Computer Science & Information Technology 57

David C. Wyld
Jan Zizka (Eds)

Computer Science & Information Technology

The Third International Conference on Computer Science & Engineering
(CSEN 2016)
Dubai, UAE, August 27~28, 2016

AIRCC Publishing Corporation

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: 978-1-921987-55-7
DOI : 10.5121/csit.2016.61001 - 10.5121/csit.2016.61012

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

Preface

The Third International Conference on Computer Science & Engineering (CSEN 2016) was held in Dubai, UAE, during August 27~28, 2016. The Second International Conference on Signal and Pattern Recognition (SIPR 2016) and The Second International Conference of Networks, Communications, Wireless and Mobile Computing (NCWC 2016) were collocated with the CSEN-2016. 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 CSEN-2016, SIPR-2016, NCWC-2016 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, CSEN-2016, SIPR-2016, NCWC-2016 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 CSEN-2016, SIPR-2016, NCWC-2016.

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

Natarajan Meghanathan	Jackson State University, USA
Dhinaharan Nagamalai	Wireilla Net Solutions, Australia

Program Committee Members

Abd El-Aziz Ahmed	Cairo University, Egypt
Abdolreza Hatamlou	Islamic Azad University, Iran
Abe Zeid	Northeastern University, USA
Aiden B.Lee	University of La Jolla, USA
Ali AL-zuky	Mustansiriyah University, Iraq
Ali Dorri	Islamic Azad University, Iran
Ali Zaart	Beirut Arab University, Lebanon
Aloizio	Aeronautic Institute of Technology, Brasil
Amani Samha	Amani samha researcher QUT, Australia
Amol D Mali	University of Wisconsin, USA
Ankit Chaudhary	Truman State University, USA
Bai Li	Woodside Energy Ltd, Australia
Barbaros Preveze	Cankaya University, Turkey
Braham Barkat	The petroleum Institute, Saudi Arabia
Chandan Kumar Karmakar	University of Melbourne, Australia
Cheng fang	Zhejiang University, China
Chih-Lin Hu	National Central University, Taiwan
Chin-Chih Chang	Chung Hua University,Taiwan
Chiranjib Sur	University of Florida, US
Christian Esposito	National Research Council, Italy
Dac-Nhuong Le	Haiphong University, Vietnam
Danda B.Rawat	Georgia Southern University, USA
Derya Birant	Dokuz Eylul University, Turkey
Doina Bein	The Pennsylvania State University, USA
Dongchen Li	Peking University, China
Emilio Jiménez Macías	University of La Rioja, Spain
Epaminondas Kapetanios	University of Westminster, London
Erritali Mohammed	Sultan Moulay Slimane University, Morocco
Fatih Korkmaz	Cankiri Karatekin University, Turkey
Gullanar M Hadi	Salahaddin University, Iraq
Hacene Belhadef	University of Constantine 2, Algeria
Hamdi M	National Engineering School of Tunis, Tunisia
Hesham Farouk	Electronics Research Institute, Egypt
Hossein Jadidoleslamy	University of Zabol, Iran
Houcine Hassan	Univeridad Politecnica de Valencia, Spain
Huiyu Zhou	Queen's University Belfast, United Kingdom
Hyunsung Kim	Kyungil University, Korea
Iman Saroit	Cairo University, Egypt

Isa Maleki	Islamic Azad University, Iran
Islam Atef	Alexandria University, Egypt
Israashaker Alani	Gaziantep University, Turkey
Jacques Epounde Ngalle	Robert Morris University, USA
Jan Lindstrom	MariaDB Corporation, Finland
Jerin Cyriac	Truman State University, USA
Jose Raniery	University of Sao Paulo, Brazil
Juan A. Fraire	Universidad Nacional de Crdoba, Argentina
Kassim S.Mwitondi	Sheffield Hallam University, United Kingdom
Kayhan Erciyes	Izmir University,Turkey
Kemal	Abant Izzet Baysal University, Turkey
Kenneth Mapoka	Iowa state university, USA
Li Zheng	University of Bridgeport, USA
Lorena Gonz lez Manzano	University Carlos III of Madrid, Spain
Mahdi Mazinani	IAU Shahreqods, Iran
Mahi Lohi	University of Westminster, UK
Malka N.Halgamuge	Melbourne School of Engineering, Australia
Mary M.Eshaghian-Wilner	University of Southern California, USA
Maziar Loghman	Illinois Institute of Technology, USA
Mehrdad Jalali	Mashhad Azad University, Iran
Mohamed AlAjmi	King Saud University, Saudi Arabia
Mohamed Ashik M	Salalah College of Technology, Oman
Mohamed Fahad AlAjmi	King Saud University, Saudi Arabia
Muhammad Abrar	Massey University, New Zealand
Muhammad Sarfraz	Kuwait University, Kuwait
Mujiono Sadikin	Universitas Mercu Buana, Indonesia
Nabila Labraoui	University of Tlemcen, Algeria
Nadia Qadri	University of Essex, United Kingdom
Najib A. Odhah	IBB university, Yemen
Neetesh Saxena	The State University of New York, USA
Nourddine Bouhmala	Buskerud and Vestfold University, Norway
Othmane Alaoui Fdili	Mohammed V University, Morocco
Ouarda Barkat	University Frères Mentouri, Algeria
Oussama Ghorbel	University of Troyes, Tunisia
Peiman Mohammadi	Islamic Azad University, Iran
Peter Ogedebe	BAZE University, Nigeria
Rafah M. Almuttairi	University of Babylon, Iraq
Rahil Hosseini	Islamic Azad University, Iran
Rahul Kosarwal	OAARs CORP, India
Ramayah T	Universiti Sains Malaysia, Malaysia
Ramon Adeogun	Victoria University of Wellington, New Zealand
Raveendra K Rao	University of Western Ontario, Canada
Rhattoy Abdallah	Moulay Ismail University, Morocco
Ricardo De Carvalho Destro	University Center of FEI, Brazil
Saad Darwish	Alexandria University, Egypt
Saeid Asgari Taghanaki	Azad University, Iran
Samadhiya	National Chiao Tung University, Taiwan
Sergey Muravyov	Tomsk Polytechnic University, Russia

Technically Sponsored by

Networks & Communications Community (NCC)

Computer Science & Information Technology Community (CSITC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

The Third International Conference on Computer Science & Engineering (CSEN 2016)

The Second International Conference on Signal and Pattern Recognition (SIPR 2016)

The Second International Conference of Networks, Communications, Wireless and Mobile Computing (NCWC 2016)

NOVEL LOGIC CIRCUITS DYNAMIC PARAMETERS ANALYSIS

Nicolae Galupa

Higher Colleges of Technology – Ras al Khaimah, UAE
ngalupa@hct.ac.ae
Department of Computer Engineering, Technical University Iasi – Romania
nky@cs.tuiasi.ro

ABSTRACT

Combinational logic circuit timing analysis is an important issue that all designers need to address. The present paper presents a simple and compact analysis procedure. We follow the guidelines drawn by previous methods, but we shall define new time-dependent logic variables that help us improve their efficiency. By using the methodology suggested, we shall replace a very laborious technique (pure delay circuit + time constants method) with a simpler procedure that can pinpoint the specific conditions for a logic circuit's anomalous behaviour within a few simple steps. Considering the logic function implemented the methodology presented will require analysis of only a limited number of situations/combinations to determine the presence of an anomalous behaviour. When anomalous behaviour is identified, the methodology provides a clear timing description.

KEYWORDS

Logic Design, Timing, Time Dependant Logic Variables

1. INTRODUCTION

The present work focuses on issues regarding the anomalous functioning of logic circuits. We shall address the static and dynamic hazards defined by J.Beister, E.J. McCluskey, R.F. Tinder and J. Brzozowski [1-3].

At present, we distinguish two analysis methodologies to determine and eventually describe the presence of a hazard in a logic circuit output. The first approach is a purely algebraic one that considers the logic function implemented by a logic circuit and identifies specific algebraic patterns that are responsible for the presence of a hazard. As presented by E.J. McCluskey and R.F. Tinder [2-3], these are $x+\overline{x}$, $x\cdot\overline{x}$ for a static hazard and $x+x\cdot\overline{x}$, $x\left(x+\overline{x}\right)$ for a dynamic hazard, where x is a component of the input vector driving the analysed logic function. This method reaches its goal by algebraically manipulating the logic function and using binary decision graphs, as presented by R. Bryant, S. Ackers, S.M. Nowick, C. Jeong, Berthomieu B. and J. Brzozowski [4-12]. However, this method, will not describe the hazard's evolution (at least not completely), meaning that no timing information will be revealed. Additionally, one can easily note that this method requires a high computational effort.

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 01–20, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61001

This paper presents an improvement of the above mentioned method, improvement that allows the designer to determine whether a hazard is present by simply analysing the individual terms present in the logic function's expression, either in SOP form (disjunctive form) or in POS form (conjunctive form). It is my opinion that the improvement presented, if used in conjunction with the classical method, will maintain reliability and will lower the computational effort required

The second approach considers the implementation of a logic function, so analysis will be performed on a completely defined combinational logic circuit (CLC). Basically, we use a pure delay circuit model plus individual in-out path definition, as presented by E.J. McCluskey [2]. Following this procedure, all distinct in → out paths are revealed, and we determine:

- a completely defined delay vector for the circuit,
- the association of each input variable to the path (paths) it crosses towards the output + its specific delay, and
- the ideal logic circuit implementing the logic function.

The method has been described by J. Beister [1], developed by O. Maler and A. Martello [13-15] and exceptionally applied by R.K Brayton [16]. An improvement that considers the inequality between the specific delays for "1"↓"0"and "0"↑"1" transitions has been presented by N. Galupa [17-18]. The rules applied for operating with the gate-specific delays have been presented by K.S. Stevens, R.B. Salah and M. Bogza [19-21]. Please bear in mind that the method presented is not confined only to acyclic combinational circuits, as proven by M. Riedel [22]. The method, also known as the time constants method, allows us to determine the moment of time when the circuit's output has stabilized. However, it will not easily provide information on the behaviour of the logic circuit output prior to stabilization. Should the analysed CLC be used to implement an automaton, its dynamic parameters are critical.

The second section will present a methodology that allows us to easily determine (logic computations only) the dynamic parameters of the circuit output (including active hazard).

Both procedures presented are based on describing the logic variables involved (and, of course, the associated electric signals) with respect to two notions:

- Logic Value – the normal use of a logic variable
- Time – will express the evolution of the variable with respect to time. This is why we shall refer to these variables as time-dependent logic variables (TDLVs).

The present paper is organized as follows:

- Definitions - in this section, we will define the TDLVs
- Properties – in this section, we present and prove the specific properties of TDLVs showing why these variables are useful for the analysis of logic circuit behaviour.

- Analysis procedure for a logic function. We prove, within this section, that the fulfilment of a simple condition will pinpoint the presence of a hazard on the circuit output, thus drastically reducing the computations required.

- Analysis methodology. This section will make use of the TDLVs to determine the hazard generating patterns associated with the logic patterns described by E.J. McCluskey and R.F. Tinder [2-3]

- Finally, a complete example using these methodologies.

2. DEFINITIONS

2.1. Time-Dependent Logic Variables (TDLVs)

Whenever a logic variable applied on a logic gate input changes value, it triggers a process that can be observed on the gate's output connection. However, the gate's output will maintain its previous level for a predetermined period of time – specifically, the gate's propagation time. This situation is bothersome when we expect the gate's output to switch to its complementary value as a result of the input change. Therefore, we shall define a logic variable, denoted τ, to be used for describing the gate's output level evolution as a result of an input variable change, with respect to time.

$$\tau_x=\tau\ (t_x-t)\ = \begin{cases} 0 \text{ for } t<t_x \\ 1 \text{ for } t\geq t_x \end{cases}$$

One can easily note that τ will help us describe a specific moment of time that is generally associated with a level transition in a signal. Obviously, we also need to be able to describe a time interval, so we shall define the logic variable δ as follows. Let us consider two moments of time ta and tb respecting ta<tb. Under these conditions, δ is defined as:

$$\delta(t_a,t_b)=\delta_{a,b} \quad = \begin{cases} 1 \text{ for } t_a\leq t<t_b \\ 0 \quad \text{otherwise} \end{cases}$$

2.2. Term weight

The weight of a logic term (conjunctive or disjunctive form) is a vector $w=(w_1,w_2,\ldots w_{n-1},w_n)$, where w_k {0,1}. If $w_k=0$, then the k component of the term is complemented; otherwise, it is in its direct form.

Example: w=0101 → the logic term is $\overline{a_1}\cdot a_2\cdot \overline{a_3}\cdot a_4$ (conjunctive form) or $\overline{a_1}+a_2+\overline{a_3}+a_4$ (disjunctive form)

3. PROPERTIES

3.1. Properties of τ and δ variables

Let us consider three ordered time stamps, t_1, t_2, and t_3 respecting $t_1<t_2<t_3$. We shall consider all possible logic terms that can be defined using three independent logic variables (both conjunctive and disjunctive forms) that have these three time stamps associated as switching moments. The reduced terms that result when τ and δ logic variables are used to express logic terms are presented in table 1. Proof has been provided by Galupa [18].

Table 1 τ,δ MINTERM AND MAXTERM PROPERTIES

Weight	Term
000	$\overline{\tau_1} \cdot \overline{\tau_2} \cdot \overline{\tau_3} = \overline{\tau_1}$ $\overline{\tau_1} + \overline{\tau_2} + \overline{\tau_3} = \overline{\tau_3}$
001	$\overline{\tau_1} \cdot \overline{\tau_2} \cdot \tau_3 = 0$ $\overline{\tau_1} + \overline{\tau_2} + \tau_3 = \overline{\delta_{2,3}}$
010	$\overline{\tau_1} \cdot \tau_2 \cdot \overline{\tau_3} = 0$ $\overline{\tau_1} + \tau_2 + \overline{\tau_3} = 1$
011	$\overline{\tau_1} \cdot \tau_2 \cdot \tau_3 = 0$ $\overline{\tau_1} + \tau_2 + \tau_3 = \overline{\delta_{1.2}}$
100	$\tau_1 \cdot \overline{\tau_2} \cdot \overline{\tau_3} = \delta_{1,2}$ $\tau_1 + \overline{\tau_2} + \overline{\tau_3} = 1$
101	$\tau_1 \cdot \overline{\tau_2} \cdot \tau_3 = 0$ $\tau_1 + \overline{\tau_2} + \tau_3 = 1$
110	$\tau_1 \cdot \tau_2 \cdot \overline{\tau_3} = \delta_{2,3}$ $\tau_1 + \tau_2 + \overline{\tau_3} = 1$
111	$\tau_1 \cdot \tau_2 \cdot \tau_3 = \tau_3$ $\tau_1 + \tau_2 + \tau_3 = \tau_1$

Clearly, if we consider n distinct time stamps t_1, t_2,…, t_n respecting $t_1<t_2< \ldots<t_{n-1}<t_n$ and following the same logic path as above, we will reach the reduced expressions for the n-component logic terms, as presented in table. II.

Table 2. GENERAL τ,δ MINTERM AND MAXTERM PROPERTIES

Weight		
000…000	$\overline{\tau_1} \cdot \overline{\tau_2} \cdot \overline{\tau_3} \cdot \ldots \cdot \overline{\tau_{n-2}} \cdot \overline{\tau_{n-1}} \cdot \overline{\tau_n} =$ $\overline{\tau_1} + \overline{\tau_2} + \overline{\tau_3} + \ldots + \overline{\tau_{n-2}} + \overline{\tau_{n-1}} + \overline{\tau_n} =$	$\overline{\tau_1}$ $\overline{\tau_n}$
000…001	$\overline{\tau_1} \cdot \overline{\tau_2} \cdot \overline{\tau_3} \cdot \ldots \cdot \overline{\tau_{n-2}} \cdot \overline{\tau_{n-1}} \cdot \tau_n =$ $\overline{\tau_1} + \overline{\tau_2} + \overline{\tau_3} + \ldots + \overline{\tau_{n-2}} + \overline{\tau_{n-1}} + \tau_n =$	0 $\overline{\delta_{n-1,n}}$
000…010	$\overline{\tau_1} \cdot \overline{\tau_2} \cdot \overline{\tau_3} \cdot \ldots \cdot \overline{\tau_{n-2}} \cdot \tau_{n-1} \cdot \overline{\tau_n} =$ $\overline{\tau_1} + \overline{\tau_2} + \overline{\tau_3} + \ldots + \overline{\tau_{n-2}} + \tau_{n-1} + \overline{\tau_n} =$	0 1
…	…	…

00...01...11	$\overline{\tau_1} \cdot ... \cdot \overline{\tau_{i-1}} \cdot \tau_i \cdot ... \cdot \tau_n =$ $\overline{\tau_1} + ... + \overline{\tau_{i-1}} + \tau_i + ... + \tau_n =$	0 $\overline{\delta_{i-1,i}}$
…	…	…
011…111	$\overline{\tau_1} \cdot \tau_2 \cdot \tau_3 \cdot ... \cdot \tau_{n-2} \cdot \tau_{n-1} \cdot \tau_n =$ $\overline{\tau_1} + \tau_2 + \tau_3 + ... + \tau_{n-2} + \tau_{n-1} + \tau_n =$	0 $\overline{\delta_{1,2}}$
100…000	$\tau_1 \cdot \overline{\tau_2} \cdot \overline{\tau_3} \cdot ... \cdot \overline{\tau_{n-2}} \cdot \overline{\tau_{n-1}} \cdot \overline{\tau_n} =$ $\tau_1 + \overline{\tau_2} + \overline{\tau_3} + ... + \overline{\tau_{n-2}} + \overline{\tau_{n-1}} + \overline{\tau_n} =$	$\delta_{1,2}$ 1
…	…	…
111…110	$\tau_1 \cdot \tau_2 \cdot \tau_3 \cdot ... \cdot \tau_{n-2} \cdot \tau_{n-1} \cdot \overline{\tau_n} =$ $\tau_1 + \tau_2 + \tau_3 + ... + \tau_{n-2} + \tau_{n-1} + \overline{\tau_n} =$	$\delta_{n-1,n}$ 1
111…111	$\tau_1 \cdot \tau_2 \cdot \tau_3 \cdot ... \cdot \tau_{n-2} \cdot \tau_{n-1} \cdot \tau_n =$ $\tau_1 + \tau_2 + \tau_3 + ... + \tau_{n-2} + \tau_{n-1} + \tau_n =$	τ_n τ_1

Please observe that in table 2, only the extremities (meaning the border time stamps, expressed by the terms associated with weights w=000…000 and w=111…111) are coherent when operating with τ. Therefore, we can safely state that when operating the terms present in a function equation with respect to time (meaning, using τ and δ), only $\tau_1=\tau(t-t_1)$ and $\tau_n=\tau(t-t_n)$ will be present in the final expression (according to the specific function's equation).

