in June of 2014 as an Intern. His current research interests include communications, CMOS analog design and applications of analog signal processing.



Zhiguo Lai received the B.S. degree in mechanical engineering from Tsinghua University, Beijing, China, the M.S. degree in electrical and computer engineering from University of Alaska, Fairbanks, AK, and the Ph.D. degree in electrical engineering from University of Massachusetts, Amherst, MA, in 1999, 2002, and 2007, respectively. From summer of 2007 to summer of 2009, he was a postdoctoral fellow at University of Massachusetts, Amherst, MA, working on narrowband interference mitigation for UWB systems. Since 2009, he has been with Newlans, Inc., Acton, MA. His research interests include design of programmable CMOS filters, wideband analog signal processing, and quantum computation. He is currently working on analog computation using deep-submicron CMOS technology.



Dev V. Gupta received his Ph.D. in 1977 from the University of Massachusetts, Amherst. He held various engineering positions at the Bell Laboratories in Andover, MA, from 1977 to 1985. He was the General Manager at Integrated Network Corporation, a manufacturer of DSL access products, from 1985 to 1995.



Dr. Gupta founded two companies, Dagaz Technologies and Maxcomm, which were acquired by Cisco Systems in 1997 and 1999 respectively. These companies developed and manufactured telephone exchange and voice and data equipment for DSL. He was Cisco's VP of Architecture in the access business unit between founding Dagaz Technologies and Maxcomm. In 2000, he founded Narad Networks which manufactured Gigabit Ethernet networking equipment for the cable industry. Narad Networks (renamed PhyFlex) was acquired by Cienna in 2007. Newlans was founded in 2003. He is a Charter Member of the Atlantic chapter of the Indus Entrepreneurs (TIE), an organization which promotes entrepreneurship. The World Economic Forum named him a 'Tech Pioneer' for the years 2001 and 2002. He is a Trustee of the University of Massachusetts, Amherst and a board member of the UMass Foundation. He is an Adjunct Professor at the University of Massachusetts where he created the Gupta Chair in the Department of Electrical and Computer Engineering. He has over thirty patents in communications, networking, circuit design, and signal processing. Dr. Gupta is the President and CEO of Newlans and he provides technical vision for the Company.

Jai Gupta currently serves as Chief Architect for commercialization efforts of Newlans' analog integrated circuits and wideband signal processing technology. His primary role is system architecture design and program management. He has previously held research positions at the National Institute of Standards and Technology and Huntington Medical Research Institute, worked as a systems engineer at metropolitan ad-hoc networking startup Windspeed Access, and held internship positions at broadband cable startup Narad Networks as well as Cisco Systems.

Jai recently completed the Executive MBA degree at Duke University's Fuqua School of Business with a concentration in Entrepreneurship and Innovation. He previously obtained the MSEE degree in Electrical Engineering Systems from the University of Southern California's Viterbi School of Engineering and the BSEE degree from the University of Pennsylvania's Moore School of Electrical Engineering.

INTENTIONAL BLANK

SOC NANOBASED INTEGRATED WIRELESS SENSOR SYSTEM

Penghua Sun, Maher Rizkalla, and Mohamed El-Sharkawy

Purdue School of Engineering and Technology, Indianapolis, Indiana, USA

ABSTRACT

Smart nanotechnology materials have been recently utilized in sensing applications. Carbon nanotube (CNT) based SoC sensor systems have potential applications in various fields, including medical, energy, consumer electronics, computers, and HVAC (heating, ventilation, and air conditioning) among others. In this study, a nanotechnology multisensory system was designed and simulated using Labview Software. The mathematical models were developed for sensing three physical quantities: temperature, gas, and pressure. Four CNT groups on a chip (two for gas sensor, one for temperature, and a fourth one for pressure) were utilized in order to perform sensing multiple parameters. The proposed fabrication processes and the materials used were chosen to avoid the interference of these parameters on each other when detecting one of them. The simulation results were translated into analog voltage from Labview software, transmitted via Bluetooth network, and received on desktop computers within the vicinity of the sensor system. The mathematical models and simulation results showed as high as 95% accuracy in measuring temperature, and the 5% error was caused from the interference of the surrounding gas. Within 7% change in pressure was impacted by both temperature and gas interference.

1. INTRODUCTION

1.1 Nanosensors

Nanotechnology has recently explored unique features related to smart nanomaterials such as fullerene, carbon nanotube, graphene, quantum dots, nanophotonic crystals, magneto resistance material, and nano polymers [1]. Figure 1 shows the image of the Fullerene, quantum dots and CNTs, that can sense physical quantities such as temperature, pressure, gas, and magnetic fields by changing their electrical or magnetic properties. They can change their conductivity when exposed to a gas. The capacitance of the device will also change when the CNTs are exposed to these physical quantities [2-3]. There are mainly three types of CNTs: Single-walled (SWCNT), double-walled (DWCNT) and multi-walled (MWCNT) as shown in Figure 1. Single-walled CNTs can be seen as a closed-loop of graphene (diameter of about 1 nanometer), while double-walled and multi-walled CNTs are basically several single-walled ones placed in a coaxial configuration (diameter of about 10s of nanometers) [4]. Reference [5] discussed the design of a radio frequency (RF) based system that detected the change in the reflected RF power due to the change in the device capacitance, and accordingly, to the exposed gas. The fabrication of the CNTs may be demonstrated by separation and enrichment [6].

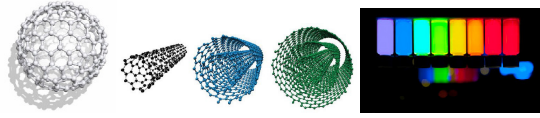


Figure 1: Image of Fullerene, Quantum Dots and CNTs.

Researchers have succeeded to implement nanotechnology sensors that sense one parameter such as temperature, gas, or pressure. There is a need, however, to have multiple sensors on a chip that can detect a combination of parameters simultaneously to perform multiple bio functions. Since the CNT material changes with all parameters when exist simultaneously, it is hard to detect the impact of each parameter on the CNT based sensor. The exposure of the CNT device to both temperature and gas present in the same chamber are conducted. The resulting change in the resistance cannot precisely predict the change of each, separately. Furthermore, the change in the pressure in the presence of the temperature and gas may not reflect the true value of the pressure quantity. A reliable system should resolve the interference in order to separately detect the impact of every parameter separately. Therefore, building three sensors in the same chip, each to detect one quantity separately is quite challenging since the three sensors are set into the same chip and exposed to all parameters' changes. Other issues may include the type of assembly of the CNT materials that may be chosen to provide the range of temperature, gas, or pressure to be suitable for Bio-applications.

1.2 Processing Unit

An interface unit may be necessary to provide the power requirement to drive the processing unit. A signal conditioning unit may be needed to provide the current levels suitable for driving the output stages. A nanoelectromechanical system (NEMS) may be implemented on board with the processing unit to provide data analysis and wireless transmission. The processing unit may take advantage of the available hardware board system associated with the Labview software. This research approach accommodated each sensor separately to sense one quantity. This was accomplished by shielding it from the interference of other quantities. The small percentage error can be incorporated in the mathematical model of the processing unit.

2. SYSTEM DIAGRAM

The temperature and pressure sensor system may be utilized in bioengineering applications, including electrocardiography (ECG), human temperature, and heart rate, among others. Figure 2 shows a general medical setting for the use of multiple sensor system via wireless network.

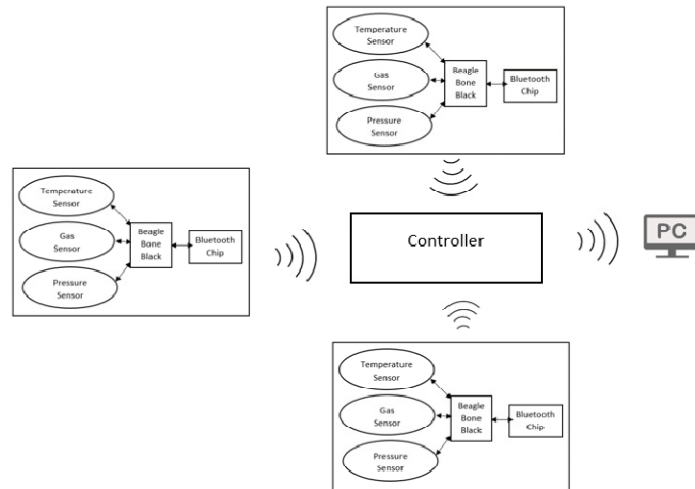


Figure 2: System Diagram.

3. SENSOR MODELS

3.1 The Pressure Sensor

The pressure sensor designed for this system was based on detecting the change in conductivity caused by pressure, resulting in a change of the material resistance. The substrate material selected for this purpose was chosen based on its linearity and pressure coefficients. PMMA (polymethyl methacrylate is a versatile polymeric material that is well suited for many imaging and non-imaging microelectronic applications. It is a common positive resist for e-beam, x-ray, and deep UV microlithographic processes) substrate material was found to be appropriate for this design [7]. Figure 3 gives the shape of the device. The detection of a clamped circular shape under a uniform pressure P is denoted by w and is given by [8]:

$$w = \frac{p a^2}{264D} [1 - (r/a)^2]^2 \quad (1)$$

Where r and a are the radial coordinate and diaphragm radius, respectively. It is clear from Figure 3 that r is the original radius and a is the radius change. D is a measurement of stiffness and is given by:

$$D = \frac{E h^3}{12(1-\nu^2)} \quad (2)$$

Where E , h , and ν are Young's modulus, plate thickness, and Poissons ratio, respectively.

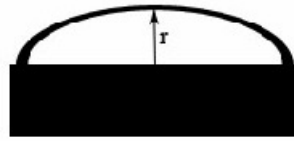


Figure 3: The Shape of PMMA Device.

The room temperature resistance at 0 kPa of a sensor was typically ranged from several kΩ to several hundred kΩ. Some works have been done previously [9] showed that the resistance across the CNTs increases linearly with applied pressure up to 70 kPa. Figure 4 gives the pressure-detection curve, showing the effect of the diameter size of the substrate

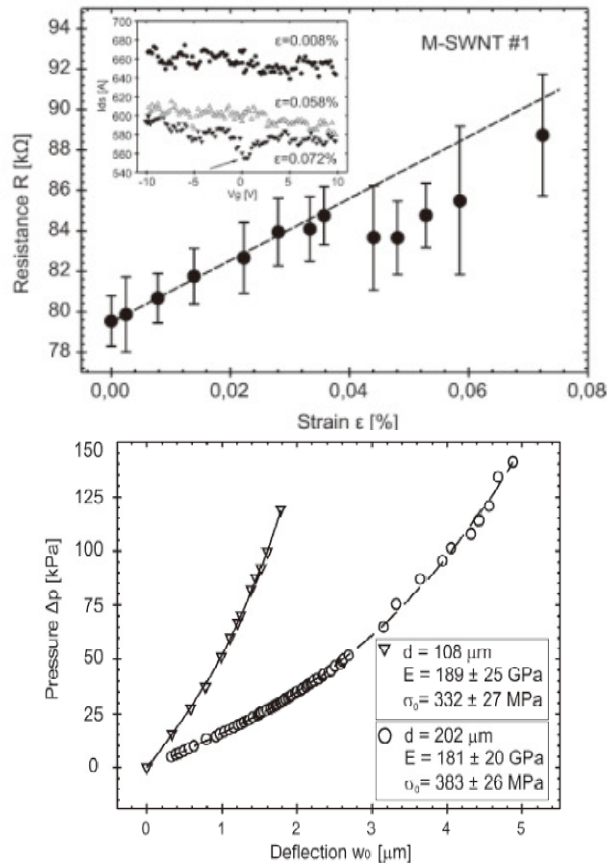


Figure 4: The Pressure/Deflection and Strain Curves.

3.2 Resistance Model

The gauge factor of the CNT is given by [2]:

$$G = \frac{\Delta R}{R} \left(\frac{1}{\epsilon} \right) \tag{3}$$

Where R and ΔR are the initial resistance of the sensor before the pressure is applied, and the resistance change of the CNT under the pressure, respectively, and ε is the strain of the sensor material.

The resistance as function of the strain parameter is given by [3,4]:

$$R(\varepsilon) = R_0 + \tilde{R}_1 \varepsilon \quad (4)$$

In this expression, R_0 is the resistance without strain, \tilde{R}_1 is the strain coefficient. The model with the strain coefficient is given by [5]:

$$R(\varepsilon) = R_s + \frac{1}{|t|^2} \frac{h}{8e^2} \left[1 + \exp\left(\frac{\tilde{E}_g \varepsilon}{k_B T}\right) \right] \approx \left(R_s + \frac{1}{|t|^2} \frac{h}{4e^2} \right) + \left(\frac{\tilde{E}_g}{|t|^2} \frac{h}{8e^2 k_B T} \right) \varepsilon \quad (5)$$

Where R_s is the series resistance of the junction due to SWCNT metal contacts, $|t|^2$ is the transmission through the nanotube, and $E_g = \tilde{E}_g \varepsilon$ is the strain-dependent band gap for metallic nanotubes, neglecting the torsion contributions. Figure 4 shows a typical resistance-stain curve, showing the effect of the diameter size in the pressure range, leading to the proper deflection that may be detected by the processing unit.

The sensitivity of CNT to detect pressure changes was found to be 54pA/mbar for three different strain coefficients averages [6]. Figure 5 shows the sensitivity curve considered for the design of the pressure sensor. Heartbeat pressure is about 170mbar, and the current change is about 10nA current change ($V_{ds}=200\text{mV}$, $V_g=4\text{V}$), which make the design of the sensor system appropriate for bio-engineering applications. This shows the appropriate sensitivity of the selected material.

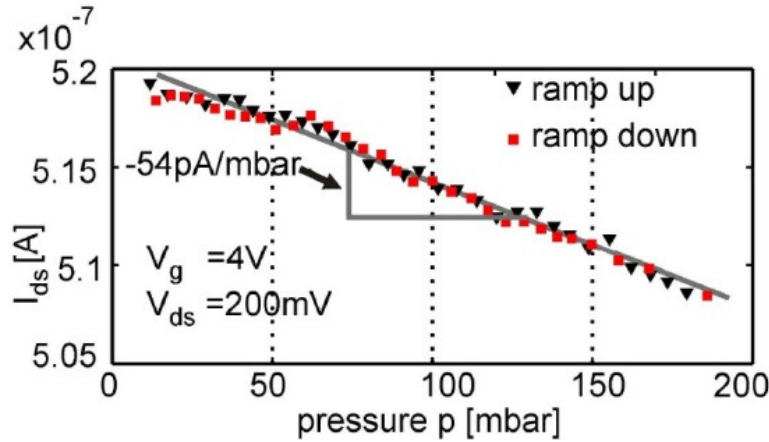


Figure 5: The Sensitivity Curve of Pressure Sensor.

As it can be seen from the data information of this sensor, the sensor is appropriate to measure up to 200 mbar linear range. The minimum current observed was near 200 mbar at $0.5\mu\text{A}$, which is still appropriate to be processed by the hardware processing unit.

3.3 The Temperature Sensor Model

3.3.1 The Concept

The concept of the temperature sensor is based on exhibiting the change in the CNT resistance material when exposed to change in temperature. The temperature change may also alter the pressure, and the thermal expansion (or contraction) of the Polyamide material.

3.3.2 The Device Model

When the temperature changes, the polyamide alters its volume, and accordingly, the pressure from it to the CNTs sensing material. Then the CNTs will exhibit a change in resistance that can be detected, and the change will be processed by the DSP unit, producing the pressure data in the form of electrical energy. The model of the pressure sensing element of the temperature sensor is the same as the pressure sensor model. Thermal stress pressure can be written as $\Delta P = E\alpha\Delta T$, Where E is young's modulus which is near 1.1 GPa for polyamide, and α is the thermal expansion coefficient and it is about 110×10^{-6} m/mK for the same material.

3.4 The Gas Sensor Model

3.4.1 The CNT Schottky Model

A Schottky barrier diodes (or MS diodes) are characterized by the so-called Schottky barrier height (SB), and denoted by the metal semiconductor potential barrier Φ_B . In our model, the CNTs were semi conductive SWCNT and the base is metal. When the CNT exposed to the gas, such as NO_2 , the S_B height increases, following the model given in [4], and [5], and given below for a P-channel device:

$$\Phi_B = \left(\Phi_{CNT} + \frac{E_g}{2} \right) - \Phi_M \quad (5)$$

Where $\Phi_{CNT} = 4.7eV$. E_g and Φ_M are the band gap and metal work function respectively. For the nanoscale devices such as in CNTs, the potential barrier as determined by [6] is modeled as:

$$\Phi_{SB} \approx \frac{kT}{\beta} \ln \left(\frac{\alpha \sqrt{\frac{E_g}{2kT}}}{\ln \alpha \sqrt{\frac{E_g}{2kT}} \frac{\Phi_B}{kT}} \right) \quad (6)$$

In the above expression, $\alpha = \frac{e^2 \left(\frac{2}{\pi}\right)^{\frac{3}{2}}}{3\sqrt{\beta} \alpha \gamma d C}$ ($\beta = 0.7$, $\alpha = 0.142$ nm C-C bond length, d = CNT diameter, $\gamma = 2.5eV$ is tight-binding overlap integral, e = electron charge, k = Boltzmann constant, $T = 300K$, and C = capacitance per unit area between metal-CNT). Assuming that $\Phi_{SB.i}$ and final $\Phi_{SB.f}$ are the initial and final value of the SB height before and after the gas adsorption, respectively, with the assumption that the band gap of the CNT does not change after gas, the $\exp \left[\frac{\Delta \Phi_{SB}}{kT} \right]$ term can be written as:

$$\exp \left[\frac{\Delta \Phi_{SB}}{kT} \right] = \exp \left[\frac{\Phi_{SB.f} - \Phi_{SB.i}}{kt} \right] \quad (7)$$

Combining equations 6 and 7, we get,

$$\begin{aligned} \exp\left[\frac{\Delta\Phi_{SB}}{kT}\right] &= \exp\left[\frac{\Phi_{SB.f}-\Phi_{SB.i}}{kt}\right] = \exp\left[\frac{1}{\beta}\ln\left(\frac{\alpha\sqrt{\frac{E_g}{2kT}}}{\ln\alpha\sqrt{\frac{E_g}{2kT}}\frac{\Phi_{B.f}}{kT}}\right) - \frac{1}{\beta}\ln\left(\frac{\alpha\sqrt{\frac{E_g}{2kT}}}{\ln\alpha\sqrt{\frac{E_g}{2kT}}\frac{\Phi_{B.i}}{kT}}\right)\right] = \\ \exp\left[\frac{1}{\beta}\ln\left(\frac{\ln\alpha\sqrt{\frac{E_g}{2kT}}\frac{\Phi_{B.i}}{kT}}{\ln\alpha\sqrt{\frac{E_g}{2kT}}\frac{\Phi_{B.f}}{kT}}\right)\right] &= \left[\frac{\ln\alpha\sqrt{\frac{E_g}{2kT}}\frac{\Phi_{B.i}}{kT}}{\ln\alpha\sqrt{\frac{E_g}{2kT}}\frac{\Phi_{B.f}}{kT}}\right]^{\frac{1}{\beta}} = \left[\frac{1}{1-\frac{\Delta\Phi_B}{kT(\ln\alpha\sqrt{\frac{E_g}{2kT}}\frac{\Phi_{B.i}}{kT})}}\right]^{\frac{1}{\beta}} \quad (8) \end{aligned}$$

It is clear from Equation 7, that the SB height change is related to the change in Φ_B . With a reasonable approximation, Equation 8 can be expressed as: $\exp\left[\frac{\Delta\Phi_{SB}}{kT}\right] = (1 - b\theta_M)^{-\frac{1}{\beta}}$, Where b is a constant dependent $\propto \frac{\Delta\Phi_B}{kT(\ln\alpha\sqrt{\frac{E_g}{2kT}}\frac{\Phi_{B.i}}{kT})}$, and θ_M is the metal surface coverage given below.