On the other hand, whenever the term in question is characterized by an ordered weight (i.e., w=000…01…111 or w=111…10…000), that term will contribute to the final expression only by pinpointing a time interval (t_{k-1}, t_k) by means of δ.

The other terms present in the analysed logic expression are either logic "1" (disjunctive) or "0" (conjunctive).

4. Analysis Procedure

Let there be a logic function y=f(a,b,c,d,…). According to the time constants method (Beister[1], McCluskey[2]), we define the individual in-out pathways by virtually replicating all gates characterized by a fan-out larger than one, replace the gates with their equivalent ideal gates (zero delay) plus their specific propagation time (delay operator) and propagate all delay operators from output → input. Finally, we will reach a structure that will present all individual in-out pathways and their specific delays followed by an ideal logic circuit.

Let us consider that logic variable a (input vector component) crosses n distinct paths towards the output, so that primary input variable a generates n distinct secondary variables $(a_1,a_2,…,a_n)$ characterized by n distinct delays $(t_1, t_2,…,t_n)$. We assume that the secondary input vector has been organized so all specific path delays respect $t_1<t_2<…<t_{n-1}<t_n$.

Therefore, a change in input variable a will generate a sequence of changes in the secondary input variables $(a_1, a_2,…,a_n)$ ordered and spaced in time according to $(t_1, t_2,…,t_n)$. Subsequently, because the secondary input vector is driving an instantaneous ideal logic circuit, we shall observe a sequence of transitions in the circuit's output.

Considering the definition of τ, we can write:

$$a_x(t)=a_{xinitial}\oplus\tau(t_x-t) \qquad (4.1.)$$

where $a_{xinitial}$ is the initial value from which a_x evolves. Note that $a_x(t)$ will maintain $a_{xinitial}$ level until we reach t_x time stamp and switches to the complementary value afterwards.

The function's expression becomes:

$$y(t) = f[a_1(t), a_2(t), ..., a_n(t)] = f[a \oplus \tau_1(t_1 - t), a \oplus \tau_2(t_2 - t), ..., a \oplus \tau_n(t_n - t)] \qquad (4.2.)$$

Ultimately, a logic function can be expressed in a conjunctive or disjunctive form, therefore, considering y(t) presented by eq. 4.2 and the properties presented in table 2, we conclude that only a strictly limited and well determined number of terms will be present.

To present how this works, we shall first consider a particular case – only three secondary variables for the primary input variable considered, a_1, a_2, and a_3, characterized by t_1, t_2, and t_3, respecting $t_1< t_2< t_3$. Afterwards, we shall generalize to n secondary variables.

The function's expression becomes:

$$y(t) = \alpha_{000}\overline{a_1(t)} \cdot \overline{a_2(t)} \cdot \overline{a_3(t)} + \alpha_{001}\overline{a_1(t)} \cdot \overline{a_2(t)} \cdot a_3(t) + ... + \alpha_{111}a_1(t) \cdot a_2(t) \cdot a_3(t) \qquad (4.3.)$$

$$y(t) = [\overline{\alpha_{000}} + \overline{a_1(t)} + \overline{a_2(t)} + \overline{a_3(t)}] \cdot [\overline{\alpha_{001}} + \overline{a_1(t)} + \overline{a_2(t)} + a_3(t)] \cdot ... \cdot [\overline{\alpha_{111}} + a_1(t) + a_2(t) + a_3(t)] \qquad (4.4.)$$

where, $\alpha_{w,y,z}$ {0,1} and is defined as follows:

- $\alpha_{w,y,z}$= 0→term characterised by weight wyz is not present;
- $\alpha_{w,y,z}$=1→term characterised by weight wyz is present.

For further computations we'll consider the function's expression presented by eq. 4.3. Computations for the other form (eq. 4.4.) are similar.

$$y(t) = \alpha_{000}\overline{a \oplus \tau_1} \cdot \overline{a \oplus \tau_2} \cdot \overline{a \oplus \tau_3(t)} + \alpha_{001}\overline{a \oplus \tau_1} \cdot \overline{a \oplus \tau_2} \cdot a \oplus \tau_3 + ... + \alpha_{111}a \oplus \tau_1 \cdot a \oplus \tau_2 \cdot a \oplus \tau_3 \qquad (4.5.)$$

with a being the initial value for the variable switching during the process. Now, we shall operate each term and reduce it according to the properties presented for τ.

$$\text{w=000}\rightarrow \overline{a \oplus \tau_1} \cdot \overline{a \oplus \tau_2} \cdot \overline{a \oplus \tau_3} = (a\tau_1 + \overline{a\tau_1})(a\tau_2 + \overline{a\tau_2})(a\tau_3 + \overline{a\tau_3}) = a\tau_1\tau_2\tau_3 + \overline{a\tau_1\tau_2\tau_3} = a\tau_3 + \overline{a\tau_1} \qquad (4.6.)$$

$$\text{w=001}\rightarrow \overline{a \oplus \tau_1} \cdot \overline{a \oplus \tau_2} \cdot a \oplus \tau_3 = (a\tau_1 + \overline{a\tau_1})(a\tau_2 + \overline{a\tau_2})(\overline{a}\tau_3 + a\overline{\tau_3}) = a\tau_1\tau_2\overline{\tau_3} + \overline{a\tau_1\tau_2}\tau_3 = a\delta_{2,3} + \overline{a}0 = a\delta_{2,3} \qquad (4.7.)$$

……………………………………………………………………………………………………..

$$\text{w=111}\rightarrow (a \oplus \tau_1) \cdot (a \oplus \tau_2) \cdot (a \oplus \tau_3) = (\overline{a}\tau_1 + a\overline{\tau_1})(\overline{a}\tau_2 + a\overline{\tau_2})(\overline{a}\tau_3 + a\overline{\tau_3}) = a\overline{\tau_1\tau_2\tau_3} + \overline{a}\tau_1\tau_2\tau_3 = a\overline{\tau_1} + \overline{a}\tau_3 \qquad (4.8.)$$

Following the same procedure and using the properties listed above, we find:

Table 3. REDUCED TIME-DEPENDENT SECONDARY TERMS

weight		
000	$\overline{a\oplus\tau_1}\cdot\overline{a\oplus\tau_2}\cdot\overline{a\oplus\tau_3} = a\tau_1\tau_2\tau_3 + \overline{a}\overline{\tau_1}\overline{\tau_2}\overline{\tau_3} =$	$a\tau_3 + \overline{a}\overline{\tau_1}$
001	$\overline{a\oplus\tau_1}\cdot\overline{a\oplus\tau_2}\cdot a\oplus\tau = a\tau_1\tau_2\overline{\tau_3} + \overline{a}\overline{\tau_1}\overline{\tau_2}\tau_3 =$	$a\delta_{2,3}$
010	$\overline{a\oplus\tau_1}\cdot a\oplus\tau_2\cdot\overline{a\oplus\tau_3} == a\tau_1\overline{\tau_2}\tau_3 + \overline{a}\overline{\tau_1}\tau_2\overline{\tau_3} =$	0
011	$\overline{a\oplus\tau_1}\cdot a\oplus\tau_2\cdot a\oplus\tau_3 = a\tau_1\overline{\tau_2}\overline{\tau_3} + \overline{a}\overline{\tau_1}\tau_2\tau_3 =$	$a\delta_{1,2}$
100	$a\oplus\tau_1\cdot\overline{a\oplus\tau_2}\cdot\overline{a\oplus\tau_3} = a\overline{\tau_1}\tau_2\tau_3 + \overline{a}\tau_1\overline{\tau_2}\overline{\tau_3} =$	$\overline{a}\delta_{1,2}$
101	$a\oplus\tau_1\cdot\overline{a\oplus\tau_2}\cdot a\oplus\tau_3 = a\overline{\tau_1}\tau_2\overline{\tau_3} + \overline{a}\tau_1\overline{\tau_2}\tau_3 =$	0
110	$a\oplus\tau_1\cdot a\oplus\tau_2\cdot\overline{a\oplus\tau_3} = a\overline{\tau_1}\overline{\tau_2}\tau_3 + \overline{a}\tau_1\tau_2\overline{\tau_3} =$	$\overline{a}\delta_{2,3}$
111	$a\oplus\tau_1\cdot a\oplus\tau_2\cdot a\oplus\tau_3 = a\overline{\tau_1}\overline{\tau_2}\overline{\tau_3} + \overline{a}\tau_1\tau_2\tau_3 =$	$a\overline{\tau_1} + \overline{a}\tau_3$

Therefore, the function's expression becomes:

$$y(t) = \alpha_{000}[a\tau_3 + \overline{a}\overline{\tau_1}] + \alpha_{001}a\delta_{2,3} + \alpha_{011}a\delta_{1,2} + \alpha_{100}\overline{a}\delta_{1,2} + \alpha_{110}\overline{a}\delta_{2,3} + \alpha_{111}[a\overline{\tau_1} + \overline{a}\tau_3] \qquad (4.9.)$$

Note that only the border terms influence the final expression ($a\cdot\tau_3 + \overline{a}\cdot\overline{\tau_1}$ if α_{000}=1, meaning a term with weight w=000 - $\overline{a_1}\cdot\overline{a_2}\cdot\overline{a_3}$ - is present, and / or $a\cdot\overline{\tau_1} + \overline{a}\cdot\tau_3$ if α_{111}=1, meaning a term with weight w=111 - $a_1\cdot a_2\cdot a_3$ - is present).

Considering that the time stamps are ordered $t_1<t_2<t_3$, the individual time intervals (t_1,t_2) and/or (t_2,t_3) will be pinpointed by $\delta_{1,2}$ and $\delta_{2,3}$ if α_{011} and α_{001}, respectively, are logic 1 and / or $\overline{\delta_{1,2}}$ and $\overline{\delta_{2,3}}$ if α_{011} and α_{001}, respectively, are logic 1.

Please observe that in the final expression, we have succeeded in significantly decreasing the number of terms involved. Also note that considering eq. 4.9, we can trace the output's behaviour with respect to time.

First, we identify the influence that each member of eq. 4.9. has on the overall output presented in Table. 4.

Table. 4. OUTPUT TRACE FOR VARIABLE a

	t_{init}	t_1	t_2	t_3	t_{final}
1. α_{000}=1⇒y(t)=	$\overline{a}$	0	0	a	
2. α_{001}=1⇒y(t)=	0	0	a	0	
3. α_{011}=1⇒y(t)=	0	a	0	0	
4. α_{100}=1⇒y(t)=	0	$\overline{a}$	0	0	
5. α_{110}=1⇒y(t)=	0	0	$\overline{a}$	0	
6. α_{111}=1⇒y(t)=	a	0	0	$\overline{a}$	

- The border terms (w=111 and w=000 – lines 1 and 6) will never generate a hazard by themselves, independent of variable a’s initial value.

	t_{init}	t_1	t_2	t_3	t_{final}
α_{000}=1and a=0⇒y(t)=	1	0	0	0	
α_{000}=1and a=1⇒y(t)=	0	0	0	1	
α_{111}=1and a=0⇒y(t)=	0	0	0	1	
α_{111}=1and a=1⇒y(t)=	1	0	0	0	

- The terms characterized by a uniform ordered weight will generate a static hazard.

	t_{init}	t_1	t_2	t_3	
α_{001}=1and a=1⇒y(t)=	0	0	1	0	Static hazard for “1”↓”0”transition
α_{011}=1and a=1⇒y(t)=	0	1	0	0	Static hazard for “1”↓”0”transition
α_{100}=1and a=0⇒y(t)=	0	1	0	0	Static hazard for ”0”↑”1”transition
α_{110}=1and a=0⇒y(t)=	0	0	1	0	Static hazard for ”0”↑”1”transition

- A dynamic hazard will be present if a combination of border terms and ordered weight terms is encountered

	t_{init}	t_1	t_2	t_3	
α_{000}=1and a=0⇒y(t)=	1	0	0	0	
α_{110}=1and a=0⇒y(t)=	0	0	1	0	
Overall behavior	**1**	**0**	**1**	**0**	**Dynamic hazard for “0”↑”1” transition**
α_{000}=1and a=1⇒y(t)=	0	0	0	1	
α_{011}=1and a=1⇒y(t)=	0	1	0	0	
Overall behavior	**0**	**1**	**0**	**1**	**Dynamic hazard for “1”↓”0” transition**
α_{111}=1and a=0⇒y(t)=	0	0	0	1	
α_{100}=1and a=0⇒y(t)=	0	1	0	0	
Overall behavior	**0**	**1**	**0**	**1**	**Dynamic hazard for “0”↑”1” transition**
α_{111}=1and a=1⇒y(t)=	1	0	0	0	
α_{001}=1and a=1⇒y(t)=	0	0	1	0	
Overall behavior	**1**	**0**	**1**	**0**	**Dynamic hazard for “1”↓”0” transition**

Additionally, please observe that the previous analysis presents us with a way to mask an existing dynamic hazard. If we encounter such a situation, all we should do is activate the appropriate term (if algebraically possible) that will fill in the glitch forming the dynamic hazard. Possible cases are listed below:

Table. 5. HAZARD MASKING CASES

CASE 1	t_{init}	t_1	t_2	t_3	
α_{000}=1and a=0⇒y(t)=	1	0	0	0	Dynamic hazard for a→”0”↑”1”
α_{110}=1and a=0⇒y(t)=	1	0	1	0	
Initial behavior	**1**	**0**	**1**	**0**	
α_{100}=1and a=0⇒y(t)=	0	1	0	0	Added term (if possible)
Final overall behavior	**1**	**1**	**1**	**0**	Hazard masked

CASE 2	t_{init}	t_1	t_2	t_3	
α_{000}=1and a=1⇒y(t)=	0	0	0	1	Dynamic hazard for a→"1"↓ "0"
α_{011}=1and a=1⇒y(t)=	0	1	0	0	
Initial behavior	**0**	**1**	**0**	**1**	
α_{001}=1and a=1⇒y(t)=	0	0	1	0	Added term (if possible)
Final overall behavior	**0**	**1**	**1**	**1**	Hazard masked

CASE 3	t_{init}	t_1	t_2	t_3	
α_{111}=1and a=0⇒y(t)=	0	0	0	1	Dynamic hazard for a→"0"↑"1"
α_{100}=1and a=0⇒y(t)=	0	1	0	0	
Initial behavior	**0**	**1**	**0**	**1**	
α_{110}=1and a=0⇒y(t)=	0	0	1	0	Added term (if possible)
Final overall behavior	**0**	**1**	**1**	**1**	Hazard masked

CASE 4	t_{init}	t_1	t_2	t_3	
α_{111}=1and a=1⇒y(t)=	1	0	0	0	Dynamic hazard for a→"1"↓ "0"
α_{001}=1and a=1⇒y(t)=	0	0	1	0	
Initial behavior	**1**	**0**	**1**	**0**	
α_{011}=1and a=1⇒y(t)=	0	1	0	0	Added term (if possible)
Final overall behavior	**1**	**1**	**1**	**0**	Hazard masked

The instrument presented above works when dealing with minterms / maxterms. Obviously, this is not always the case. However, whenever we address a term that is not complete (meaning that it lacks input components), we shall analyse whether its weight is ordered. If not, no further inquiries are required, as the term will not generate anomalous behaviour. On the other hand, if the weight is ordered, the term should be expanded to canonical form (without altering the function's truth value) and an analysis performed.