$\theta_M = (1 - e^{-k_M P t})$, $\theta_{NT} = (1 - e^{-k_{NT} P t})$, $\theta_M = \frac{K_M P}{K_M P + 1}$, $\theta_{NT} = \frac{K_{NT} P}{K_{NT} P + 1}$. The surface coverage, θ_M and θ_{NT} are obtained from the Langmuir model [6][7]. The polarizability and dipole moment of the gas molecule are determined whether b is positive or negative, which corresponds to either a decrease or increase in conductance. The resistance change includes two components: the contact resistance (R_M) and the CNT channel resistance (R_{CNT}). Assuming that the initial and final resistances are $R_0 = (R_M + R_{VNT})$ and $R_f = R_{0=M.f} + R_{CNT.f}$. Then the CNT resistance is inversely proportional with the carrier density ($R_{CNT} \propto \frac{1}{n_0}$) [5, 9].

$$\begin{aligned} \Delta R = R_f - R_0 &= R_M \left(\frac{R_{M.f}}{R_M} - 1\right) + R_{CNT} \left(\frac{R_{CNT.f}}{R_{CNT}} - 1\right) = R_M \left[\frac{n_0}{n_f} e^{\frac{\Delta\Phi_{SB}}{k_B T}} - 1\right] + R_{CNT} \left[\frac{n_0}{n_f} - 1\right] \\ &= R_M \left[(\delta\theta_{NT} + 1)^{-1} (1 - b\theta_M)^{-\frac{1}{\beta}} - 1\right] + R_{CNT} \frac{-\delta\theta_{NT}}{\delta\theta_{NT+1}} \quad (9) \end{aligned}$$

$$\frac{\Delta G}{G_0} = \frac{G_f - G_0}{G_0} = \frac{G_f}{G_0} - 1 = \frac{R_0}{R_f} - 1 = \frac{R_0 - R_f}{R_f} = \frac{-\Delta R}{R_0 + \Delta R} = \frac{\left[1 - (\delta\theta_{NT} + 1)^{-1} (1 - b\theta_M)^{-\frac{1}{\beta}}\right] R_M + R_{CNT} \frac{\delta\theta_{NT}}{\delta\theta_{NT+1}}}{(\delta\theta_{NT} + 1)^{-1} (1 - b\theta_M)^{-\frac{1}{\beta}} R_M + R_{CNT} \frac{1}{\delta\theta_{NT+1}}} \quad (10)$$

G is the conductance.

4. WIRELESS TRANSMISSION

In this section, the complete wireless transmission based on BBB (developed by N.I.) and the Bluetooth techniques were introduced.

4.1 Hardware Board System

The assembly of the hardware system consists of three main components: BBB board, Bluetooth chip, and the control unit. The micro USB port of BBB simply works as a power source port; the micro HDMI port was linked to a monitor; the micro SD card was attached to the reader, which worked as a hard disk to the system. The standard USB port was interfaced to the USB hub and then attached to the USB based Bluetooth chip, and the mouse and keyboard adapter. Figure 6

shows the BBB board used in this work. In this application, the BBB board functions as a mini PC that communicates with the Bluetooth system. It receives the data from the Labview software and transmits it to another Bluetooth device.

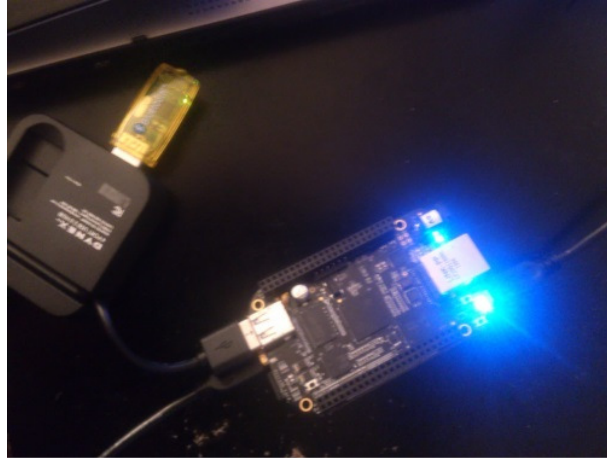


Figure 6: Connection Diagram of BBB Board.

5. RESULTS & DISCUSSIONS

In this section, we report simulation results from Labview for the integrated sensor system, including the temperature, gas, and pressure sensors. The wireless transmission for the dynamic responses of the sensors was verified. The simulation here presents data for one sensor as sample data for the multisensory system.

5.1. The Temperature Sensor

The range of the temperature used for testing the sensor was from 283.15 to 353.15 K, which corresponds to 10 and 80 °C, respectively. This range was chosen in order to perform linear response with the CNT resistance. Figure 7 shows the details of the response following the mathematical model. As it can be seen, the curve fits linearly well within the temperature range of 283 to 353K, and beyond this range, the response is not linear. This is the useful range of Bio and HVAC applications. The x axles in Figure 7 are time (s). The resistance is changing from 75 to 700 k Ω and the sensitivity range of the CNT pressure sensing part is 0~8400 kPa. Figure 7 describes the resistance-temperature direct relationship. The slope of the curve (within the desired range of temperature) gives the sensitivity determined by $\Delta R/\Delta T$. In this case, the sensitivity is 0.7 k Ω /K. This indicates that the sensor can detect as small as milli K change, since the corresponding change in resistance is still in the 10s k Ω range, and this can be easily detected by the processing unit of the system.

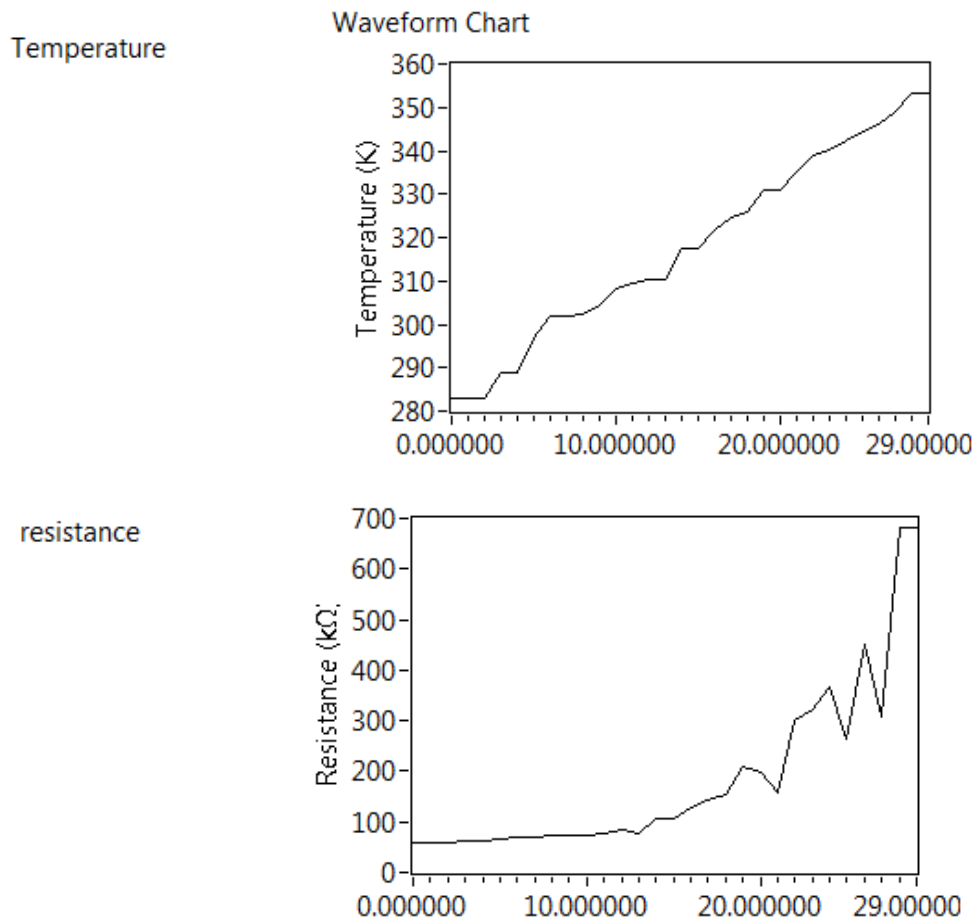


Figure 7: The Simulation Result of the Temperature Sensor.