In case of an n-component secondary input vector (a_1, a_2,…,a_n) generated by primary input variable a, characterized by t_1, t_2, …, t_n ordered timestamps ($t_1<t_2<\ldots<t_n$), the function's expression becomes:

$$\begin{aligned} y(t) = {} & \alpha_{000...000}[a\tau_n + \overline{a}\overline{\tau_1}] + \alpha_{000...001}\delta_{n-1,n} + ... + \alpha_{000...01...111}a\delta_{i-1,i} + ... + \alpha_{011...11...111}a\delta_{1,2} + \\ & + \alpha_{100...000}\overline{a}\delta_{1,2} + ... + \alpha_{111....10..000}\overline{a}\delta_{j-1,j} + ... + \alpha_{111...110}\overline{a}\delta_{n-1,n} + \alpha_{111...110}[a\overline{\tau_1} + \overline{a}\tau_n] \quad (4.10.) \end{aligned}$$

Therefore, the function's output trace is presented by Table.6.

Once again, note that we have succeeded in significantly decreasing the number of terms involved. Just identifying the presence of specific terms (border and/or weight ordered) will be enough to conclude whether anomalous behaviour is present. Furthermore, in particular situations, we can also pinpoint the methodology for masking the anomalous behaviour, should this be our goal.

Table. 6. GENERAL OUTPUT TRACE FOR VARIABLE a

	t_{init}	t_1	t_2	t_3	…	t_{n-2}	t_{n-1}	t_n	t_{final}
$\alpha_{000...000}=1 \rightarrow y(t)=$	$\overline{a}$	0	0	0	…	0	0	a	
$\alpha_{000...001}=1 \rightarrow y(t)=$	0	0	0	0	…	0	a	0	
$\alpha_{000...011}=1 \rightarrow y(t)=$	0	0	0	0	…	a	0	0	
……………………………………………………									
$\alpha_{001...111}=1 \rightarrow y(t)=$	0	0	a	0	…	0	0	0	
$\alpha_{011...111}=1 \rightarrow y(t)=$	0	a	0	0	…	0	0	0	
$\alpha_{100...000}=1 \rightarrow y(t)=$	0	$\overline{a}$	0	0	…	0	0	0	
$\alpha_{110...000}=1 \rightarrow y(t)=$	0	0	$\overline{a}$	0	…	0	0	0	
……………………………………………………									
$\alpha_{111...100}=1 \rightarrow y(t)=$	0	0	0	0	…	$\overline{a}$	0	0	
$\alpha_{111...110}=1 \rightarrow y(t)=$	0	0	0	0	…	0	$\overline{a}$	0	
$\alpha_{111...111}=1 \rightarrow y(t)=$	a	0	0	0	…	0	0	$\overline{a}$	

The singular presence of a term with the weight ordered is a sufficient condition for the presence of a static hazard.

5. CIRCUIT ANALYSIS METHODOLOGY

The previous paragraph presented how to determine whether a logic function output exhibits anomalous behaviour. Another approach is required, if our goal is to determine the occurrence of an anomalous pulse(s) on a logic circuit's output and eventually determine its dynamic parameters.

Basically, we shall determine the secondary input vector and its associated delays vector using already established methods (J. Beister[1], McCluskey[2]). At this point, we will use the TDLVs (τ and δ) to express all variables considering the time variable. We reduce the logic function's expression to a minimal one (Boolean rules), and we will attempt to identify the hazard generating patterns (static or dynamic).

Notation rules:

- delay for NOT gate → t_N / and associated $\tau_N=\tau\ (t_N-t)$
- delay for AND gate → t_A / and associated $\tau_A=\tau\ (t_A-t)$
- delay for OR gate → t_O / and associated $\tau_O=\tau\ (t_O-t)$
- delay for NAND gate → t_{NA} / and associated $\tau_{NA}=\tau\ (t_{NA}-t)$
- delay for NOR logic gate → tNO / and associated $\tau_{NO}=\tau\ (t_{NO}-t)$
- $t_N+t_O = {}_{tN+O}$ / and associated $\tau_{N+O}=\tau\ (t_{N+O}-t)$, so.on.

To determine the time-dependent hazard generating patterns, we shall start from the circuit patterns as presented by E.J. McCluskey and R.F. Tinder [2-3].

5.1. Static - Case 1

Algebraic description → $y = x \cdot \bar{x}$

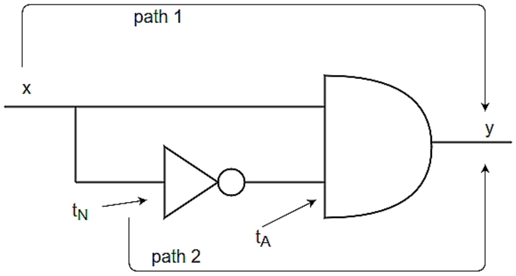


Fig.1.Hazard generating circuit – case 1

Delays: → path 1- t_A / associated $\tau_A=\tau(t_A-t)$; non inverting
→ path 2- $t_N+t_A=t_{N+A}$ / associated $\tau_{N+A}=\tau(t_{N+A}-t)$; inverting

$$y(t) = (x \oplus \tau_A) \cdot (\bar{x} \oplus \tau_{N+A}) = (x\overline{\tau_A} + \bar{x}\tau_A)(\bar{x}\overline{\tau_{N+A}} + x\tau_{N+A}) = x\overline{\tau_A}\tau_{N+A} + \bar{x}\tau_A\overline{\tau_{N+A}} \qquad (5.1.)$$

Considering that $t_A<t_N + t_A=t_{N+A}$ according to table 1, we know that $\overline{\tau_A} \cdot \tau_{N+A} = 0$ and $\tau_A \cdot \overline{\tau_{N+A}} = \delta_{A,N+A}$, thus rendering the above equation as follows:

$$y(t) = \bar{x} \cdot \delta_{A,N+A} \qquad (5.2.)$$

Eq. 5.2. provides us with the static "0" hazard pattern.

One can easily observe that x=0 → $y(t) = \delta_{A,N+A}$ and x=1 → $y(t)=0$, meaning that for transition x "0"↑"1", the circuit's output will exhibit a pulse valued "1" beginning at t_A and ending t_{N+A}, while for transition x "1"↓"0", the circuit's output will maintain a constant "0" value.

5.2. Static - Case 2

Algebraic description → $y = x + \bar{x}$

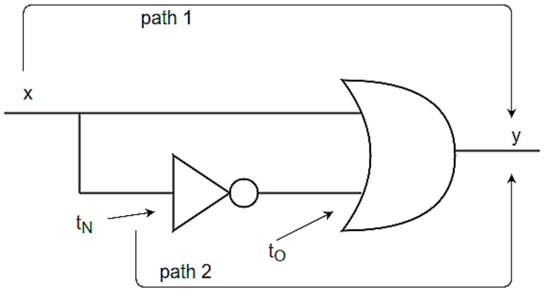


Fig.2.Hazard generating circuit – case 2

Delays: →path 1- t_O / associated $\tau_O=\tau(t_O-t)$; non inverting
→path 2- $t_N + t_O = t_{N+O}$ / associated $\tau_{N+O}=\tau(t_{N+O}-t)$; inverting

$$y(t) = (x \oplus \tau_O) + (\bar{x} + \tau_{N+O}) = x\overline{\tau_O} + \bar{x}\tau_O + x\tau_{N+O} + \bar{x}\overline{\tau_{N+O}} = x(\overline{\tau_O} + \tau_{N+O}) + \bar{x}(\tau_O + \overline{\tau_{N+O}}) \qquad (5.3.)$$

Considering that $t_O<t_N+t_O=t_{N+O}$ according to table 1, we know that $\overline{\tau_O}+\tau_{N+O}=\overline{\delta_{O,N+O}}$ and $\overline{\tau_O}+\tau_{N+O}=1$, thus rendering the above equation as follows:

$$y(t)=\bar{x}+x\cdot\overline{\delta_{O,N+O}} \tag{5.4.}$$

Eq. 5.4. provides us with the static "1" hazard pattern.

One can easily observe that x=0→*y(t)=1* and x=1→ $y(t)=\overline{\delta_{O,N+O}}$, meaning that for transition x "1"↓"0", the circuit's output will exhibit a pulse valued "0" beginning at t_O and ending at t_{N+O}, while for transition x "0"↑"1", the circuit's output will maintain a constant "1" value.

5.3. Dynamic - Case 3

Algebraic description → $y=x+x\cdot\bar{x}$

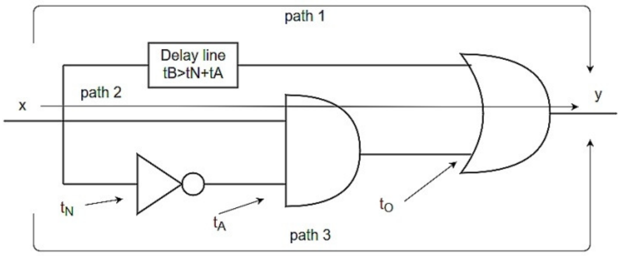


Fig.3.Hazard generating circuit – case 3

Delays: →path1– t_B+t_O / associated $\tau_{B+O}=\tau(t_{B+O}$ -t); non inverting
→path2- $t_A+t_O=t_{A+O}$ / associated $\tau_{A+O}=\tau(t_{A+O}$-t); non inverting
→path3-$t_N+t_A+t_O=t_{N+A+O}$/associated$\tau_{N+A+O}=\tau(t_{N+A+O}$-t); inverting

$$y(t)=(x\oplus\tau_{B+O})+(x\oplus\tau_{A+O})\cdot(\bar{x}+\tau_{N+A+O})=(x\overline{\tau_{B+O}}+\bar{x}\tau_{B+O})+(x\overline{\tau_{A+O}}+\bar{x}\tau_{A+O})\cdot(x\tau_{N+A+O}+\overline{\bar{x}\tau_{N+A+O}})=$$
$$=x(\overline{\tau_{B+O}}+\overline{\tau_{A+O}}\tau_{N+A+O})+\bar{x}(\tau_{B+O}+\tau_{A+O}\overline{\tau_{N+A+O}}) \tag{5.5.}$$

Considering that $t_{A+O}<t_N+t_A+t_O=t_{N+A+O}$ according to table 1, we know that $\overline{\tau_{A+O}}\tau_{N+A+O}=0$ and $\tau_{A+O}\overline{\tau_{N+A+O}}=\delta_{A+O,N+A+O}$, rendering the above equation as follows:

$$y(t)=x\overline{\tau_{B+O}}+\bar{x}(\tau_{B+O}+\delta_{A+O,N+A+O}) \tag{5.6.}$$

Eq. 5.6. provides us with the dynamic "1" hazard pattern. One can easily observe that:

x="1"→ $y(t)=\overline{\tau_{B+O}}$, meaning that for transition x"1"↓"0", the circuit's output will switch to "1"↓"0" after t_{B+O}, and no further transitions will occur afterwards (normal output transition – no hazard present)

x="0"→ $y(t)=\tau_{B+O}+\delta_{A+O,N+A+O}$, meaning that for transition x"0"↑"1" and considering that $t_B>t_{N+A}$ (see the figure above), the circuit's output will present a

- "0" until t_{A+O},
- "1" from t_{A+O} to t_{N+A+O} (term $\delta_{A+O,N+A+O}$),
- "0" between t_{N+A+O} → t_{B+O}, and
- "1" after t_{B+O}(term τ_{B+O}).

Please observe that if t_B≤tN+A, the output will switch to “1” after t_{B+O} (term τ_{B+O})≤t_{N+A+O} (term $\delta_{A+O,N+A+O}$), rendering the hazard invisible (masked but not non-existent).

5.4. Dynamic - Case 4 Algebraic description → $y = x \cdot (x + \overline{x})$

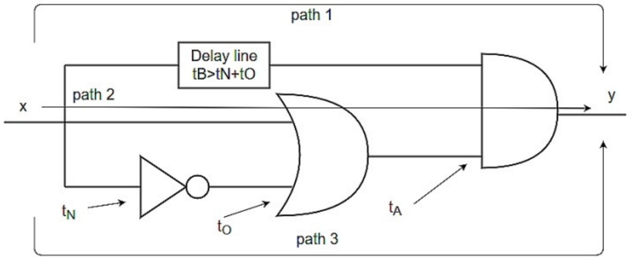


Fig.4.Hazard generating circuit – case 4

Delays: →path 1– t_B+t_A / associated $\tau_{B+A}=\tau(t_{B+A}$ -t); non inverting
→path 2- $t_O+t_A=t_{O+A}$ / associated $\tau_{O+A}=\tau(t_{O+A}$-t); non inverting
→path3-$t_N+t_O+t_A=t_{N+O+A}$/associated $\tau_{N+O+A}=\tau(t_{N+O+A}$-t); inverting

$$y(t) = (x \oplus \tau_{B+O}) \cdot [(x \oplus \tau_{O+A}) + (\overline{x} \oplus \tau_{N+O+A}) = (x\overline{\tau_{B+O}} + \overline{x}\tau_{B+O}) \cdot (x\overline{\tau_{O+A}} + \overline{x}\tau_{O+A} + x\tau_{N+O+A} + \overline{x}\overline{\tau_{N+O+A}}) =$$

$$= x\overline{\tau_{B+O}}(\overline{\tau_{O+A}} + \tau_{N+O+A}) + \overline{x}\tau_{B+O}(\tau_{O+A} + \overline{\tau_{N+O+A}}) \qquad (5.7.)$$

Considering that $t_O+t_A=t_{O+A}<t_N+t_O+t_A=t_{N+O+A}$ according to table 1, we know that $\overline{\tau_{O+A}} + \tau_{N+O+A} = \overline{\delta_{O+A,N+O+A}}$ and $\tau_{O+A} + \overline{\tau_{N+O+A}} = 1$, thus rendering the above equation as follows:

$$y(t) = x \cdot \overline{\tau_{B+O}} \cdot \overline{\delta_{O+A,N+O+A}} + \overline{x} \cdot \tau_{B+O} \qquad (5.8.)$$

Eq. 5.8. provides us with the dynamic “1” hazard pattern. One can easily observe that:

x=”0”→$y(t)=\tau_{B+O}$, meaning that for transition x “0”↑”1”, the circuit’s output will switch to “0”↑”1” after t_{B+O}, and no further transitions will occur afterwards (normal output transition – no hazard present)

x=”1”→ $y(t) = x \cdot \overline{\tau_{B+O}} \cdot \overline{\delta_{O+A,N+O+A}}$, meaning that for transition x “1”↓”0” and considering that $t_B>t_{N+A}$ (see fig. 5.4.), the circuit’s output will present a

- “1” until t_{O+A},
- “0” from t_{O+A} to t_{N+O+A} (term $\overline{\delta_{O+A,N+O+A}}$),
- “1” between t_{N+O+A} → t_{B+O},
- “0” after t_{B+O} (term τ_{B+O}).

Please observe that if $t_B \le t_{N+A}$, the output will switch to “0” after t_{B+O}(term $\overline{\tau_{B+O}}$)≤t_{N+A+O} (term $\overline{\delta_{O+A,N+O+A}}$), rendering the hazard invisible (masked but not non-existent)

As a conclusion, we can state the following:

- Let $t_1,t_2,\ldots,t_n$, be n timestamps respecting $t_1< t_2<\ldots< t_n$,

- Let there be a logic circuit implementing a logic function $f(x,a_0,a_1,\ldots,a_{m-2})$, where $\{x,a_0,a_1,\ldots, a_{m-2}\}$ is the logic function's m-component input vector and x is the input variable to be analysed.
- Let us assume that input variable x is characterized by n distinct in→out paths, therefore generating an n-component secondary input vector $\{x_1, x_2,\ldots, x_n\}$ with associated specific propagation delays of $\{t_1, t_2, t_3,\ldots, t_n\}$
- Rewrite the secondary logic function's equation:

$$F= f(x_1, x_2,\ldots, x_n, a_0, a_1,\ldots, a_{m-2}) \qquad (5.9.)$$

with respect to TDLV (τ,δ) using $x_k=x\oplus\tau_k$, for $k\in\{1,2,\ldots, n\}$

$$F= f(x\oplus\tau_1, x\oplus\tau_2,\ldots, x\oplus\tau_n, a_0, a_1,\ldots, a_{m-2}) \qquad (5.10.)$$

and reduce its expression according to (τ,δ) properties:

$$F= f(x, \tau_1, \tau_2,\ldots, \tau_n, \delta_{1,2},\ldots, \delta_{i-1,i},\ldots, \delta_{n-1,n}, a_0, a_1,\ldots, a_{m-2})$$

where x is the initial value for input variable x.

By determining a specific combination for the remainder of the input vector $\{a_0, a_1,\ldots, a_{m-2}\}$ that would render the logic function's expression identical to one of the patterns presented above, we can state that hazard exists and we can specify the moment of the time when it occurs.

Assuming that $t_{init}<t_\alpha<t_\beta<t_\omega<t_{final}$, the logic function's expression is reduced to:

- $F=\overline{x}\cdot\delta_{\alpha,\beta}$ - static "0" beginning at t_α and ending at t_β
- $F=\overline{x}+x\cdot\overline{\delta_{\alpha,\beta}}$ - static "1" beginning at t_α and ending at t_β
- $F=x\cdot\overline{\tau_\varpi}+\overline{x}\cdot(\tau_\varpi+\delta_{\alpha,\beta})$ - dynamic "0" begins at t_α and ends at t_ω
- $F=x\cdot\overline{\tau_\varpi}\cdot\overline{\delta_{\alpha,\beta}}+\overline{x}\cdot\tau_\varpi$ - dynamic "1" begins at t_α and ends at t_ω

6. ANALYSIS EXAMPLE

6.1. Analysis example 1:

$$f(a,b,c)=a\cdot(b+c)+\overline{a}\cdot\overline{b}\cdot c$$

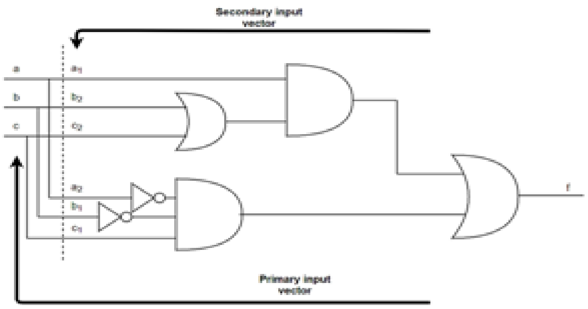


Fig.5. Analysis example – circuit 1

Each component of the primary input vector is characterized by two distinct in→out paths. However, analysis will be performed only for variables a and b. Variable c, although characterized by two paths, does not exhibit a complementary relationship between them, so it will not generate a hazard (E.B.EICHELBERGER [6]). We have computed for each path its specific propagation delay time and defined the secondary input vector by associating with each path a unique secondary variable derived from a primary one (as presented in fig.5.). For presentation purposes we've considered the circuit to be implemented using classic logic gates. The method works the same for any technology chosen to implement the circuit analysed.

Table. 7 CIRCUIT 1-PATH DELAYS FOR VARIABLE a and b

Path delay	FAST Schottky	LS TTL
t_{a1}	13.2 ns	37 ns
t_{a2}	18.2 ns	52 ns
t_{b1}	18.2 ns	52 ns
t_{b2}	19.8 ns	59 ns

Please note that $t_{a1} < t_{a2}$ and $t_{b1} < t_{b2}$.

6.1.1. Analysis for input (a) transition

Function's expression considering the secondary inputs becomes:

$$f(a_1, a_2, b, c) = a_1 \cdot (b + c) + \overline{a_2} \cdot \overline{b} \cdot c \qquad (6.1.)$$

Function's expression with respect to TDLV (τ,δ) is:

$$\begin{aligned} f(a, \tau_{a1}, \tau_{a2}, b, c) &= (a \oplus \tau_{a1}) \cdot (b + c) + \overline{(a \oplus \tau_{a2})} \cdot \overline{b} \cdot c = \\ &= a \cdot [(b + c) \cdot \overline{\tau_{a1}} + \overline{b} \cdot c \cdot \tau_{a2}] + \overline{a} \cdot [(b + c) \cdot \tau_{a1} + \overline{b} \cdot c \cdot \overline{\tau_{a2}}] \end{aligned} \qquad (6.2.)$$

Now, we can perform the analysis on variable a by assuming values for the *(b+c)* and $\overline{b} \cdot c$ terms:

- $(b + c) = \overline{b} \cdot c = 0$ → f(a,τ_{a1},τ_{a2},b,c)=0; NO output anomalies present (output is a constant "0")
- $(b + c) = 0, \overline{b} \cdot c = 1$ → NO solution for the logic equations system → this situation will never occur
- $(b + c) = 1, \overline{b} \cdot c = 0 \rightarrow f(a, \tau_{a1}, \tau_{a2}, b, c) = a \cdot \overline{\tau_{a1}} + \overline{a} \cdot \tau_{a2} \rightarrow$ both transitions for a will provide a singular transition on the output (τ_{a1} respectively $\overline{\tau_{a1}}$) → NO output anomalies present

$(b + c) = \overline{b} \cdot c = 1 \rightarrow f(a, \tau_{a1}, \tau_{a2}, b, c) = a \cdot \left(\overline{\tau_{a1}} + \tau_{a2}\right) + \overline{a} \cdot \left(\tau_{a1} + \overline{\tau_{a2}}\right) \rightarrow$ Considering $t_{a1}<t_{a2}$ and the (τ,δ) properties, we know that $\overline{\tau_{a1}} + \tau_{a2} = \overline{\delta_{1,2}}$ and $\tau_{a1} + \overline{\tau_{a2}} = 1$, thus rendering the function's equation to be:

$$f = a \cdot \overline{\delta_{1,2}} + \overline{a} \qquad (6.3.)$$

Eq. 6.3 presents a pattern identifying a static “1” hazard for input transition **a** “1”↓”0” and b=0, c=1. By placing the time origin at the moment variable a switches, the output will evolve as presented in fig.6.:

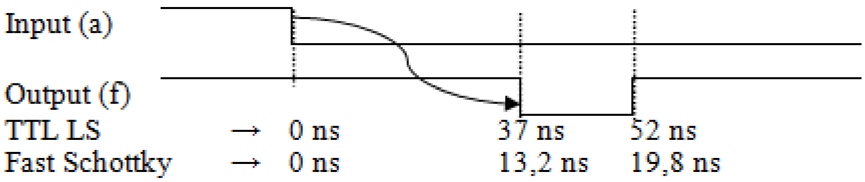

Input (a)
Output (f)
TTL LS → 0 ns 37 ns 52 ns
Fast Schottky → 0 ns 13.2 ns 19.8 ns

Fig.6. Output waveform for input a transition “1”↓”0”

The analysis for input b follows a similar path

6.2. Analysis example 2:

$$f(a,b,c) = a\cdot(b+c)\cdot\overline{(a+c)\cdot b}$$

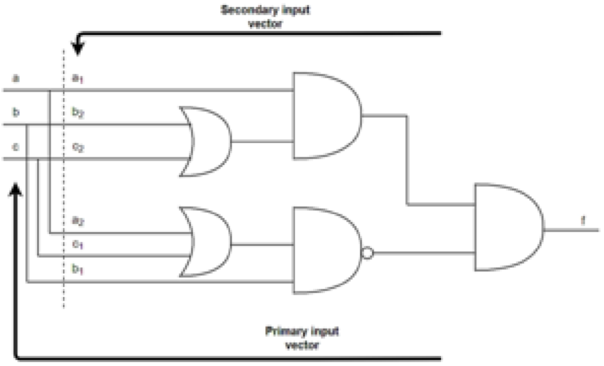


Fig.7. Analysis example – circuit 2

Each component of the primary input vector is characterized by two distinct in → out paths.We have computed a specific propagation delay time for each path and defined the secondary input vector by associating with each path a unique secondary variable derived from a primary one (as presented in fig.7.).

Table. VIII CIRCUIT 2-PATH DELAYS FOR VARIABLES **a,b,c**

Path delay	FAST Schottky	LS TTL
t_{a1}	13.2 ns	30 ns
t_{a2}	18.2 ns	52 ns
t_{b1}	18.2 ns	30 ns
t_{b2}	19.8 ns	52 ns
t_{c1}	19.2 ns	52 ns
t_{c2}	19.8 ns	52 ns

Please note that $t_{a1}<t_{a2}$ and $t_{b1}<t_{b2}$. However, we encounter a most interesting situation when we consider variable c. In the case of LS TTL implementation, both paths are characterized by the same propagation delay time, meaning that although this input variable respects all conditions that would make it a candidate for hazard analysis, we do not need to analyse the circuit’s behaviour when variable c switches because any complementary output switch should occur at the same moment of time, therefore cancelling each other. In the case of FAST Schottky implementation, tc1 < tc2, so analysis should be performed.

6.2.1. Analysis for input (c) transition

Function's expression considering the secondary inputs becomes:

$$f(a,b,c_1,c_2) = a \cdot (b + c_2) \cdot \overline{(a + c_1) \cdot b} \qquad (6.4.)$$

Function's expression with respect to TDLV (τ,δ) is:

$$f(a,b,c,\tau_{c1},\tau_{c2}) = a \cdot (b + c \oplus \tau_{c2}) \cdot \overline{(a + c \oplus \tau_{c1}) \cdot b} == a \cdot \overline{b} \cdot (c \oplus \tau_{c2}) =$$
$$= c \cdot a \cdot \overline{b} \cdot \overline{\tau_{c2}} + \overline{c} \cdot a \cdot \overline{b} \cdot \tau_{c2} \qquad (6.5.)$$

In this case, only τ_{c2} is present in the function's expression, meaning that only one delay will be visible at the circuit output (t_{c2}) if the proper conditions are met (a=1, b=0)

→ No anomalous behaviour possible →

variable c switch will not generate an output anomalous behaviour under any circumstances.

6.2.2. Analysis for input (a) transition

Function's expression with respect to the secondary input vector is:

$$f(a_1,a_2,b,c) = a_1 \cdot (b + c) \cdot \overline{(a_2 + c) \cdot b} = a_1 \cdot \overline{a_2} \cdot b \cdot \overline{c} + a_1 \cdot \overline{b} \cdot c \qquad (6.6.)$$

Function's expression with respect to TDLV (τ,δ) is:

$$f(a,\tau_{a1},\tau_{a2},b,c) = (a \oplus \tau_{a1}) \cdot \overline{(a \oplus \tau_{a2})} \cdot b \cdot \overline{c} + (a \oplus \tau_{a1}) \cdot \overline{b} \cdot c =$$
$$= a \cdot \left(b \cdot \overline{c} \cdot \overline{\tau_{a1}} \cdot \tau_{a2} + \overline{b} \cdot c \cdot \overline{\tau_{a1}}\right) + \overline{a} \cdot \left(b \cdot \overline{c} \cdot \tau_{a1} \cdot \overline{\tau_{a2}} + \overline{b} \cdot c \cdot \tau_{a1}\right) \qquad (6.7.)$$

Now, we can perform the analysis on variable a by assuming values for $b \cdot \overline{c}$ and $\overline{b} \cdot c$ terms:

- $b \cdot \overline{c} = \overline{b} \cdot c = 1$ → NO solution for the logic equations system
- $b \cdot \overline{c} = \overline{b} \cdot c = 0$ → f(a,τ_{a1},τ_{a2},b,c)=0; NO output anomalies
- $b \cdot \overline{c} = 0, \overline{b} \cdot c = 1$ → both transitions for a will provide a singular transition on the output (τ_{a1} respectively $\overline{\tau_{a1}}$) → NO output anomalies present
- $b \cdot \overline{c} = 1, \overline{b} \cdot c = 0$ → Considering t_{a1}<t_{a2} and the (τ,δ) properties, we know that $\overline{\tau_{a1}} \cdot \tau_{a2} = 0$ and $\tau_{a1} \cdot \overline{\tau_{a2}} = \delta_{1,2}$, thus rendering the function's equation to be:

$$f = \overline{a} \cdot \delta_{1,2} \qquad (6.8.)$$

Eq. 6.8. presents a pattern identifying a static "0" hazard for input transition **a** "0"↑"1" and b=1, c=0. By placing the time origin at the moment variable a switches, the output will evolve as presented in fig.8.:

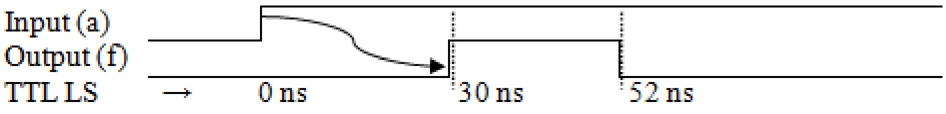


Fig.8. Output waveform for input a transition "0"↑"1"

7. CONCLUSION

Both approaches presented will either improve or ease the use of the classic methods while maintaining their reliability. Please bear in mind that it is not our intention to state that the classic techniques are failing, but we simply wish to demonstrate that the same results may be reached using less computation, and if we use the second approach, the outcome will provide more information as far as timing is concerned.

The first approach presented may be used when the behavior of a combinational logic circuit needs to be analyzed. One can easily determine if the CLC's output may present a hazard simply by identifying specific terms in its expression. At this point, we may choose not to use that circuit or perform structural changes (if possible) to mask the hazardous behavior. However, choosing the second approach (masking anomalous behavior) comes with a cost, as masking is performed by entering redundant terms into the logic function equation or by using dummy gates to equalize the different path propagation delays. That means an increase in the implementation cost and a decrease in speed.

The second approach, if used to analyze an already implemented CLC, will provide detailed information on the output's behavior with respect to the time axis. Having this information available, we shall be able to design a hierarchically superior circuit that uses the present circuit's outputs in such a manner that the time frames, when the circuit output is malfunctioning, are not to be considered.

The second approach also presents us with a possible development path in the area of asynchronous automata. It is well known that the proper design and operation of these devices is dependent on strict timing specifications. The proposed approach is to design the automaton as a synchronous machine, use the methodology presented to map the outputs of the input group of functions CLC (next state CLC – Mealy or Moore) and define and design a variable time pulse generator to be used as a synchronizing signal (instead of a fixed parameter clock signal) that will use the earliest mapped moment of time to change the state of the automation. Thus, the automaton will not be an asynchronous one but rather a pseudo synchronous one. The advantages would be the ease of design and less susceptibility to timing issues while retaining most of the advantages of the asynchronous structure, such as high speed and fast response.

REFERENCES

[1] Beister J.: "A unified approach to combinational hazards", IEEE Trans.Comp.VolC-23, pp. 566-575, 1974, 10.1109/T-C.1974.223996

[2] McCluskey E. J.:"Logic Design Principles",Prentice-Hall, Englewood Cliffs, NJ, 1986.

[3] Tinder Richard F.: "Engineering Digital Design", Second Edition, Elsevier - ACADEMIC PRESS, 2001

[4] Bryant R.:"Graph-based Algorithms for Boolean Function Manipulation," IEEE Trans.Comp., 677-691 (1986), 10.1109/TC.1986.1676819

[5] Akers S.: "Binary Decision Diagrams," IEEE Trans.Comp., C-27, 509-516 (1978). DOI:10.1109/TC.1978.1675141

[6] Nowick S.M.,O'Donnell C.W.:"On the existence of hazard-free multi-level logic" Proceedings / IEEE ASYNC 03, May12-16,2003, DOI: 10.1109/ASYNC.2003.1199171

[7] Jeong C., Nowick S.M.: "Fast hazard detection in combinational circuits" ACM DAC 04, June 7-11, 2004, 10.1109/DAC.2004.240453

[8] Berthomieu B., Diaz M.:"Modeling and Verification of Time Dependent Systems using Time Petri Nets", IEEE Transactions on Software Engineering 17,259-273, 1991DOI: 10.1109/32.75415

[9] J.A. Brzozowski and C.J.H. Seger: "Advances in Asynchronous circuit theory Part II - Bounded Inertial Delay Model, MOS Circuits, Design Techniques", EATCS Bulletin 43, 199-263, 1991

[10] J. Brzozowski, B. Li, Y. Ye: "On the Complexity of the Evaluation of Transient Extensions of Boolean Functions", 12th International Workshop on Descriptional Complexity of Formal Systems, DCFS 2010,

[11] Y. Ye., J. Brzozowski: "Covering of Transient Simulation of Feedback-Free Circuits by Binary Analysis",Int. J.Found. Comput. Sci.17,949-973, 2006

[12] M. Gheorghiu, J. Brzozowski:"Simulation of Feedback-Free Circuits in the Algebra of Transients"Internat. J.Found. Comput.Sc.,14,1033-1054, 2003.

[13] Oded Maler, Amir Pnueli: "Timing Analysis of Asynchronous Circuits using Timed Automata", Correct Hardware Design and Verification Methods, Volume 987, 1995,pp189-205,Springer2005,DOI: 10.1007/3-540-60385-9_12

[14] Alan R. Martello, Steven P. Levitan: "Temporal analysis of time bounded digital systems", Correct Hardware Design and Verification Methods, Volume 683, Springer 1993, pp 27-38, DOI: 10.1007/BFb0021712

[15] Asarin A., Maler O. and Pnueli A.: "Symbolic Synthesis of Discrete and Timed Systems" in A. Nerode (Ed), Hybrid Systems II, Springer LNCS, 1995, DOI: 10.1.1.43.5633 1995

[16] William K.C. Lam, Robert K. Brayton: "Timed Boolean Functions – A unified formalism for exact timing analysis", Kluwer Academic Publishers, 1994, ISBN 0-7923-9454-2

[17] Galupa N.: "Increase of Sequential Systems Performance Using Digital Hazard Analysis", ICCS/ISITA '92, 1096-1100, ISBN: 0-7803-0803-4, 10.1109/ICCS.1992.255092

[18] Galupa N.: "TIME/LOGIC VARIABLES USED FOR DIGITAL HAZARD SEARCH,"Proceedings of IEEE CCECE 2008, DOI: 10.1109/CCECE.2008.4564553

[19] Stevens K.S., Ginosar R., and Rotem S.: Relative timing [asynchronous design], in IEEE Transactions on VLSI Systems, pp. 129-140, ISSN 1063-8210, 2002, 1.1109/TVLSI.2002.801606

[20] R. Ben Salah, Bozga M. and Maler O.: "On Timing Analysis of Combinational Circuits," In FORMATS'03, LNCS 2791, pages 204-219. Springer (2003)

[21] R. Ben Salah, Bozga M. and Maler O.: "On Timed Components and their Abstraction," In SAVCBS'07 Workshop, ACM ISBN 978-1-59593-721-6/07/0009 (2007)

[22] Riedel M.D., Bruck J.: "Timing Analysis of Cyclic Combinational Circuits," Technical Report, Parallel and Distributed Systems Group, CaltechPARADISE:2004.ETR060, 25 Feb 2014

AUTHORS

Nicolae Galupa (B.Sc and M.Sc. Hons 1988, PhD 2003) received his BSc and MSc degree from Technical University Iasi, Faculty of Automatic Control and Computer Engineering, Iasi, Romania in 1988 after a 5 years training programme and PhD degree from the same University in 2003 after a 6 years training programme.