5.2. Discussions

The design of the integrated sensor system built on the same chip, detecting temperature, gases, and pressure that interfere with each other is demonstrated. The nanodevices utilized CNT smart nanomaterials that are appropriate for bio-engineering applications as evidenced by the power level have been determined. The range of detection is determined from the linear part of the sensor characteristics. The error factor from interference was compensated within the DSP unit. The proposed system can be expanded to include more sensors such as liquid and gas flow or blood sugar detectors. For the temperature sensor, an indirect detection method was introduced. The CNTs worked as the pressure sensor to detect the volume change of the thermal expansion elements to avoid the gas influence. In this case, a read-out circuit was to be interfaced to the sensor in order to process the data.

For the gas sensor, the CNTs coating with SS-DNA device was given. The two sequences of SS-DNA made the CNTs significantly more sensitive to the gas. To raise the accuracy of the sensor, the twin-sensor with both sequence 1 and sequence 2 SS-DNA coated CNTs was used. The DSP chip could be modified to incorporate both sequence 1 and sequence 2 $\% \Delta I/I$ data, into the table look up. For the pressure sensor, the CNTs placed between source and drain of the CMOS device

was introduced. To reducing the effect of the temperature and gas, the CNTs were coated with Parylene C material.

REFERENCES

- [1] Christofer Hierold, CONCEPTS FOR CARBON NANOTUBE SENSORS, Micro and Nanosystems, ETH Zurich.
- [2] Verma, R. ; Dept. of Electr. & Comput. Eng., Indiana Univ.-Purdue Univ. Indianapolis (IUPUI), Indianapolis, IN, USA ; Said, K. ; Salim, J. ; Kimathi, E. Carbon nanotube-based microstrip antenna gas sensor, Circuits and Systems (MWSCAS), 2013 IEEE 56th International Midwest Symposium on, 4-7 Aug. 2013.
- [3] M. Fujioka, H. Watanabe, Y. Martin, M. Nakano, Separation and enrichment of semiconducting carbon nanotubes and its application to highly sensitive carbon nanotube gas sensor, 2011 IEEE Nanotechnology Materials and Devices Conference.
- [4] Cheng Yung Kuo, Chia Lang Chan, Chie Gau, Chien-Wei Liu, Shiuan Hua Shiau, and Jyh-Hua Ting, Nano Temperature Sensor Using Selective Lateral Growth of Carbon Nanotube Between Electrodes, IEEE TRANSACTIONS ON NANOTECHNOLOGY, VOL. 6, NO. 1, JANUARY 2007.
- [5] Cosmin Roman, Florin Ciontu, Bernard Courtois, Single molecule detection and macromolecular weighting using an all-carbon-nanotube nanoelectromechanical sensor, 2004 4th IEEE Conference on Nanotechnology.
- [6] Carmen K. M. Fung, Maggie Q. H. Zhang, Rosa H. M Chan and Wen J. Li, A PMMA-BASED MICROPRESSURE SENSOR CHIP USING CARBON NANOTUBE SENSING ELEMENTS, Centre for Micro and Nano Systems, The Chinese University of Hong Kong, Hong Kong SAR.
- [7] Brendan Crawford, Dan Esposito, David Pelletier, Vishal Jain, Flexible Carbon Nanotube Based Temperature Sensor for Ultra-Small-Site Applications.
- [8] Suehiro, G. Zhou and M. Hara, Fabrication of a carbon nanotube-based gas sensor using dielectrophoresis and its application for ammonia gas sensor using dielectrophoresis and its application for ammonia detection by impedance spectroscopy, Journal of Physics D, Vol. 36, pp. 109-114 (2003).
- [9] Sun Penghua, "MS Thesis" Department of ECE, Purdue School of Engineering and Technology, May 2014.

ANALYSIS OF SPECTRUM SENSING TECHNIQUES FOR DETECTION OF DVB- T SIGNALS IN GAUSSIAN AND FADING CHANNELS

Ireyuwa Igbinosa¹ Olutayo Oyerinde² and Stanley Mneney¹

School of Electrical, Electronic and Computer Engineering,¹
University of KwaZulu-Natal, Durban, 4041, South Africa.
Tel: +27 74 4200748

School of Electrical and Information Engineering,²
University of the Witwatersrand, Johannesburg, 2050, South Africa.
yvwa123@gmail.com¹; olutayo.oyerinde@wits.ac.za²;
mneney@ukzn.ac.za¹

ABSTRACT

Spectrum sensing is an essential concept in cognitive radio. It exploits the inefficient utilization of radio frequency spectrum without causing destructive interference to the licensed users. In this paper we considered spectrum sensing of Digital Video Broadcast Terrestrial (DVB-T) signal in different scenario. We compared various spectrum sensing algorithms that make use of the second order statistics; the energy detector was also included for comparison. The results show that it is possible to obtain good detection performance by exploiting the correlation method.

KEYWORDS

Spectrum sensing, DVB-T, OFDM, Cognitive Radio

1. INTRODUCTION

Spectrum sensing can be said to be the process of performing measurement on a part of a spectrum and thereby forming a decision based upon measured data [1]. Spectrum sensing is an essential operational block of the cognitive radio which consists of spectrum sensing, management and spectrum mobility. Measurement of spectrum has shown unused spectrum resources in frequency, time and space [2]. The frequency bands of the wireless communication are currently not efficiently used. This is due to the strict frequency allocation policy. The issue of spectrum utilization brought about the cognitive radio concept. This concept has proven itself as a promising technique to improve spectrum utilization by exploiting the spectrum holes. However, the introduction of cognitive Radio in an existing network increases interference. Hence the impact on the primary network must be kept as low as possible. Therefore the

secondary users must sense the spectrum and detect whether the primary user is occupied. For this to be achieved, the secondary users should be able to detect very weak primary user signals [4].

A notice of proposed the rulemaking (NPRM) issued by the U.S. federal communication commission (FCC) in 2004 indicated that the unutilized TV Channels in both very high frequency (VHF) and ultra high frequency (UHF) bands could be used for fixed broadband access [5]. This factor increased the interest within the research community to develop a standard for wireless regional area network (WRAN) system operating on a TV white space making use of cognitive Radio techniques [6]. According to the IEEE 802.22 standard a secondary user should be able to detect a primary user of DVB-T signal with the probability of detection of at least 0.9 and the probability of false alarm of not more than 0.1, at -22.2dB SNR [7-8]. The cognitive functionality of this standard is channel sensing, channel classification and maintenance of channel information. A description of all the functionalities of an 802.22 WRAN is given in [7].

In this research, we evaluated various detection algorithms for detecting primary DVB-T signals in different scenarios. We considered a number of feature detectors and compared them with energy detectors.

The rest of this paper is organized as follows, section II, show characteristics of DVB-T signals, section III, problem formulation and section IV shows the various detection algorithms, section V shows the simulation results and finally section VI concludes this paper.

2. CHARACTERISTICS OF DVB-T SIGNALS

From the draft of the European Telecommunication Standard Institute (ETSI), the characteristics of the physical layer were discussed in detailed [9]. But in the scope of this work, we are going to summarize our main area of interest which is the channel coding and modulation. The system input stream which is organized in fixed length packets of the digitized multiplexed MPEG-2 signal that carries the payload data is divided in lengths of 188bytes [9]. The major modulation constellation used are QPSK, 16 QAM, 64 QAM, non-uniform 16 QAM, or non-uniform 64 QAM. Details can be found in section 4.3.5 in [9]. The modulated signals are determined by three main parameters which include; bandwidth, mode and length of the cyclic prefix. The bandwidth usually takes one of the following values 5MHz, 6MHz, 7MHz or 8MHz. The number of subcarriers that are used in the OFDM modulation are determined by the mode parameters which can take values of 2k or 8k mode which is 2048 or 8192 respectively. The cyclic prefix is used in order to avoid inter symbol interference (ISI) in wireless transmission. The length of the cyclic prefix is determined in terms of fractions of the duration of OFDM symbol part and assigns the values 1/4, 1/8, 1/16 and 1/32. A sequence of 68 consecutive OFDM symbols constitutes a frame and four of such frames are gathered to form a super frame. The continuous DVB-T signal in time domain can be defined as:

$$s(t) = e^{j2\pi f_c t} \sum_{m=0}^{\infty} \sum_{i=0}^{67} \sum_{k=0}^{k_{max}} c_{m,i,k} \varphi_{m,i,k}(t) \quad (1)$$

3. SYSTEM MODEL

It is assumed that the DVB-T Signal mode is known; hence the parameter of the DVB-T signal that we want to detect is known. Let $x(t)$ be the received continuous baseband signal, then the time received sequence becomes:

$$s[k] = x(kT_e), k = 0, 1, \dots, M - 1,$$

where M is the total number of samples. The difficulty of spectrum sensing is to decide whether there is transmitted signal or not. Hence we have to be able to discriminate between the following hypotheses:

The H_0 and H_1 correspond to either the absence or presence of DVB-T signal. The sequence $s[k]$ is the sampled version of the signal $s(t)$ defined in equation (1). Where $f_c = 0$ and sampling rate is $1/T_e$. The amplification factor h_i and the delays k_i describe the multipath fading environment. The model assumes that the signal propagates through multipath fading environment. The noise $n(k)$ is assumed to be complex white zero-mean Gaussian with variance σ_n^2 . The decision on the presence or absence of DVB-T signal is based on a test statistics γ which is a result of the received sequence $\{x[k]\}_{k=0}^{M-1}$, i. e

$$\gamma = f(x[0], x[1], \dots, \dots, \dots, x[M - 1]) \quad (2)$$

In order to decide upon a hypothesis we compare γ to a threshold (μ) and decide whether the signal is present. The probability of false alarm (Pfa) and the probability of detection are defined as:-

$$Pfa = Pr\{\gamma > \mu | H_0 \text{ is true}\}$$

and $Pd = Pr\{\gamma > \mu | H_1 \text{ is true}\}$

These are the two main detector performance indicators. The test statistic can be chosen in different ways in the rest of the paper we would show some of the choices.