He has been with the Faculty of Automatic Control and Computer Engineering, Technical University Iasi, Iasi, Romania since 1990. Also since 2006 he joined the digital circuits design team of DICODE Ltd. and focused on timing issues in digital circuits.

He joined IEEE in 2006 and has been a member of IEEE Region 8 ever since.

DIGITAL INTERACTIVE STORYTELLING APPROACHES: A SYSTEMATIC REVIEW

Islam Sharaha[1] and Amal AL Dweik[2]

[1]Master of Informatics
islamsharaha@gmail.com
[2]Dept. of Computer Engineering and Sciences
Palestine Polytechnic University (PPU) Hebron, Palestine
amal@ppu.edu

ABSTRACT

Interactive Digital Storytelling (IDS) is concerned with the creation of a new media art form that allows real-time interaction with a developing narratives. IDS is important learning, training, testing and entertainment tool. This paper makes a systematic review that compares several approaches used in (IDs) in terms of user interaction type, degree of interaction importance, classification of approaches types, and comparing approaches in terms of some performance factors.

KEYWORDS

Digital Interactive Story Telling, time- interaction, degree of interaction.

1. INTRODUCTION AND THEORETICAL BACKGROUND

Storytelling is the process of creating narrative structures or engaging with them, which is pervasive in many aspects of children's life. It has different definitions such as: "Storytelling is one of the oldest methods of communication and learning" [Tharrenos et al, 2015]. As [Nuri et al, 2013], the storytelling is an important way to share experiences, thoughts, and imaginations between people in term of verbal statements. In a child's world, the storytelling is a great tool to reflect children's feelings. A storytelling is a good way for learning about identity and communication as it enables the exploration of one's inner world [Benjamin, 1998]

Interactive Digital Storytelling (IDS) is concerned with the creation of a new media art form that allows real-time interaction with a developing narrative [Stefan Rank et al, 2012]. [Lathem SA, 2005] defined Digital storytelling as a combination of traditional, oral narration with different types of multimedia (like: image, text, video and music) with communication tools. As [Benjamin, 1998], there are three functions that narrative should serve, and must be carefully analyzed to produce a good story: cognitive, social and emotional function.

There are three levels of the story creation. The first level is the storyline or plot; which is a series of chronology and causally related events that make up the story's content. Storyline can be a

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 21–30, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61002

script-based or a character-based. The second level is the narrative, which is a representation of the plot from a particular point of view. The third level is the presentation, which is a realization of the story in a particular medium [Mariat et al, 2002]. Good story line should adhere of Fraytiys triangle [Jeroen, 2012].

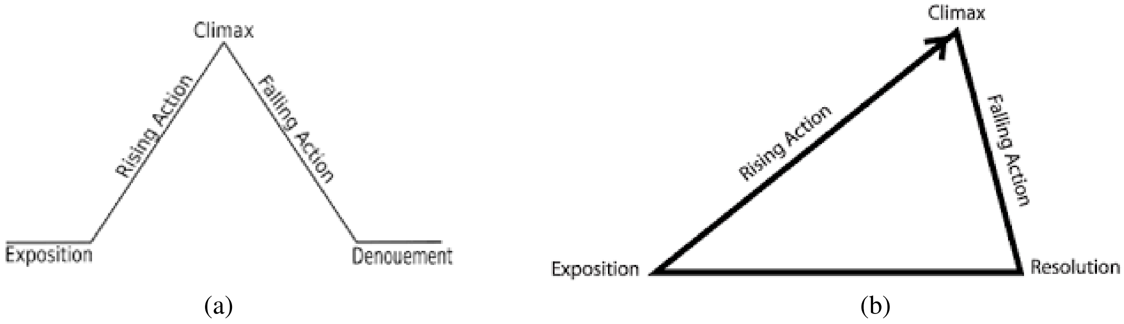


Figure 1: (a) Standard Fraytiys triangle. (b) Another form of Fraytiys triangle.

As shown in figure 1, rising action describes the events that occur and actions that are undertaken previous to the story's climax. The climax features the highest peak in dramatic tension. Thereafter, there is a falling action. Lastly, denouement addresses the resolution of the conflict and the final release of tension.

As [Edrilei, 2014], the story telling systems can follow three basic approaches: plot-based, character-based and hybrid approach.

The rest of the paper is organized as follows: the next section introduced the literature review. A discussion of storytelling approaches in terms of types, user interactivity types, approaches tools, implementation and testing, and performance factors are presented next. In the final section, the conclusion and future work is introduced.

2. RELATED WORK

Several researchers interested in the interactive storytelling field. Some of them considered the interactive storytelling as a good entertainment tool as [Edirlie, 2014]. This is accomplished by allowing the user easily to interact with the system through the "paper and pencil" approach. Other researchers used interactive stories as an educational tool, such as [Nuri, 2013]. They proposed StoryTech, which is a smart storytelling toy that features a virtual space and a real space. [Raidle, 2007] proposed an approach that combines believable agents and intelligent scenario direction. This was used for social and cultural training, which consider the use of the storytelling techniques as a training tool. [M. Seif Al-Nasr, 2013] made questioner on narrative interactive to determine the user background in interactive storytelling. The user interpretations, emotions and behavioral response to Façade are analyzed. Where Façade, which is one of the most well-known interactive storytelling systems, depends on the drama manager that manages the narrative progression. It does so by trying to module the narrative so that it corresponds to a desired story arc such as the one of Freytag's triangle

Marc Gavazza and others described the planning techniques to control autonomous characters in order to make interaction with virtual character [M Gavvzza et al, 2005].

Rafael Perez described a computer model for plot generation based on emotion and tension between characters, implemented using MEXICA which is a computer model of a cognitive account of creative writing purposed by Rafael Perez in 2001 [P.Y Perez, 2007].

Edirilie Soares presented paper and pencil approach as a storytelling system that is based on augmented reality and used SVM to recognize the user sketches [Edirlie et al, 2014].

Nuri Kara introduced StoryTech, which is a smart storytelling toy that features a virtual space that includes computer based graphics and characters, and a real space, which includes plush toys, background cards, and a communication interfaces based on mixed reality [Nuri et al, 2013].

Mariet et al used agent techniques to produce a virtual storyteller, where the storyline created by the action of characters is guided by director agent. Yundong et al used agent technology to present DIRACT, which is an approach to create characters that do not make a difference between director or actors' characters [Yundong et al, 2010], [Mariat et al, 2002].

3. DISCUSSION

The following section discusses the different digital story telling approaches.

3.1. Interactive Storytelling Approaches Types.

As shown in table 1, the digital story telling approaches are classified into three types: Character-based approaches, Plot-based approaches and Hybrid approaches.

Table.1 interactive storytelling approaches.

Author	**Approach Title**	**Storytelling approach type**	**Application based app**
[Yundong et al, 2010]	DIRACT	Character-based	Real-time
[Edirlie et al, 2014]	Paper and pencil	Hybrid approach	Real-time
[Polbo et al, 2005]	CBR Plot Generation	Character-based	Natural language recognition
[P.Y Perez, 2007]	MEXICA	Character-based	Natural language recognition
[Nuri et al, 2013]	StoryTech	Plot-based	Real-time

The approaches were classified based on the story derivative way. In the Character- based approaches, the story development depends on character decision. The main disadvantage of this system is that it is less adhere of Fraytyis triangle. In Plot-based approach, the characters have no autonomy and they are less consistence in the scene since the characters are often interchangeable. The Hybrid approaches are used to bridge the gap between the plot-based approach and character-based approach.

3.2 User Interactivity Types

As [Linssen, 2012] Interactivity means having at least some control over the narratives. From this definition, we can say that interactivity has different degrees. As shown in the table 2, the user interaction types are divided into three levels or degrees: limited, medium and high. By limited interaction, we mean that the user influences little parts in the story or storytelling level. Where in medium interaction, the user influences a whole level of the storytelling. High interaction means that the user can influence all the story levels or parts.

The user interaction in most approaches compared in table 2 is limited or low in the approaches that are used or meant by speech recognition. This is due to the difficulties in the recognitions of different languages and child's speech. Where the interaction is high in real time approaches, such as: [Edirlie et al, 2014],[Yundong et al, 2010] and [Nuri et al, 2013].

Table 2. Comparison between interactive storytelling approaches in terms of user interaction.

Author	**Approach title**	**User interaction type**	**Degree of importance**
[Y.- G Cheongh et al ,2008]	Framework for authoring interactive narrative	Participants (Speech)	Limited
		Author (authoring tool)	Medium
[Han YU et al, 2008]	Goal-oriented system	Character creation (as agents)	Medium
[M Riadle et al, 2003]	Automated Scenario Director	Speech to be changed	Limited
[M Gavazza et al, 2004]	Interacting with Virtual Characters	Physical interaction Speech	Medium
[M Gavazza et al, 2005]	Dialogue Generation in Character-based Interactive Storytelling	(Future work) Embodying the user as one of virtual characters	High
[Edirlie et al, 2014]	Paper and pencil	Drawing on paper	High: real-time
[M.O Raidle et al, 2007]	Interactive narrative system	As one of virtual characters	High

[P.Y Perez, 2007]	MEXICA	Give the initial state	Limited
[C.B Callaway, 2002]	Narrative prose generation	Give the story request	Limited
[Nuri et al, 2013]	StoryTech	Put objects on the receive panel	High: real-time
[U Spierling, 2002]	Setting the scene	Give the story requests	Limited
[David et al, 2009]	Learning to Influence Emotional Responses	Answering some given question to drive the narrative	Limited
[Yundong et al, 2010]	DIRACT	As a director virtual character (agent)	High: real-time
[Mariat et al, 2002]	Virtual story teller	Create characters and give priority	Medium
[Polbo et al, 2005]	Plot generation based on CBR	Query for a new story from old others	Medium

3.3 Tools Classification

The interactive storytelling approach can benefit user in many ways. The interactive storytelling used techniques can be classified as a tool of authoring, education, entertainment, and training. The classified techniques can be further classified into several story level influences, as shown in table. 3.

Table 3. Comparison between interactive storytelling approaches in term of technique used and how to benefit the user.

Author	**The approach used**	**Used technique**	**Tool**	**Story level influence**
[Y.- G Cheongh et al ,2008]	Framework for authoring interactive narrative	Branching graphs and AI planning	Authoring (Storyline)	Storyline (Script-based)
[Han YU et al, 2008]	Goal-oriented system	Multi agent system	Authoring(Character creation)	Presentation

[Edirlie et al, 2014]	Paper and pencil	Augmented reality, svm	Entertainment tool	Presentation
[Nuri et al, 2013]	RFID interactive panel	Mixed reality	Entertaining, educational and measuring tool	Presentation
[David et al, 2009]	Influence emotional responses	YouTube video And pre-authoring text	Tool to derive narrative by player (one time)	Narrative
[Yundong et al, 2010]	DIRACT	Multi agent system (inheritance)	Authoring (character creation)	Presentation
[Mariat et al, 2002]	Virtual story teller	Multi agent system (intelligent agent)	Authoring (automatic story line)	Storyline (Character-based)
[M.O Raidle et al, 2007]	Believable agents	Combine narrative control, believable character agents and drama manager	Authoring and training tool	Storyline (Character-based)
[Polbo et al, 2005]	Plot generation based on CBR	Case-Based Reasoning CBR, Natural Language Generation NLG	Authoring tool	Storyline (Character-based)

As shown in table 3, the interactive storytelling techniques can be used as a very good authoring tool and can help the user with low experience in authoring to create their own stories such as those used in [Yundong et al, 2010], [M.O Raidle et al, 2007] and [Mariat et al, 2002]. From table 3, you can notice that most of interactive storytelling authoring tool approaches are using agent techniques; that is the behavioral and emotional agent designed to serve the storytelling attracting the users and giving them more chances and abilities to produce a good story. On the other hand, the approaches that used a simple way to interact at most real-time approaches are considered as a good entertainment tool.

3.4. Interactive storytelling approaches implementation and testing.

Some of the approaches used or suggested to make interactive storytelling are implemented and tested. Others were either implemented or just tested. As shown in table 4.

Table 4. Implemented and tested approaches.

Author	Approach title	Implemented	Tested
[Edirlie et al, 2014]	Paper and Pencil	Yes	Yes (for effectiveness and satisfaction)
[Han YU et al, 2008]	Goal-oriented system	Yes	Yes (usability)
[Polbo et al, 2005]	Plot generation based on CBR	No	No
[P.Y Perez, 2007]	MEXICA	Yes	Yes (interestingness, novelty, predictability)
[Yundong et al, 2010]	DIRACT	No	Technical test (usability)

[Edirlie et al, 2014] mentioned that their approach was implemented and tested by questionnaire contains 54 questions derived from the IRIS evaluation Toolkit, and the participants were high school students. Where [P.Y Perez, 2007], made a questionnaire using a story developed in MEXICA and MINSTREL in order to compare MEXICA and MINSTREL in terms of interestingness, novelty and predictability.[Yundong et al, 2010] made a case study to evaluate the approach usability.

3.5 Comparison of performance factors.

Table 5, summarized the surveyed approaches according to the performance factors: speed, accuracy, usability and reuse. Speed means the time needed to respond to user interaction. Accuracy is the ability to satisfy user purpose. Usability means the easiest to use the approach. Reuse is the ability of the use of elements from old stories to create a new story.

Table 5: comparison between approach in term of speed, accuracy and user interface

Author	**Approach title**	**Speed**	**Accuracy**	**Usability**	**Reuse**
[Edirlie et al, 2014]	Paper and pencil	Good (one interaction per minute)	83% because of some limitations in recognition algorithm	Satisfaction user interface because of the easy way to interact with the system	No

[P.Y Perez, 2004]	MEXICA	----	More than MINSTREL	------	Yes
	MINSTREL	-----	Less than MEXICA	-------	Yes
[Han YU et al, 2008]	Goal-oriented	-----	Good	Friendly user interface saves development time and cost	No
[Polbo et al, 2005]	CBR plot generation	Not implemented	Not implemented	Not implemented	Yes
[Yundong et al, 2010]	DIRACT	Not implemented	Not implemented	Ease to use	Yes

[P.Y Perez, 2004], made a comparison between MEXICA and MINSTREL, where those two approaches were the most farmhouse at that time. He mentioned that the stories produced by MEXICA were more interesting, but some time they were poorly written. As we can see in the table, paper and pencil approach seems to be good in usability. However, for the accuracy, there were some limitations because of recognition algorithm, or some time in user sketches. Approaches in [Polbo et al, 2005] and [Yundong et al, 2010] was not implemented, but it was mentioned that these approaches allowing the reuse by their construction.

4. CONCLUSION

Storytelling is the oldest way to communicate, learn, entertain and share experiences and thoughts among people. Furthermore, it is an effective way to reflect the feeling and people social background. Interactive Digital Storytelling (IDS) is the way to share stories over the world. There are many approaches used in IDS. They are classified into three basic groups: character-base, plot-based and hybrid approach, depending on the way that they follow to build the story. There are several approaches in terms of user interaction types and tools; where it is found that the most important are: low, medium, and high levels. Some of the approaches were implemented and tested, while others were not. The paper compares different approaches in terms of accuracy, speed, reuse, and user satisfaction.

As a conclusion, we can conclude that the real time approaches were the best for entertainment; because of the high interaction with it. Where the most appropriate approach for authoring and training are better to be implemented using agent technology. In terms of usability, the approaches with friendly user interface, had the better satisfaction.

REFERENCES

[1] Perez Y Perez, "Employing emotions to drive plot generation in a computer-based storyteller", cognitive system research 8 (2007) 89-109.

[2] Edirilie Soares de Lima, et al, "Draw your own story: paper and pencil interactive storytelling", Entertainment Computing, 5 (2014) 33-41.

[3] Mark O. Riedle, Andrew Stern, "Believable agents and intelligent story adaptation for interactive storytelling", 2007.

[4] Marc Gavvaza, Fred Charles, "Dialogue generation in character-based interactive storytelling", American Association for Artificial Intelligence, 2005.

[5] Han Yu, et al, "A goal-oriented development tool to automate the incorporation of intelligent agents into interactive digital media applications", Theoretical and Practical Computer Applications in Entertainment, Vol. 6, No. 2, 2008.

[6] Mark O. Riedle, Andrew Stern, "Believable Agents and Intelligent Scenario Direction for Social and Cultural Leadership Training", 2007.

[7] Marc Cavazza, et al, "Interacting with Virtual Characters in Interactive Storytelling ", AAMAS'02, July 15-19, Bologna, Italy, 2002.

[8] Magy Seif El-Nasr, et al, "Experiencing interactive narrative: A qualitative analysis of Facad ", Entertainment Computing 4 (2013) 39-52.

[9] Pablo Gervas, et al, "Story plot Generation based on CBR", Knowledge-Based Systems 18 (2005) 235-242.

[10] yun-Gyung Gheong, et al, "A framework for authoring Interactive narrative" LNCS 5334, 297-308, 2008.

[11] Mark Riedl, et al, "Managing Interaction Between users and Agents in a Multi-agent Storytelling Environment", AAMAS'03, 2003.

[12] Charless B. CHallaway, et al, "Narrative prose generation", Artificial Intelligence 139 (2002)203-252.

[13] Ulrike Spierling, et al, "Setting the scene: playing digital director in interactive storytelling and creation", Computers & Graphics 26 (2002) 31–44.

[14] Nuri Kara, et al, "Investigating the Activities of Children toward a Smart Storytelling Toy", Educational Technology & Society, 16 (1), 28–43, 2013.

[15] YundongCai, et al, "DIRACT: Agent-based Interactive Storytelling", IEEE/ WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010.

[16] DavidL.Robert, et al, "Learning to Influence Emotional Responses for Interactive Storytelling", Association for the Advancement of Artificial Intelligence, 2009.

[17] Mariët Theune, et al, "The Virtual Storyteller: story creation by intelligent agents", 2002.

[18] MarkO.Riedl, "Interactive Narrative: A Novel Application of Artificial Intelligence for Computer Games", Association for the Advancement of Artificial Intelligence, 2012.

[19] Jeroen Linssen, "A Discussion of Interactive Storytelling Techniques for Use in a Serious Game", 2012.

[20] Tharrenos Bratitsis, et al, "From early childhood to special education: Interactive digital storytelling as a coaching approach for fostering social empathy", Procedia Computer Science 67 (2015) 231 – 240.

SECURITY SYSTEM FOR DATA USING STEGANOGRAPHY AND CRYPTOGRAPHY (SSDSC)

Ayman Wazwaz[1], Khalil Abu-haltam[2], Sawan Atawneh[2], Areej Idaes[2], Dalal Salem[2]

[1]Electrical Engineering Department, Palestine Polytechnic University
aymanw@ppu.edu
[2]Electronics and Communications Engineering

ABSTRACT

Security System for Data using Steganography and Cryptography (SSDSC) is a set of hardware and software components that will be used to send secured documents through the internet. Some of the software will be loaded into a microcontrollers in order to increase the complexity and security. The data will be encrypted using the Advanced Encryption Standard (AES) algorithm with a key from the Raspberry PI microcontroller and hide it inside an image using Least Significant Bit (LSB) algorithm, the data will be invisible. The image will be transmitted and received through the internet, the receivers will extract the hidden data from the image and decrypt it to have the original data with the image.

Complicating the steps of hiding and encryption will reduce the possiblity of intrusin of secured documents, and the process will be trasparent to the user to increase security without affecting the normal steps and the behavior in secured documents exchange.

KEYWORDS

Steganography, Cryptography, LSB algorithm, AES algorithm..

1. INTRODUCTION

Steganography is the art and science of concealing communication. The goal of steganography is to hide the existence of information exchange by embedding messages into unsuspicious digital media covers [1]. Cryptography, or secret writing, is the study of the methods of encryption, decryption, and their use in communications protocols [2]. Both techniques manipulate data to ensure the security of information, but the concept of steganography differs from cryptography. Cryptography obscures the meaning of a message, but it does not conceal the fact that there is a message. The goal of cryptography is to make data unreadable by a third party, whereas the goal of steganography is to hide the data from a third party. Both techniques have an ancient origin, but the modern field is relatively new. Cryptography and steganography are fundamental components of computer security [1]. In the current trends of the world, the technologies have advanced so much that most of the individuals prefer using the internet as the primary medium to

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 31–38, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61003

transfer data from one end to another across the world. Data security basically means protection of data from unauthorized users or hackers and providing high security to prevent data modification.

In this paper, we present a system that is able to encrypt documents, hide it in an image and send it in one side, and then retrieve it in the other side. It uses a graphical interface to facilitate the use of the system and have the ability to transmit and receive through the public internet while keeping documents secured.

2. PROBLEM STATEMENT

The SSDSC can be used in many applications across the world, such applications will include the general secondary exams in Palestine, it will reduce the burden of carrying and distributing exams to hundreds of schools, other beneficiaries of the SSDSC are Banks, between branches and between the bank and its clients. No matter how large or small your company is, you need to ensure the security of your information assets, so you can use this system to protect your information. Other applications of SSDSC is online elections, internet banking, medical-imaging and others.

3. RESEARCH METHODOLOGY & SOLUTION

The SSDSC system consists of software and hardware. The software part will be loaded into a microcontroller attached to the computer at both sides in order to increase its complexity, and to secure the process of extracting data and decryption.

Raspberry Pi is small single-board Linux computer that is used to execute codes instead of computers in some environments. Here, it is used to store keys to be used in data exchange, and codes to implement the encryption and digital steganography.

Figure1 shows the system components and processes that consists of five parts namely: Text, Image, PC, Microcontroller (Raspberry Pi) and internet.

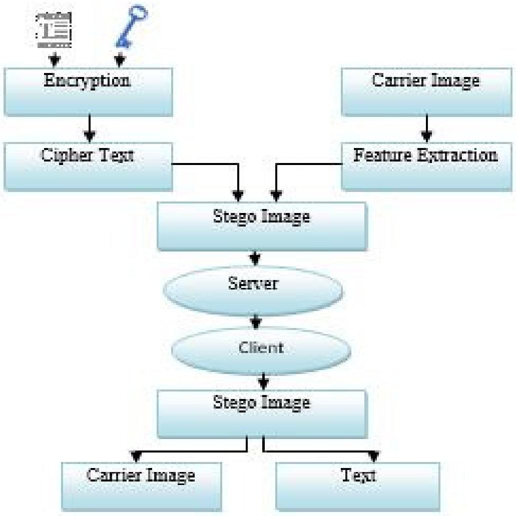


Figure1: SSDSC System Components

The system works according the following steps:

- First step, Insert the text and image to the application.
- Secondly, the text will be encrypted and then hidden within the image.
- Thirdly, connect the microcontroller with the computer, and then the microcontroller work on the text and apply encryption.
- Fourthly, upload the image that contains the information encoded on the server.
- Fifthly, the client access to the server and download the images that contain encoded information.
- Finally, the computer will work on the extraction of encrypted data of the image, then break this encryption to have the original data (file).

4. SOFTWARE AND HARDWARE IMPLEMENTATION

The data will be encrypted on the Raspberry Pi using AES algorithm. The AES stand for Advanced Encryption Standard which is substitution/permutation network cipher. It take 128-bit plaintext and 128, 196 or 256 bit key, depending on the number of rounds [2].

Advanced Encryption Standard, also known as Rijndael is a symmetric-key block cipher published by the National Institute of Standards and Technology (NIST) [2].

The criteria is defined by NIST for selecting AES falls into three areas: Security, Cost, and Implementation. The AES algorithm is used in this system to encrypt the data which is called the cipher data. The key is 256 bits size, it is generated on the Raspberry Pi.

The cipher data will be hidden in image. The most important method will be lagged the hiding is the implementation of the feature extraction of the images in which algorithms are used to detect and isolate various desired portions or shapes (features) of an image. The selected portion is not important in scene. It is particularly important in the area of recognition. The algorithm is Object detection in a cluttered scene using point feature matching. This process is implemented using Matlab code.

A popular digital steganography technique is so-called least significant bit (LSB).Least significant bit (LSB) insertion is a common, simple approach to embedding information in a cover image [3]. The least significant bit (the 8th bit) of some or all of the bytes inside an image is changed to a bit of the secret message. When using a 24-bit image, a bit of each of the red, green and blue color components can be used, since they are each represented by a byte. In other words, one can store 3 bits in each pixel. An 800 × 600 pixel image, can thus store a total amount of 1,440,000 bits or 180,000 bytes of embedded data [4].

The object detection algorithm used to extract the features of the image. It is detect the object which is not important in scene .Then the cipher data will be hidden using LSB algorithm in this object.

The system will read the image and convert it into grayscale. The features will be extracted using object detection in a cluttered scene using point feature matching algorithm to determine the parts which is not important in scene. If the capacity of these parts is suite for hiding the cipher data in it, the system will choose the pixel and hide the data. The cipher data will be divided into sub data when the capacity is not enough and testing the condition again. The data will be hidden in the image which is called stego-image .The changes in color of the bits will not able to be notice from the users.

Virtual Private Network is a client server application based on tunneling protocols that are used in making private connection based on public infrastructure like the internet. Here, it is assumed that the sender and the receivers use the internet as an infrastructure, and VPN software is installed to assure privacy.

5. RESULTS

Here, the results of the steps explained earlier to extract the features of the image; object detection algorithm code detected a specific object (Engineering building in Palestine Polytechnic University campus). The ciphered data is hidden in this image.

Figures 2,3,4 and 5 are the results of selecting image and hiding the secured data in specific parts according the features of the selected images.

Figure 2 shows the reference image containing the object of interest (Engineering building) and the target image containing a cluttered scene (image of a different objectives).

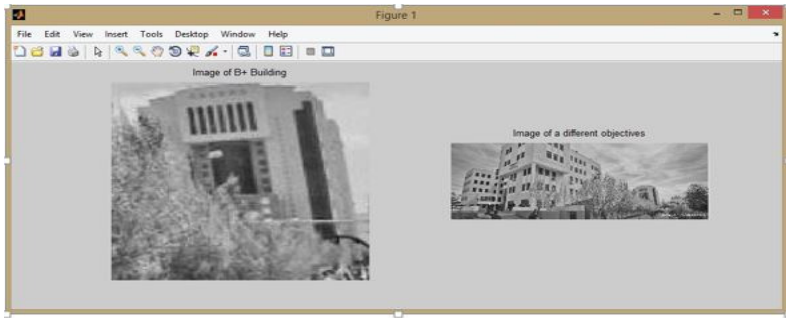


Figure 1
File Edit View Insert Tools Desktop Window Help
Image of B+ Building
Image of a different objectives

Figure 2 The reference image and the target image

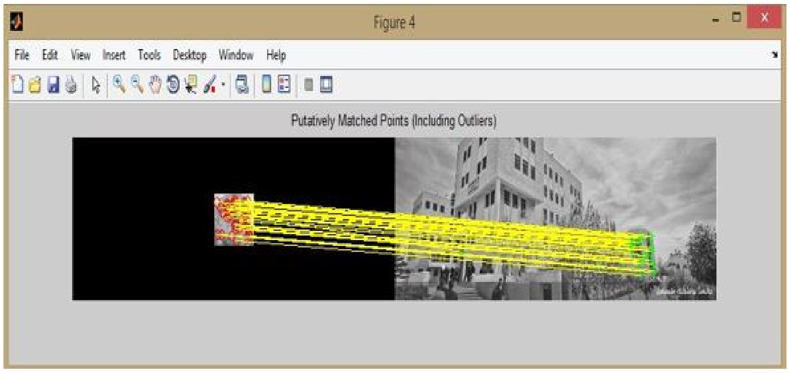


Figure 4
File Edit View Insert Tools Desktop Window Help
Putatively Matched Points (Including Outliers)

Figure 3: Display matched features

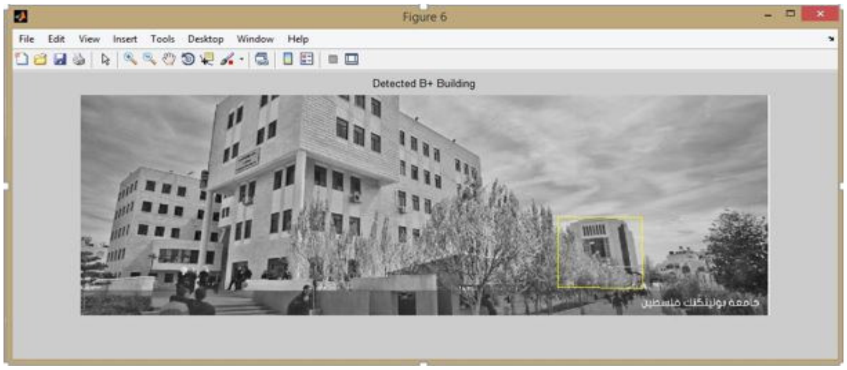


Figure 4: Display the detected object.

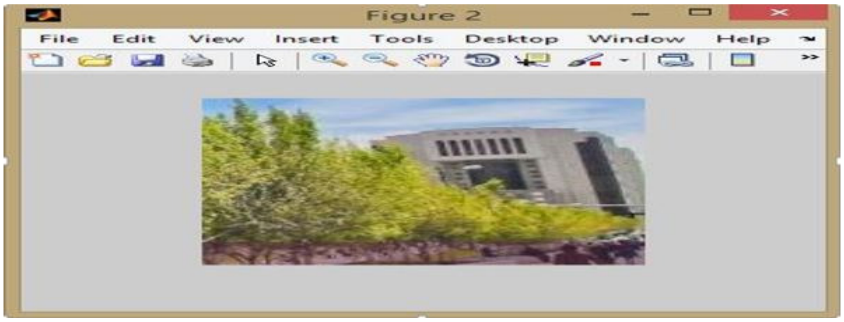
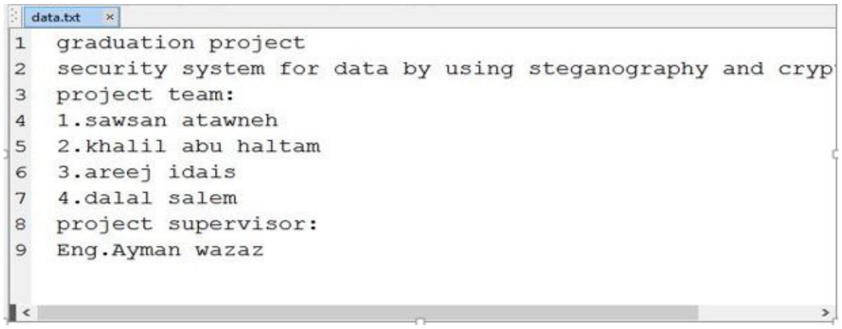


Figure 5: Detected object image which is used to hide the data

After implementing the LSB algorithm, we obtained the following results:



```
graduation project
security system for data by using steganography and cryp
project team:
1.sawsan atawneh
2.khalil abu haltam
3.areej idais
4.dalal salem
project supervisor:
Eng.Ayman wazaz
```

Figure 6: The hidden text.

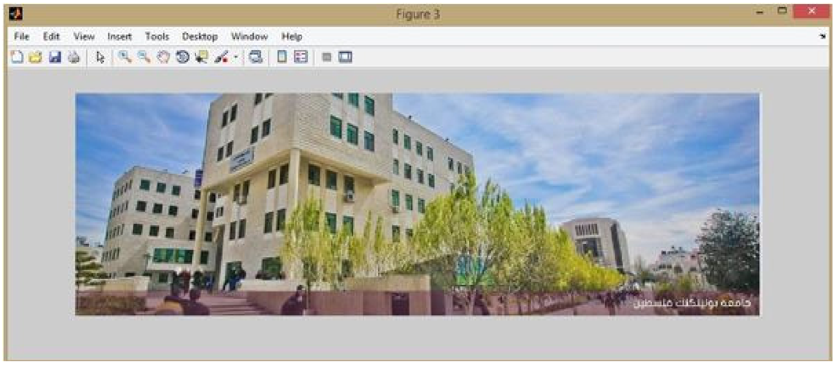


Figure 7 : Stego-image.

6. TESTS

When the system hided the data, the images before and after the hiding is approximately the same. The difference between the images locate the pixels position of the hidden data.

In image comparison, figures contains the first image (carrier image), the second image (stego-image), and the histogram for the both and result of the test.

A histogram is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable) in an image processing context, the histogram of an image normally refers to a histogram of the pixel intensity values. This histogram is a graph showing the number of pixels in an image at each different intensity value found in that image [5].