4. SPECTRUM SENSING ALGORITHMS

This section presents the various algorithms that were considered for the detection of the DVB-T signals.

4.1. Energy Detection

Energy detection is an optimal way to detect primary signals when prior information of the primary signal is unknown to the secondary user. It measures the energy of the received signal waveform over a specified observation time [10-11] and compares it to a predefined threshold. The test statistic for energy detector is:

$$Y_{ED} \triangleq \sum_{k=0}^{M-1} x[k]x^*[k] \quad (3)$$

The performance of the energy detector is well known and can be written in closed form. The probability of false alarm P_{FA} is then given as

$$P_{FA} \triangleq P_r (Y_{ED} > \mu_{ED} | H_0) = 1 - F_{x_{2M}^2} \left(\frac{2\mu_{ED}}{\sigma_n^2} \right) \quad (4)$$

Where $F_{x_{2M}^2}(\cdot)$ is the cumulative distribution function of a x^2 - distributed random variable with $2M$ degrees of freedom. Therefore, given a false alarm. Then we can derive the threshold μ_{ED} from the following equation :-

$$\mu_{ED} = F_{x_{2M}^2}^{-1} (1 - P_{FA}) \frac{\sigma_n^2}{2} \quad (5)$$

The probability of detection is then defined as:-

$$P_r (Y_{ED} > \mu_{ED} | H_1) = 1 - F_{x_{2M}^2} \left(\frac{2\mu_{ED}}{\sigma_n^2 + \sigma_s^2} \right), \quad (6)$$

where σ_s^2 the average is received signal power. From equation 5 it shows that the energy detector requires the noise power σ_n^2 to be known, else the detector would perform poorly. However, it is a known fact that the energy detector deteriorates when noise power estimate is imperfect [12].

5. FEATURE DETECTORS BASED ON SECOND ORDER STATISTICS

The term feature detector is mainly used in the context of spectrum sensing and usually refers to the exploitation of known statistical properties of the signal [13]. The features of the transmit signals are results of the redundancy added by coding of the modulation and formatting schemes used at the transmitter end. For example OFDM modulation adds a cyclic prefix that manifest itself through a linear relationship between the transmitted samples. It is also known that most communication system multiplex known pilot symbol into transmitted data stream or better still superimpose pilot symbol on the transmitted signals and by so doing it results into very destructive signal feature.

In this section we would show the simulation results of some state of the art detectors that exploit signal features suitable for spectrum sensing application in cognitive radio. The methods proposed here are recent advances in spectrum sensing and there are still ongoing researches in this area.

A. Autocorrelation Based Detection

The autocorrelation detector is a simple and computationally efficient spectrum sensing scheme for OFDM based primary user signal using its autocorrelation coefficient. OFDM signals have a very explicit correlation structure imposed by the insertion of a cyclic prefix (CP) at the transmitter [14]. This method exploits the correlation of an OFDM signal using knowledge of the

Td. The method proposed in [14], uses the empirical mean of the sample value product $r[k]$, normalized by the received power as test statistic. The theoretical approximation of the values of the threshold and probability of detection are given in equation 20-21 [14]. The noise variance does not need to be known to make a decision. The test statistics is defined in [14] is given as-

$$Y_{ACD} = \frac{\frac{1}{M - T_d} \sum_{k=0}^{M-T_d-1} Re(r[k])}{\frac{1}{M \sum_{k=0}^{M-1} |x[k]|^2}} \quad (7)$$

B. Detection Based on Cyclic Prefix Sliding Window Correlation

To protect the DVB-T signal against ISI, a CP which copies the last part of the symbol is added at the beginning of each OFDM symbol. The duration of the CP could be 1/4, 1/8, 1/16 or 1/32 of the original symbol duration. The detector of [15] uses a sliding window that sums T_c which is the consecutive samples. The approach of taking the sum of the real part was proposed in [16] the test statistics is given as:-

$$Y_{SW} = \max_{\theta} Re \left(\sum_{k=0}^{\theta+T_c-1} R[k] \right) \quad (8)$$

C. Detection based on second order statistics

The repetition of data in the CP causes a structure to an OFDM signal. This structure can be exploited for detection. The detector proposed in [13] is assumed to know the size of the OFDM symbol that is T_c and T_d and exploits the structure obtained by the CP, where S_{θ} indicates the T_c consecutive correlated samples. Detailed equation is given in [13]. Contrary to the Energy detector, this does not require any prior knowledge of the noise power. The test statistic for this proposed detector is given as:-

$$Y_{2nd} = \max_{\theta} \left(\sum_{k \in S_{\theta}} Re(R[k]) \right)^2 \quad (9)$$

6. SIMULATIONS AND RESULTS

In this section, we present the simulated results of the spectrum sensing algorithms described in previous sections showing their comparisons. Five scenarios were considered, the DVB-T signal model is the same for all scenario. The following setting was used for all scenarios:

- DVB-T signal model: 2k
- Cyclic prefix: $T_c/T_d=1/4$
- Probability of false Alarm: 0.05

- DVB-T bandwidth: 8 MHz
- Number of realization for simulated Pd: 1000

Abbreviation	Detector
2 nd Order	Detection based on second order statistics
ACD	Autocorrelation based detector
SW	Detection based on cyclic prefix sliding window
ED	Energy Based detector

Table 1. Detectors Simulated

Scenario 1

In this scenario AWGN channel was considered with the sensing time of 10ms and Pfa of 0.05 using 1000 simulations. The results are shown in figure 1

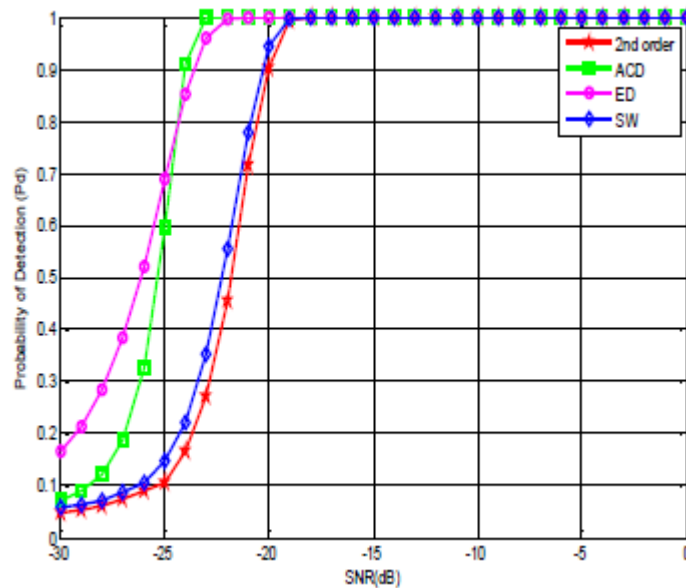


Fig.1. Probability of detection for AWGN channel sensing time 10ms

Scenario 2

In this scenario the AWGN channel was considered using the sensing of 50ms and Pfa of 0.05 using 1000 simulation

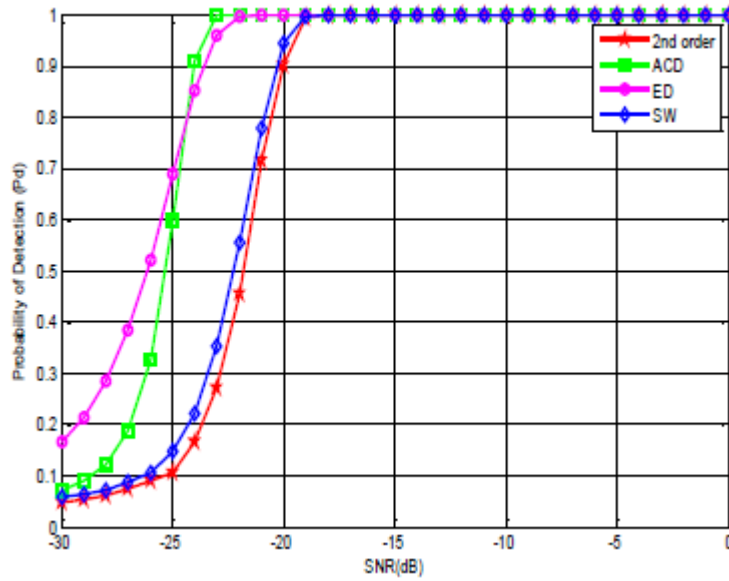


Fig.2. Probability of detection for AWGN channel sensing time 50ms

Scenario 3

In this case a Rayleigh flat fading channel was considered using the built in Matlab fading generators with a sensing time of 10ms and Pfa of 0.05 using 1000 simulation result are shown in figure 3

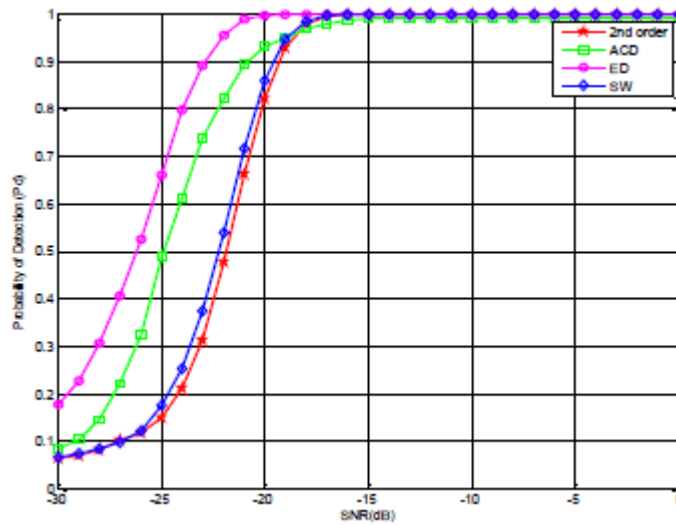


Fig.3. Probability of detection for Rayleigh flat fading 10ms sensing time

Scenario 4

In this case a Rayleigh flat fading channel was considered using the built in Matlab fading generators with a sensing time of 50ms Pfa of 0.05 using 1000 simulation result are shown in figure 4.