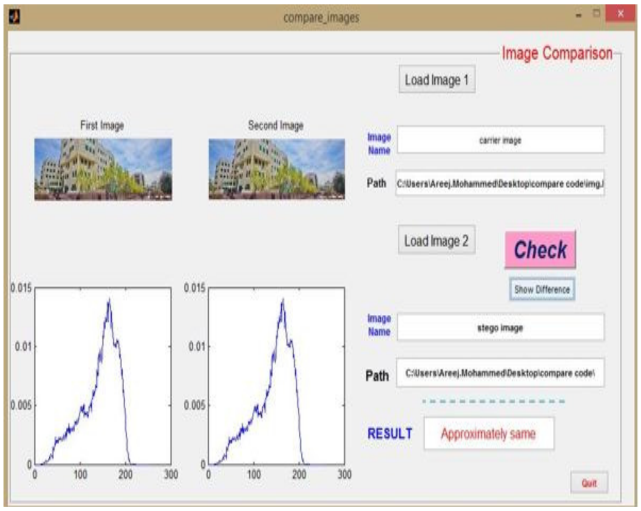


Image Name: carrier image

Path: C:\Users\Areej.Mohammed\Desktop\compare code\img.

Image Name: stego image

Path: C:\Users\Areej.Mohammed\Desktop\compare code\

RESULT: Approximately same

Figure 8: The result of the comparison between the carrier image and stego-image which is approximately the same.

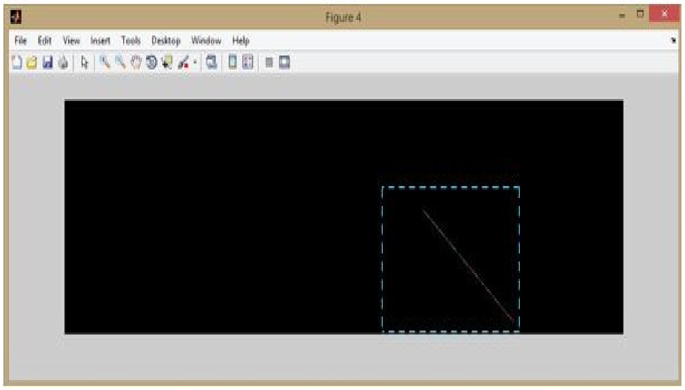
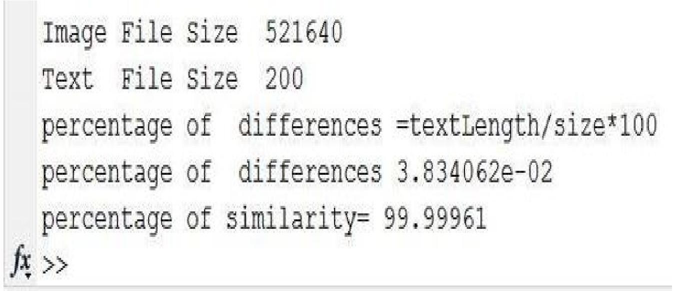


Figure 9: The difference between the carrier image and stego-image.



```
Image File Size  521640
Text  File Size  200
percentage of  differences =textLength/size*100
percentage of  differences 3.834062e-02
percentage of similarity= 99.99961
fx >>
```

Figure 10: The percentage of similarity and differences between the two images.

Figure 9 shows a small sequence of dots resulted from the difference between the two images, and figure 10 shows the number of hidden bytes compared to the image size, and the percentage of similarity between the two images.

A website will be used by the users to upload and download the files. The administrator of the webserver need to login to the website using a username and a password. The files will be added or removed by the administrator who is responsible for the distribution of these files to the other authenticated users. Other users will use the same website to login in and download the documents. Figure 11 shows the login page on the web server.

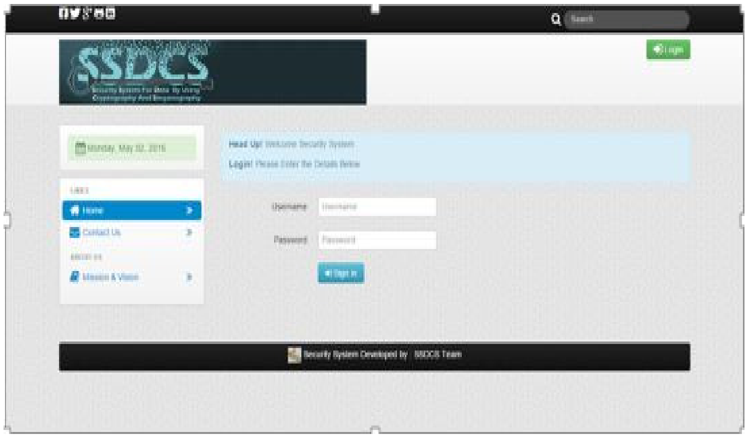

Search
Login
SSDCS
Security System For Data By Using Cryptography And Steganography
Monday, May 02, 2016
Head Up! Welcome Security System
Login! Please Enter the Details Below
Home
Contact Us
Mission & Vision
Username
Password
Sign In
Security System Developed by SSDCS Team

Figure 11: Access to the web server

The security in this system is integrated from different protocols and services:

- Encryption with keys using AES.
- Steganography using images
- Hardware security using microcontrollers.
- Signing in using username and password.
- Virtual Private Network (VPN) to open a private connection using the internet as a public service.

7. CONCLUSION

The “Security System for Data using Steganography and Cryptography” has been designed and tested. The system designed to encrypt the data using AES algorithm, extract the features of the image to detect the places that are suitable to hide the cipher data in. The data will be extracted from the image and retrieve the original data and image.

Adding hardware components made the system more secured because the encryption keys are distributed with hardware itself, this added value will make it difficult for intruders to have a copy of this hardware including the software embedded in the system.

REFERENCES

[1] D.C. Lou, J.L. Liu, H.K. Tso, Evolution of information – hiding technology, in H. Nemati (Ed.), Premier Reference Source–Information Security and Ethics: Concepts, Methodologies, Tools and Applications, New York: Information Science Reference, 2008, pp. 438-450.

[2] Carl H. Meyer and Stephen M. Matyas, Cryptography: A New Dimension in Computer Data Security, John Wiley & Sons, New York, 1982.

[3] P.C. Wu, W.H. Tsai, A stenographic method for images by pixel-value differencing, Pattern Recognition Letters 24 (2003) 1613-1626.

[4] R. Ibrahim and T.S. Kuan, Steganography imaging system (SIS): hiding secret message inside an image, Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2010, San Francisco, USA, 2010, pp. 144-148.

[5] Freedman, David; Diaconis, P. (1981). "On the histogram as a density or: L2 theory”.Zeitschrift furWahrscheinlichkeitstheorie und verwandte Gebiete 57 (4): 453–476.Doi:10.1007/BF01025868

REMOTE NETWORK LABORATORY TO SUPPORT MULTI-LEVEL NETWORK PRACTICE

Ninghan Zheng, Chengbin Quan, Xiaojun Wu, Shanshan Li,
Yongqiang Chen

Department of Computer Science and Technology,
Tsinghua University Beijing, 100084, China
zhengnh@tsinghua.edu.cn, quancb@tsinghua.edu.cn,
xjwu@tsinghua.edu.cn,
lishanshan@tsinghua.edu.cn,chenyongqiang@tsinghua.edu.cn

ABSTRACT

The Computer Network Laboratory of Computer Science department in Tsinghua University provides experimental environment for undergraduate computer network serial courses offered by CS Dept. In order to meet the demand of different elective students, we combed experimental teaching system of computer network serial courses, then designed and implemented a remote network laboratory supported multi-level network practice. We mainly deployed three series of experimental system: Virtual Network Experiment System based on commercial network equipment, Network Protocols Experiment System named NetRiver and IPv4/IPv6 Transition Technologies Experiment System based on 4over6 tunneling technology. With the help of our network laboratory, the following goals can be achieved: 1) Remote. Students can access experimental environment from classroom, dormitory and etc.; 2) Multi-level. Experimental contents include both network principle and commercial realization, include both classical sliding window protocol, IPv4 or IPv6 data transceiver protocol, IPv4 or IPv6 data forwarding protocol, routing protocol and up-to-date IPv4/IPv6 Transition Technologies. Every network course can construct experimental content by combination of them; 3) Up-to-date. We introduced the latest technology in the field of computer network research, such as IPv6 network protocol and 4over6 tunneling technologies.

KEYWORDS

remote laboratory, laboratory education, experimental equipment, computer network

1. INTRODUCTION

The Computer Network Laboratory is an important part of the Computer Teaching Experiment Center in Computer Science department of Tsinghua University. The Lab provides experimental environment for undergraduate computer network serial courses offered by CS Dept. At present, the development of the laboratory mainly considers the following two aspects: Multi-level network practice and Innovative experimental method.

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 39–53, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61004

The consideration of the multi-level network practice comes from the fact that the lab serves different courses which have different target students and different teaching goals. We mainly divided the courses into three categories: Computer Network Curriculum as selective course for all the non-computer major students; Computer Network Principles Curriculum as compulsory course for the computer major students and Special Training in Computer Network Curriculum as limited optional course for the computer major students. The experimental teaching objectives of the three types of curriculums are different from each other, simultaneously they have common parts. So in order to achieve the goal of supporting different courses, we can deploy or implement serials of computer network experimental equipment and design variety of candidate experimental projects which can be effectively selected and combined.

Innovative experimental methods meet the demands of the tremendous development of electronic and computer techniques. The development of Internet technology, especially the development of the Cloud Computing technology, provides a strong support for the innovation of experimental methods. The experimental combination of the PCs and special experimental equipment in the past was gradually replaced by more flexible experimental methods. The remote experimental platforms based on C/S mode or B/S mode are more popular with students. With the help of them, students can flexibly arrange experiment time and perform the experiment in classroom, dormitory, library and everywhere can access Internet.

Based on the above two points, we combed the knowledge of different computer network courses, designed and implemented three kinds of network experimental platforms to support multi-level network practice. These experimental platforms support remote experiment and be scalable to support the different number of elective students.

In this paper, we will introduce our attempt in trying to improve the computer network laboratory education in Computer Science department of Tsinghua University. We will first discuss the teaching system of computer network curriculums in section 2 secondly proposed our principles of designing computer network laboratory platforms in section 3, finally present the implementation of computer network laboratory education in Tsinghua University in section 4, and multi-level network practice supported by the principles will be presented simultaneously.

2. TEACHING SYSTEM OF COMPUTER NETWORK EXPERIMENT

Computer network experimental courses is not isolated, must be closely combined with the computer teaching curriculum system. Computer science emphasizes practice, so most of computer courses of CS Dept. in Tsinghua are provided with the corresponding experimental part. Figure 1. provides a brief description on the teaching system of computer experiment.

We used a tree-shaped model to describe the teaching system and divided computer experiment into five parts: Basic Experiment as the trunk which is the necessary prerequisite to other parts; four directions including Computer System Experiment, Software System Experiment, Computer Network Experiment and Application Technologies Experiment as leaves which represent the four secondary disciplines of CS and can be studied simultaneously.

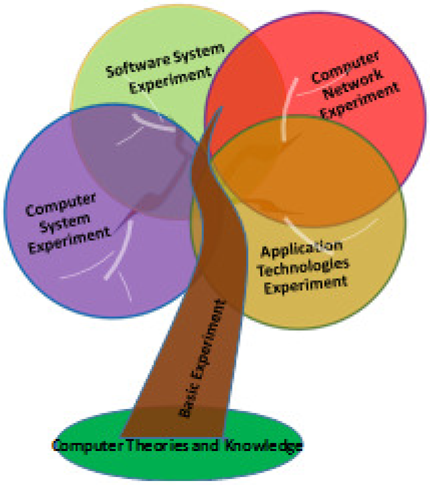

Software System Experiment
Computer Network Experiment
Computer System Experiment
Application Technologies Experiment
Basic Experiment
Computer Theories and Knowledge

Figure 1. Teaching system of computer experiment

The Computer Network has been an independent secondary discipline of CS in 2015 in China because it plays a more and more important role in the economic development of our country. In the past we placed computer network courses into Application Technologies direction. Like MIT [1] and Stanford [2], we set the Computer Network Principles Curriculum as follow-up course of Computer Compose Principle Curriculum. Now we adjust them into parallel courses based on the following understanding:

- With the data structure knowledge and programming skills obtained in Basic Experiment, students can study network principles and protocols well, while Computer Compose Principle Curriculum focus on the principles of one local computer system include processor, memory, and etc.

- As an independent branch of teaching system, only a 48 hours required course is not sufficient. We added Special Training Serial Curriculums in which more deep practice is carried out as follow-up course of main curriculums in different directions such as Special Training in Operating System Curriculum followed Operating System Curriculum, Special Training in Compilation Principles Curriculum followed Compilation Principles Curriculum. Meanwhile we added Specialty Practice Serial Curriculums in junior summer semester, in which students complete a project with integrated knowledge of each direction. So Computer Compose Principle Curriculum belongs to Computer System Experiment and have its own subsequent Special Training Serial Curriculums and Specialty Practice Serial Curriculums.

Now we can draw up curriculum system of computer network experiment in Figure 2. In this figure, Computer Network Curriculum as selective course for all the non-computer major students is not drew up because of its separation from the teaching system of CS.

Computer Network Experiment spent 8 credit laboratory hours is a part of Computer Network Curriculum. This course is target at non-computer major students who are not expected to master thorough computer knowledge and programming skills. The curriculum teaches the principles of computer networks rather than network engineering or network buildup. So observation and analysis on network protocols rather than programming and implementation are expected in the experiment.

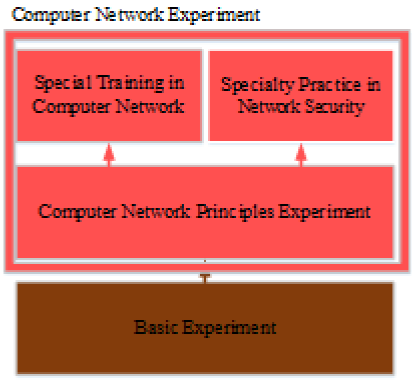

Computer Network Experiment
Special Training in Computer Network
Specialty Practice in Network Security
Computer Network Principles Experiment
Basic Experiment

Figure 2. Curriculum system of computer network experiment

Computer Network Principles Experiment belong to Computer Network Principles Curriculum and worthy of 24 credit laboratory hours. It is a compulsory course for the computer major junior students in fall semester. The curriculum teaches the principles of computer networks with the classics textbook of "Computer Network" edited by Andrew [3]. Elective students with a strong foundation in computer programming are willing to comprehensively study the key network techniques and re-implement the important network protocols by hand.

Special Training in Computer Network followed Computer Network Principles Experiment is a full-time practical training course spend 48 hours. 24 credit laboratory hours spent in Computer Network Principles Experiment are not sufficient to do comprehensive practice in network protocols and cannot accommodate frontier network knowledge and technology. Special Training in Computer Network aims to be the appropriate successor.

Specialty Practice in Network Security is also a full-time practical training course. The 200 hours will be used to complete a project for Network Security. A set of special experimental equipment is dedicated to the course. This is beyond the scope of this paper.

3. PRINCIPLES OF DESIGNING COMPUTER NETWORK LABORATORY PLATFORM

In view of the discussion above, we propose several principles in designing computer network laboratory platforms (Figure 3.):

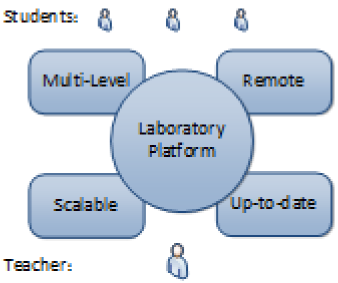


Figure 3. Principles in designing computer network laboratory platform

- Multi-level

We dedicate to design laboratory platform for three main courses: Computer Network Experiment, Computer Network Principles Experiment and Special Training in Computer Network. Computer Network Experiment pay attention to observation and analysis on network protocols. Computer Network Principles Experiment encourage reimplementation the important network protocols. Specialty Practice in Network Security continue to strengthen the former mission and introduce new network technology.

- Remote

Internet technology, especially Cloud Computing technology, create possibilities for remote experiment and cloud-based experiment [9-11]. Students can flexibly arrange experiment time and perform the experiment in classroom, dormitory, library and anywhere can access Internet. We wouldn't consider software-based virtual network devices like [9, 10] because of their unreal results. We'd like to publish real physical network devices to students in Internet, thus they can interact with real equipment.

- Scalable

Capacity of experiment course, especially the elective courses, fluctuates wildly from year to year. The laboratory platforms should be scalable according to student number. As remote experiment resource according to the planning requirements of the last section, it should be easier to be deployed and released.

- Up-to-date

Network technology have been developing rapidly and new technologies continue to appear. Students like to practice on rich and up-to-date network technologies including IPv6 network protocol, IPv4/IPv6 Transition Technologies and wireless network technology. In the original laboratory, these experimental contents could not be supported. So we must continuously design and implement new laboratory platforms to keep pace with the times.

In the past work, there are many kinds of experimental methods:

- Commercial network equipment: Students are curious of how the real network instruments including switches and routers work. Although network buildup which should be included in engineering elective course is not the focal point of our experiment, interacting with them we can analysis the work principles of network protocols through packet capture software.

- Visual learning tools: The tools use images or Flash to display the results of interaction with the simulated network device [4, 5]. They are often used to rapidly verify the correctness of operating and configuration on device and are suitable for novice learning such as Cisco Networking Academy [6]. But we cannot capture the intermediate results of interaction, and then the protocol analysis cannot be carried out.