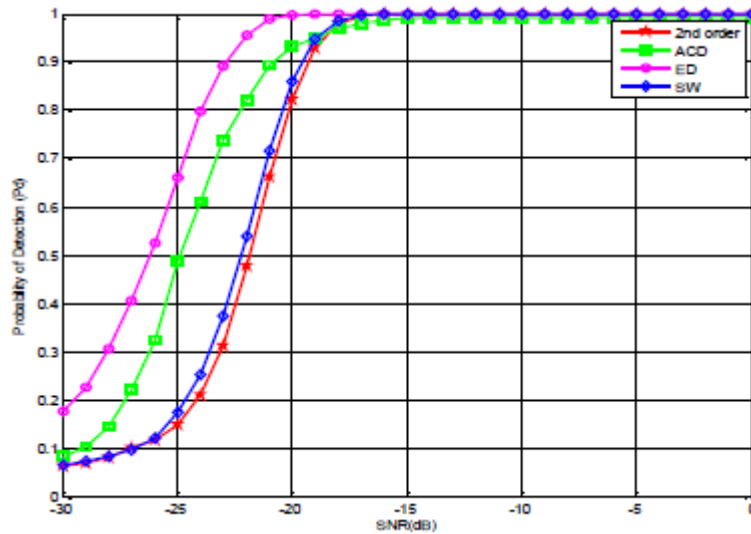


Fig.4. Probability of detection for Rayleigh flat fading 50ms sensing time

Scenario 5

In this section a Rayleigh flat fading channel with shadowing was considered with 10ms sensing time. The standard deviation of the log – normal shadowing is 10dB. The results are shown in figure 5

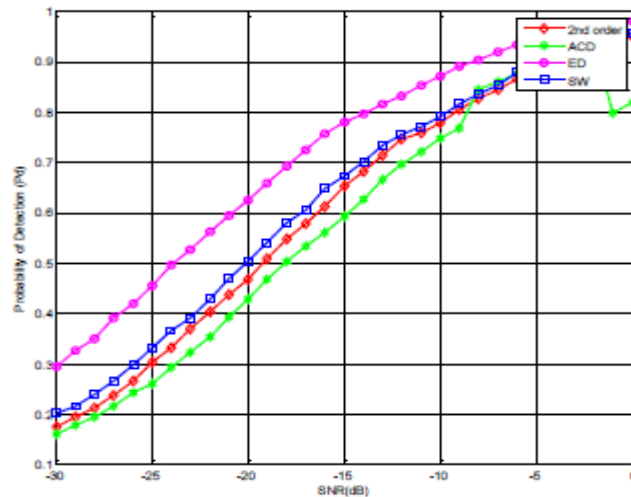


Fig.5. Probability of detection for Rayleigh flat fading channel with shadowing 10ms sensing time

Scenario 6

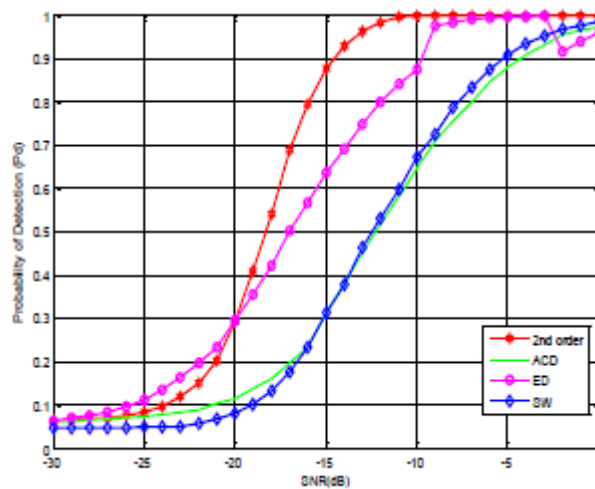


Fig.6. Probability of detection for Rayleigh multipath fading with 10ms sensing time

In this scenario the Rayleigh multipath fading channel was considered using the extended vehicular model using the sensing 10ms. The simulated channel model was done by the built in functions in MATLAB for Rayleigh fading environment.

The presented algorithms do not rely on any information about the structure of the waveform; the only assumption is the knowledge of the CP and the total duration of OFDM symbol of the feature detectors while the energy detector does not require prior knowledge of the signal. Though the energy detector tends to show better performance among the others, its major drawback is that it assumes perfect knowledge of the noise floor level. Therefore, if the noise power is known the energy detector works excellently well. The performance of the energy detector is drastically affected if the noise power knowledge is erroneous. In practical scenario the noise power is never perfectly known hence estimated. Also if there is interference i.e. a signal from another secondary user, the feature detectors would be able to distinguish between the primary signal and the interfering signal whereas the energy detector would not be able to make the detection. In the case where the knowledge of the noise power is completely unknown only the second order statistics and the autocorrelation detectors would be able to work effectively. From observation of the graphs it can be seen that the second order statistics outperforms the autocorrelation based detector in all scenarios in the region of $P_d \geq 0.15$.

When we observe the result of the Rayleigh multipath fading channel, we can see that the channel does not much influence on the performance of the presented detectors because a frequency selective channel, exploits the inherent correlation of the OFDM signal obtained by repetition of data caused by the cyclic prefix is then exploited for detection.

7. CONCLUSION

In this paper we considered various sensing algorithms which use the second order statistics for the detection of the DVB-T signal and compared the result to the energy detector. The second order based statistics method performed well with less information and they would make good

candidates for implementation in cognitive Radio spectrum sensor. In this work we did not consider the carrier frequency offset (CFO) because such imperfect knowledge would decrease the correlation structure of the OFDM signal. In future we would make comparison of spectrum sensing algorithms under such conditions.

ACKNOWLEDGEMENT

The Author would like to appreciate the centre for postgraduate funding (CEPS) for their funding and support towards this work.

REFERENCES

- [1] End to End Efficiency (E3) white paper [http:// www.ict-e3.eu](http://www.ict-e3.eu), 2004.
- [2] FCC, "Spectrum policy task force report", Technical report 02-135, Federal communication commission, Nov.2005, available http://transition.fcc.gov/sptf/files/SEWGFfinalReport_1.pdf.
- [3] S.Haykin, "Cognitive radio: brain-empowered wireless communication," IEEE Journal on selected areas in communication, vol.23, no.2, pp.201-220, Feb. 2005.
- [4] A. Sahai, N.Hoven and R.Tandra, "Some fundamental limits on cognitive Radio", in Allerton Conference on communication control and computing, Oct.2004, pp.1662-1671.
- [5] U.S. FCC, ET Docket 04-186, "Notice of proposed rulemaking in the matter of unlicensed operation in TV Broadcast Bands", May 2004.
- [6] G.ko,A.A.Franklin, S-J You, J-S Pak, M-Y Song and C-J Kim, "Channel management in IEEE 802.22 WRAN systems", IEEE Communication magazine, Sep. 2010.
- [7] C.Stevenson et al., "IEEE 802.22: The first Cognitive Radio wireless regional area network standard", IEEE Communication magazine, vol.47, no.1, pp.130-138, Jan.2009.
- [8] S.Shellhammer and G.chouinard, "Spectrum Sensing Requirements Summary", IEEE Standard, 8022.22-06/0089r4. June 2006.
- [9] ETSI EN300 744 VI.6.1 (2009-01), 'Digital Video Broadcasting (DVB), framing structure, channel coding and modulation for digital terrestrial television.' Technical Report, ETSI, 2009.
- [10] H. Urick, "Energy detection of Unknown deterministic signals", Proceedings of IEEE, vol.55, no.4, pp.523-531, Apr. 1967.
- [11] M. Hoyhtya, A. Hekkala, M.Katz and A. Mammela, "Spectrum Awareness: Techniques and challenges for active spectrum sensing", Cognitive wireless networks, pp. 353-372, 2007.
- [12] R. Tandra and A.Sahai, "Fundamental limits and detection in low SNR under noise uncertainty", in IEEE international conference on wireless networks communication and mobile computing, June 13-16, 2005, vol.1, pp.464-469.
- [13] E. Axell, "Spectrum Sensing Algorithm Based on Second Order Statistics" Ph.D. Thesis, Division of communication system, Department of Electrical Engineering, linkoping University, 2012.
- [14] S.Chaudhari, V.koivunen and H.V.Poor "Autocorrelation based decentralized sequential detection of OFDM signals in cognitive radio", IEEE Transaction on signals processing, vol.57, no.7, pp.2690-2200, July 2009.
- [15] Huawei Technologies and UESTC "Sensing Scheme for DVB-T" IEEE Standard 802.22-06/0127-1, July 2006.
- [16] D. Danev, "On Signal Detection techniques for DVB-T Standard" Proceeding of the 4th International Symposium on communication control and regional Processing (CIP) 2010.
- [17] H.L.Van Trees. "Detection, Estimation and Modulation Theory: Part 1, John Wiley and sons Inc. 1968.

AUTHORS

Ireyuwa Eghosa Igbinsa received his Diploma in computer engineering from the University of Benin, in 2003, B.Eng. and M.Sc by Research in Manchester Metropolitan University in Manchester United kingdom in 2011 and currently working on a Ph.D. at the University of KwaZulu-Natal, Durban. His research interest include cognitive radio application, wireless system, Spectrum sensing and OFDM systems.

Olutayo Oyeyemi Oyerinde received the B.Sc. (Hons.) and the M.Sc. from Obafemi Awolowo University, Ile-Ife, Nigeria, in 2000 and 2004, respectively, and the Ph.D. degree from the School of Engineering, University of KwaZulu-Natal (UKZN), Durban, South Africa, in 2010. He is currently a Telecommunications lecturer in the School of Electrical and Information Engineering, University of the Witwatersrand, South Africa. His research interests includes multiple antenna systems, orthogonal frequency division multiplexing system and channels estimation, and signal processing techniques.

Stanley Henry Mneney received the B.Sc. (Hons.) Eng. degree from the University of Science and Technology, Kumasi, Ghana, in 1976 and the M.A.Sc. from the University of Toronto, Toronto, Ontario, Canada, in 1979. In a Nuffic funded project by the Netherlands government he embarked on a sandwich Ph.D programme between the Eindhoven University of Technology, Eindhoven, Netherlands and the University of Dares Salaam, Dares Salaam, Tanzania, the latter awarding the degree in 1988. He is presently a Professor of Telecommunication and Signal Processing at the University of KwaZulu-Natal, Durban, South Africa. His research interests include theory and performance of telecommunication systems, low cost rural telecommunications services and networks, channel modelling and digital signal processing applications.

INTENTIONAL BLANK

COMPARISON OF FILTERING AND CLUSTERING TECHNIQUES IN DIAGNOSIS OF INFANTS RETINOPATHY RISK

Niousha Hormozi¹, Seyed Amirhassan Monadjemi² and Gholamali Naderian³

¹Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran

²Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran

³Isfahan University of Medical Sciences, Isfahan, Iran

¹Niousha.hormozi@yahoo.com, ²amonadjemi@yahoo.co.uk,

³naderianus@yahoo.com

ABSTRACT

ROP is an eye disease in premature infants. In infants who are born earlier than normal, retinal vessel growth stops. Early treatment is very crucial in this case as it can end up to blindness if the diagnosis has not done in a short time. The purpose of this research is to design an intelligent automated system for the early detection of disease at an early stage to prevent these babies from dangerous consequences. In this study, we analyzed the images in the Lab color space, and evaluated the efficiency of applying filters named, Canny Laplacian and Sobel. The results indicate relatively higher efficiency and quality of the Laplacian filter in ROP diagnosis.

KEYWORDS

Retinopathy of prematurity, Canny filters, Laplacian filters, Sobel operator, Lab color space, ROP

1. INTRODUCTION

ROP is an eye disease in premature infants which is due to the abnormal growth of retinal blood vessels that leads to scarring in the retina eventually. The main cause of visual impairment and blindness in ROP is retinal detachment which is the subsequence of the scars in turn. Swollen and twisted veins contribute to Plus-disease (PD) that is considered as the most severe type of ROP[4]. In Figure 1, retinal detachment of scleral membrane due to the stretching of the wound contraction caused by abnormal blood vessel growth is shown.

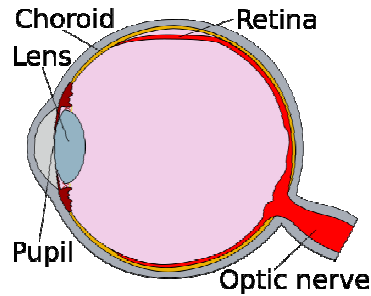


Figure 1. Retinal detachment caused by abnormal vessel growth [1].

In infants born earlier than normal (premature infants) the growth of retinal vessel stops before the retinal surface can completely be covered by retinal vessels. So the uncompleted sections would be unable to receive enough oxygen and food by blood circulation that consequently causes the disease. It is more common in infants weighing less than 1500 gr and gestational age less than 31 weeks[5]. Thus, the ROP normal growth of blood vessels stops and abnormal blood vessels grow, if left untreated it can lead to vision loss or blindness. Severity of ROP is determined by following factors:

- 1- Which area the new vessels have been located. Figure 2 illustrates the defined areas.
- 2- How much retinal vessel network is grown?
- 3- How much swollen the veins are?
- 4- The presence or absence of plus-disease.

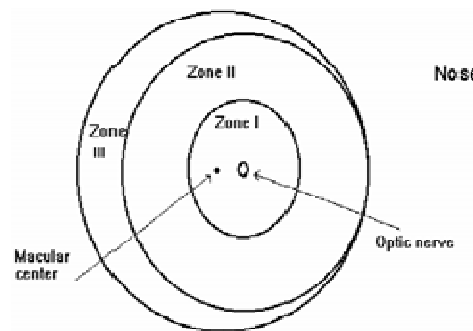


Figure 2. Defined areas of the eye [1]

Structural changes in the vessels can be studied in several ways:

- Examining by ophthalmoscope directly
- By examining images of the retina

Ophthalmoscope is an instrument that helps the physician so that they can see a perspective of the retina, it is challenging to use though, due to the need for expertise and experience. Figure 3 displays a typical ophthalmoscope.



Figure 3. Ophthalmoscope[2].

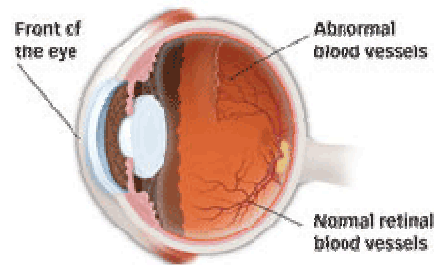
So it is important to have a system that can be used to increase the accuracy of the examination done by physicians or practitioners (e.g. by replacing trained nurses). Figures 4, 5 and 6 show abnormal growth, tortuosity of vessels and scarring of the retina. Given these figures, it seems it is possible to extract vessels automatically and use it to diagnose ROP.



Figure 4. Retinal vessel tortuosity and scars



Figure 5 Incomplete vascular growth



abnormal growth of blood vessels [1]

Factors that affect vessel segmentation are as follows:

- Vessel's width, shape and colour are not the same. Vessel's widths differ from one pixel to 12 pixels.
- There are other structures in images similar to vessel structures, such as retinal disc boundary, or nerves.
- Crossing points and bifurcations could confuse the techniques.
- Edge of the disk may be classified as vessel incorrectly.
- Sometimes the local contrast between the vessels and the background is very low. Particularly, thin veins have less contrast with their background [4, 6].

Despite many successful works in the field of image processing techniques performed on adult retinal images, most of these techniques fail when it comes to infants images. That's because several parameters are different in adults and infants images, such as resolution of images, blood vessels thickness and noise ratio. Vessel detection in images is more difficult in infants than in adults; however, both are equally important issues[7]. Figure 7 and 8 show structural differences between the infants and adult images.

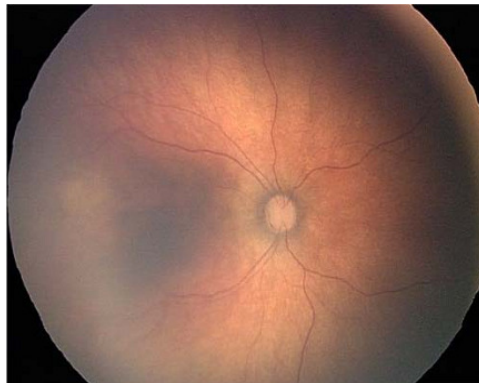


Figure7. An examples of infant retinal image

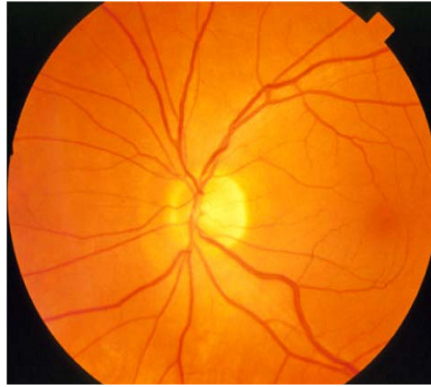


FIGURE 8. AN EXAMPLE OF ADULT RETINAL IMAGE [3].

2. MATERIALS AND METHODS

In this paper, the image is segmented in Lab color space. For this, clustering is applied on channels a and b in the Lab color space. The number of clusters are considered two, constituting of veins and background. After separating the gray segment, the ratio of the number of pixels in gray segment is evaluated to the total pixels in the image. This represents the area in which vessels are not fully grown and need regular checkups to ensure that it is fully developed in time and if not, the urgent treatments are needed to help the vascular structure grow normally. Figures 9, 10 and 11 show the clustering. The original image can be seen in Figure 9. Figure 10 and 11 represent the clusters including the background and the vessels respectively. The second cluster of pixels ratio to the total number of image pixels demonstrates the performance of this function in isolating the area of the vessels. In figure 11 this ratio was 2.17.

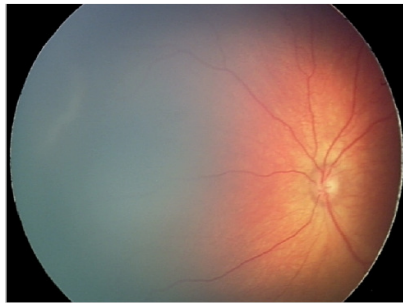


Figure 9. The original image

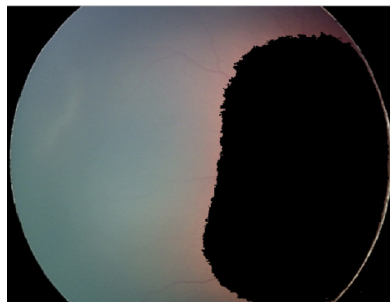


Figure 10. The first cluster, consisting undeveloped vessels

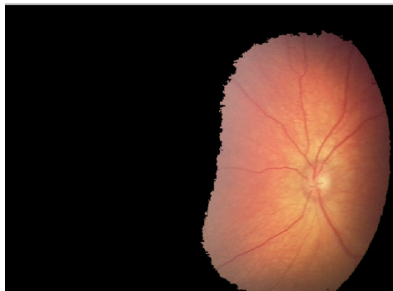


Figure 11. the second cluster with the ratio of 2.17

Table 1 illustrates the results of applying a threshold 0.5 to the ratios in terms of accuracy, specificity and sensitivity.

Table1. The result of thresholding the ratios with threshold 0.5

Sensitivity	Specificity	Accuracy
%74	%50	%69

In this paper, we also used the three operators named; Canny, Laplacian and Sobel operator to compare the results of edge detection in finding vessels the performance of these three are shown in Figures 12 to 18. After applying Laplacian operator, two different morphological operators used for the sake of noise removal; morphological opening with 3×3 squares and lines with a difference of 15 degrees (0, 15, 30, 60, 75, 90, 105, 120, 135, 150, 165) and 17 pixels length.

Laplacian filter can be seen in Figure 12, it has been accurately able to isolate the vessel. The morphological operators are used to remove noise from the image.

Figure 13 shows the result of opening morphological operator on Figure 12.

Figure 14 shows the result of applying lines 15 degree difference lines of length 12 to open Figure 12.

Figure 15, 16 are the result of applying Sobel and Canny operators respectively.



Figure 1. The result of applying Laplacian filter

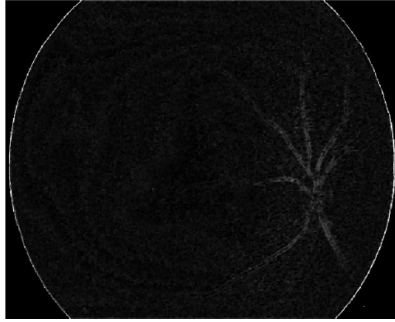


Figure 13. The result of opening by the 3 3 square

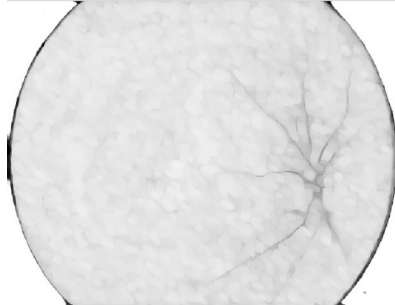


Figure 14. The result of opening by lines with a difference of 15 degrees

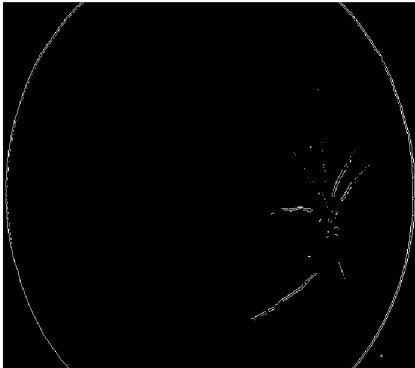


Figure 15. The result of applying Sobel operator

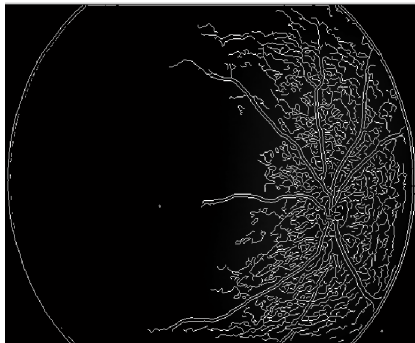


Figure16. The result of Canny operator

3. CONCLUSIONS

In this paper we compared methods of edge detection for extracting vessels. We also segmented the retinal vessels in Lab color space images. It is conducted from the results that noise is removed followed by Laplacian filters as well as the opening operator on Laplacian filtered images of the retina. Thus, the detection of retinal vessel network provides determining the possibility of ROP risk. This plays an important role for newborn infants who born with undeveloped retinal vessels to be followed in a regular basis so that it would prevent them from blindness in future.

REFERENCES

- [1] Available: http://en.wikipedia.org/w/index.php?title=Retinopathy_of_prematurity&oldid=621118923
- [2] facts about retinopathy of prematurity (ROP)," National Eye Instiute, october 2009. [Online]. Available: <http://www.nei.nih.gov/health/rop/rop.asp>. [Accessed 2012 may 14]. .
- [3] K. A. Vermeer, F. M. Vos, H. Lemij, and A. M. Vossepoel, "A model based method for retinal blood vessel detection," *Computers in Biology and Medicine*, vol. 34, pp. 209-219, 2004.
- [4] L. Gang, O. Chutatape, and S. M. Krishnan, "Detection and measurement of retinal vessels in fundus images using amplitude modified second-order Gaussian filter," *Biomedical Engineering, IEEE Transactions on*, vol. 49, pp. 168-172, 2002.
- [5] D. K. Wallace, "Diagnostic and Treatment Advances in Retinopathy of Prematurity," 2007.
- [6] M. Sofka and C. V. Stewart, "Retinal vessel centerline extraction using multiscale matched filters, confidence and edge measures," *Medical Imaging, IEEE Transactions on*, vol. 25, pp. 1531-1546, 2006.
- [7] L. Sukkaew, B. Uyyanonvara, S. S. Makhanov, S. Barman, and P. Panguthipong, "Automatic tortuosity-based retinopathy of prematurity screening system," *IEICE transactions on information and systems*, vol. 91, pp. 2868-2874, 2008.

AUTHOR INDEX

- Abdehamid Abdelhadi Mansor 255
Abdelhak Mansoul 105
Abdelkader Benyettou 01
Abdelmajid Hajami 177
Abdelmonaime Lachkar 83
Abduladhim Ashtaiwi 305
Abdusadik Saoud 305
Abirami S 21
Ahmed Mohammed Elsayi 255
Akhil Anjekar 97
Ameen Chilwan 207
Amel Boufrioua 189
Amir Hassan Monadjemi 61
Anna Filasova 267
Anna Filasova 279
Ashok Kumar J 21
Azade Rezaeezade 61
Baghdad Atmani 105
Baghdad Atmani 115
Bartolome T. Tanguilig 123
Besma Chaar Fayech 51
Davoodabadi M.R 231
Dev Gupta 317
Dusan Krokavec 267
Dusan Krokavec 279
Eugene Lim 169
Eugenia Litvinova 73
Eui Kyoung Shin 169
Farahani S.D 231
Fatih Korkmaz 13
Hakim Allali 177
Hayati Mamur 13
Hazim El-Mounayri 151
Ibrahim A.B 135
Ibrahim Almerhag 305
Ireyuwa Igbinosa 343
Ismail Nojavani 61
Ismail Topaloglu 13
Jai Gupta 317
Joao Jose Neto 291
Jose Maria Novaes dos Santos 291
Kashif Mahmood 207
Kody Varahramyan 151
Latesh Malik 245
Laura Felice 219
Lorena W. Rabago 123
Madhu D 197
Maher Rizkalla 333
Marcela Ridao 219
Maria Carmen Leonardi 219
Maria Virginia Mauco 219
Meryem Saadoune 177
Michael Jarschel 207
Miran Choi 169
Mohamed Benamina 115
Mohamed El-Sharkawy 333
Mohammed Bekkali 83
Murugappan S 21
Narasimhamu K.L 33
Narendra Kumar G 197
Nihar Athreyas 317
Noria Benyettou 01
Olav N. Østerbø 207
Olutayo Oyerinde 343
Parastou Shahsamandi E 43
Penghua Sun 333
Prashant Dahiwalé 245
Prashant Dahiwalé 97
Raghuwanshi M M 245
Rao C.S.P 33
Rapeepan Promyoo 151
Risty Moyo-Acerado 123
Rostamiyan Y 231
Sadka A.H 135
Santhoshkumar M K 197
Sofia Benbelkacem 115
Soheil Sadi-nezhad 43
Souad Mekni 51
Stanley Mneney 343
Suchita Tarare 97
Sunyoung Joung 169
Svetlana Chumachenko 73
Swarnalatha Srinivas 197
Venugopal Reddy V 33
Vincent Rodin 01
Vladimir Hahanov 73
Wajeb Gharibi 73
Wan Mohd Nasir Wan Kadir 255
Yilmaz Korkmaz 13
Young Ju Joo 169
Zhiguo Lai 317
Niousha Hormozi 355
Seyed Amirhassan Monadjemi 355
Gholamali Naderian 355