- Network programing: Operating system provides API for user-end network programing such as Winsock and Linux socket. But the core network protocol stacks are encapsulated in system modules and transparent to user. Students are difficult to operate the network protocols.

- Network simulation tools: There are so many tools of this type: OPNET, J-Sim, NS-2, NS-3, KivaNS and etc. They play a great role in scientific research. But we think they are not suitable to undergraduate experimental teaching no matter use them directly [7] or wrap them with a friendly front end [8]. These tools are not intuitive and require deep knowledge and experience. Students have to spend a lot of time to learn the manuals.

- Virtual machines: Use VMs, we can do network experiment remotely [9, 10]. But also we need guest OS to provide software-based virtual router, firewall and VPN.

4. Implementation of Computer Network Laboratory Education

4.1. Computer Network Laboratory Equipment

Based on the above discussion, we propose three categories of experimental methods to support multi-level experiment:

4.1.1. Virtual Network Experiment System

We deployed cloud-based virtual network experimental platform of Ruijie Corporation [13] to support remote experimental environment based the real network devices. Figure 4. depicts one experimental group with Virtual Network Experiment System and now we deployed 8 groups. Each group consists of a set of real physical network devices include one router, one layer 3 switch and one wireless AP.

Cloud-based virtual network experimental system is a special industrial control computer using virtualization technology which can simultaneously load multiple VM images running Windows or Linux. There are 10 physical connection channels between the system and layer 3 switch. So it can simulate 10 PCs.

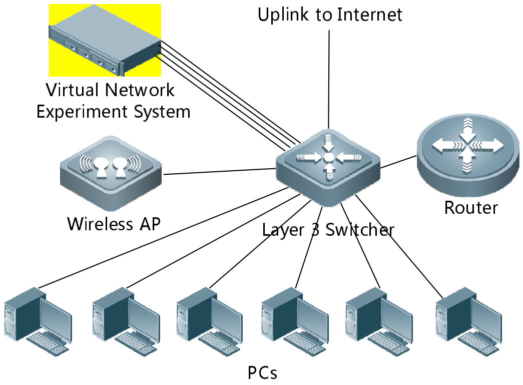


Figure 4. A experimental group with Virtual Network Experiment System

There are two main ways to do experiment: The first way is using PC to do the traditional local experiment. The other way is using virtual terminals produced by virtual network experimental system to do remote experiment. Students remotely access Virtual Network Experiment System through browser and log in the VM OS generated by system. For students, it seems that the real network devices are around them now.

It seems strange in Fig. 4. that the router is out of proper place. In a classic network topology, the router should be laid between the uplink and switch. But when students do router experiment remotely, they will probably lose the link because the router not retain it. So our solution is using the lookback interface as experimental interface instead of uplink interface when do the router experiment.

With the help of packet capture software like WireShark [14], students can observe and analysis explicit working process of the network protocols in our experimental environment. A group package analysis based experiment supported are list in Table 1:

Table 1. Experiment supported by virtual network experiment system

Tag	Experiment Name	Experiment Object
111	FAT AP with single SSID	Configuration on wireless AP with single SSID in FAT mode
112	FAT AP with multiple SSID	Configuration on wireless AP with multiple SSID in FAT mode
113	AP bridge in one vlan	Bridge the same vlan using two APs
114	AP bridge in different vlans	Bridge different vlans using two APs
121	ARP protocol	Analyse principle of ARP protocol
122	802.3 package	Analyse package of 802.3
123	802.1Q package	Analyse package of 802.1Q
124	STP BPDU	Analyse package of Spanning Tree protocol
125	MSTP BPDU	Analyse package of Multiple Spanning Tree protocol
126	Loop effect	Analyse Loop effect without Spanning Tree protocol
127	LLC forward	Analyse package forwarding of ARP protocol
131	Static route	Config and Analyse static route
132	RIP	Config and Analyse RIP protocol
133	OSPF	Config and Analyse OSPF protocol
141	TCP	Analyse TCP protocol
142	UDP	Analyse UDP protocol
151	DHCP	Config and Analyse DHCP protocol
152	DHCP relay	Config and Analyse DHCP relay protocol
153	FTP	Config and Analyse FTP protocol
154	HTTP	Config and Analyse HTTP protocol

“Tag” in this table plays a role in category: The first number distinguishes the platforms: “1” for Virtual Network Experiment System, “2” for Network Protocols Experiment System and “3” for IPv4/IPv6 Transition Technologies Experiment System; the second represents the layer of TCP/IP 5-layer network protocols, for example “2” means data link layer; the last one is sequence number of the same layer experiment in one system.

The pros and cons of the system include: 1) It suits the non-computer major students without strong programming skills to study basic network principles. And also can be a foundation correlation method to the majors when they realization network protocols by programing; 2) It creates possibilities for remote access to real physical network equipment instead of software-based virtual device. Meanwhile traditional local experimental environment reserved; 3) It is scalable by adding the number of groups; 4) With the help of Ruijie wireless AP and 4G switch, wireless experiment is added to the laboratory.

4.1.2. Network Protocols Experiment System (NetRiver)

We designed and implemented the Network Protocols Experiment System named NetRiver to support programing experiment. Early version named NetRiver 2000 was released in 2007 [12]. New version of NetRiver 3000 was released in 2015. The topology is depicted in Figure 5.

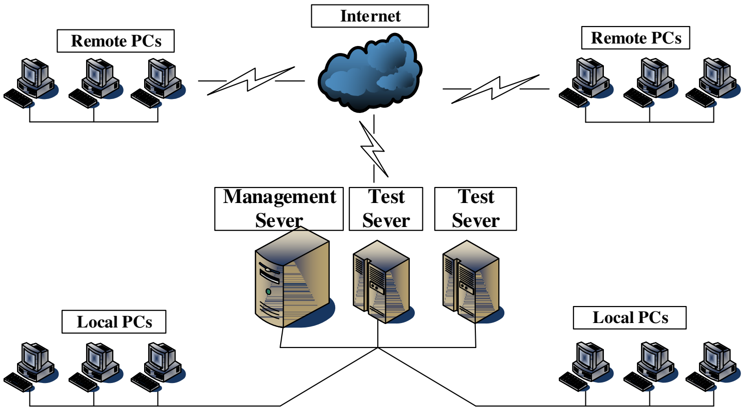


Figure 5. A experimental group with Virtual Network Experiment System

The system consists of a management server, variable number of test server and a user-friendly integrated development software client.

Management server is used to manage the students’ information, experiment content. The results of experiment can be submitted to the server, then the statistical results are generated automatically for instructors. The visual client provides edit, compile, execute and debug environment. Test server acts as the key of the system. The TCP/IP 5-layer network protocols have been realized in it, meantime a serial of APIs are open to students. Students can conveniently focus on specified segments of these network protocols according to experimental requirements, without concerning about the context. After that students can test their tasks with the test server, contrary to which, in a real network environment this type of experimental data packages will be discard as error data.

Students can edit, compile, execute and debug offline in any PC client. They realize the experiment code of network protocol segment, call the APIs provided by the system to constitute the complete protocol stack. Now they can login the management server and get the allocated test server. Test server will verify the task program by interacting with client.

This system uses scalable script description to define experiment content. Teachers can easily add new experiment content to it. Now the network protocols programing-based experiment are list in Table 2.

Table 2. Experiment supported by virtual network experiment system

Tag	Experiment Name	Experiment Object
221	Sliding window protocol	Realize sliding window protocol in Andrew textbook
231	IPv4 sending and receiving	Achieve the function of IPv4 Protocol in host protocol stack
232	IPv6 sending and receiving	Achieve the function of IPv6 Protocol in host protocol stack
233	IPv4 transmission experiment	Achieve the function of IPv4 in router protocol stack
234	IPv6 transmission experiment	Achieve the function of IPv6 in router protocol stack
235	RIP protocol	Achieve RIP protocol in router protocol stack
236	Protocol state machine	Achieve state machine in BGP protocol
237	IPSec	Achieve IPSec protocol
241	TCP	Achieve encapsulation and sending, receiving of TCP, analyzing process of TCP message

The pros and cons of the system include: 1) It is a protocols programing-based system. It helps students to study key network technologies which are hidden in system core of host OS or network devices' OS and difficult to practice. So it can be the key chain of practice for network direction computer major undergraduates; 2) Experimental resources of the system is being released online, students can finish the job remotely; 3) The main performance bottleneck locate in test server. But the number of test servers can also be increased by register in management server. 4) The achievement comes from IPv6 core router finished by Tsinghua, so it maintains the advanced nature of the technology.

4.1.3. IPv4/IPv6 Transition Technologies Experiment System

We developed new experiment system because IPv4/IPv6 Transition Technologies have emerged as hot researches. The work principles of IPv4/IPv6 Transition are also hidden in OS of hosts and network device. Some configuration-based experiment can be carried out with real commercial equipment. But programing-based experiment are seldom touched. Figure 6. describes process of the new system.

Main-body text is to written in fully (left and right) justified 11 pt. Times New Roman font with a 6pt. (paragraph) line spacing following the last line of each paragraph, but a 12pt. (paragraph) line spacing following the last paragraph. Do not indent paragraphs.

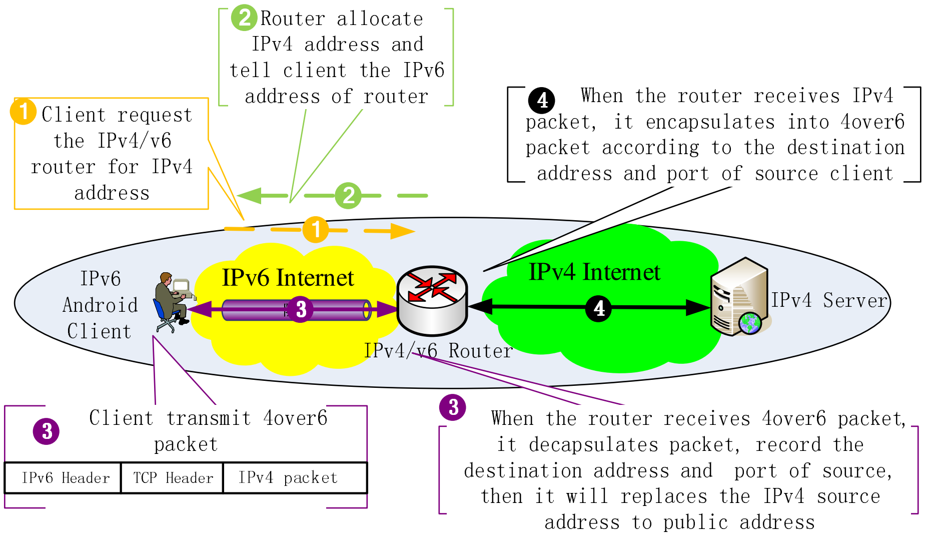

2 Router allocate IPv4 address and tell client the IPv6 address of router
1 Client request the IPv4/v6 router for IPv4 address
4 When the router receives IPv4 packet, it encapsulates into 4over6 packet according to the destination address and port of source client
2
1
IPv6 Android Client
IPv6 Internet
3
IPv4/v6 Router
IPv4 Internet
4
IPv4 Server
3 Client transmit 4over6 packet
IPv6 Header
TCP Header
IPv4 packet
3 When the router receives 4over6 packet, it decapsulates packet, record the destination address and port of source, then it will replaces the IPv4 source address to public address

Figure 6. IPv4/IPv6 Transition Technologies Experiment System

In this scene, a user who has only IPv6 Internet wants to access the IPv4 server in IPv4 Internet. The work-flow will be:

- Client request the IPv4/v6 router for IPv4 address by broadcasting.
- Router responds to the request and allocate IPv4 address to client and tell client the IPv6 address of router.
- Client transmit 4over6 packet.
- When the router receives 4over6 packet, it decapsulates this packet, records the destination address and port of source, then it will replaces the IPv4 source address to public IPv4 Internet address.
- When the router receives IPv4 packet from IPv4 server, it encapsulates into 4over6 packet according to record of the destination address and port of client.

Now we have two parts to work at: IPv6 client and IPv4/IPv6 router. We call the IPv6 client as client of IPv4/IPv6 Transition Technologies Experiment System, IPv4/IPv6 router as the server.

In order to supplement the Android programing skill of undergraduates who have seldom practice on it in other courses. We deliberately user an Android mobile device as the client that a PC is also ok.

So two important experimental projects come out from Figure 6 (List in Table 3):

In the last semester, we only let the students to realize the client. The server was made into an Industrial Personal Computer which is more stable and reliable and deployed in campus network of Tsinghua. So the students can test their client remotely with the server.

Table 3. Experiment supported by ipv4/v6 transition technologies experiment system

Tag	Experiment Name	Experiment Object
331	Client implementation	Design and Implementation of client of IPv4/IPv6 Transition Technologies Experiment System
332	Sever implementation	Design and Implementation server of IPv4/IPv6 Transition Technologies Experiment System

The client running in Android OS can be divided into two parts: front-end and back-end.

Front-end is the UI programing in JAVA. The main functions include:

- Monitor the network and get the IPV6 address of the uplink physical interface.
- Start the back-end.
- Set the timer and refresh UI.
- Show the network information.
- Star the VPN service of Android OS.

Back-end programing in c interacts with the server. The main functions include:

- Access the server.
- Get the IPV4 address of the downlink virtual interface.
- Read/write virtual interface.
- Decapsulate IP packets.
- Communicate with the server by IPv6 socket
- Send keeplive message to the server

The server can also be deployed in Linux system. The main functions include:

- Create IPV6 socket, listen to the communication between the client and IPv4 server.
- Maintain the virtual interface, implement I/O on virtual interface.
- Maintain IPV4 address pool, allocate address to new client.
- Maintain the information table of clients, save the mapping relations between IPv4 address and IPv6 socket.

- Read the data in IPV6 socket from client, handle the control message.
- Implement decapsulation function to data received by IPV6 socket, then write them to the virtual interface.
- Implement encapsulation function to data received by virtual interface, then send them to client.
- Listen to clients and send keeplive message.

The pros and cons of the system include: 1) The programing-based practice on IPv4/IPv6 transition technologies give a great challenge to computer major students. It can be good start to be network professionals; 2) The server of the system can be deployed in Internet, students can implement the client locally and test it with the server remotely; 3) It is very easy to add the new server if the server resource is insufficient; 4) IPv4/IPv6 transition technologies are new and popular.

4.2. Computer Network Laboratory Curriculum

With the support of laboratory equipment above, we can arrange the proper combination of experimental contents for the three Computer Network Laboratory Curriculum. Table 5 list the experimental contents in the latest semester.

Table 5. Experimental arrangement for computer network laboratory curriculum

	Computer Network Experiment		**Computer Network Principles Experiment**		**Special Training in Computer Network**	
Layer	**Tag**	**Time (hour)**	**Tag**	**Time (hour)**	**Tag**	**Time (hour)**
1	111	1			111	1
	114	1			114	1
2	126	0.5	221	8		
	127	0.5				
3	132	1	232	8	132	1
	133	1	233	8	133	1
					234	8
					237	8
					331	28
4	141	0.5				
	142	0.5				
5	151	1				
	153	1				
Amount		**8**		**24**		**48**

Computer Network Experiment face the non-computer major students. Package analysis based experiment are more suitable. So they only get in touch with Virtual Network Experiment System. The experimental project covers all layers of the network protocols, and as many kinds of real network device as possible are introduced to students.

Computer Network Principles Experiment focus on working principles of network protocols. As a computer major student, he'd better to grasp the essentials of relative theoretic curriculum. Programing-based experiment are more suitable. Network Protocols Experiment System (NetRiver) is the best choice. But because of the limited time, only the most critical experimental contents are required.

Special Training in Computer Network is the continuation of Computer Network Principles Experiment. Two aspects are taken into account: More challenging tasks and neglected basic operation skills. For the former, IPv4/IPv6 transition technologies experiment which will take at least 20 hours for a well-skilled student are introduced. For the latter, small-scale network build-up experiment and package analysis based experiment are required. This is a result of feedback analysis on past students. Quite a few computer major students don't know how to build up a LAN when they go to work. Some of them cannot distinguish wireless AP from wireless router.

More experimental projects are supported by the three major system. It gives us the opportunity to alternate experimental contents every year. The projects list in TABLE IV can be replaced by equivalent ones.

Text	**Alingnment**	**Font**	**Followed by:**
Title	Centre	20 pt. TNR, bold, small-caps	24 pt. line sp.
Authors	Centre	13 pt. TNR	12 pt. line sp.
Addresses	Centre	12 pt. TNR	
emails	Centre	11 pt. italic TNR	18 pt. line sp. (last)
Abstract heading	Left	13 pt. bold italic TNR, small caps	6 pt. line sp.
Abstract text	Left	10 pt. italic TNR	12 pt. line sp.
Keywords heading	Left	13 pt. bold italic TNR, small caps	6 pt. line sp.
Keywords	Left, left, ..	10 pt. italic TNR	18 pt line sp.
Section headings	Left	14 pt. bold TNR, small caps	6 pt. line sp.
Sub-section heads	Left	12 pt. bold TNR	6 pt. line sp.
Sub-sub-sections	Left	11 pt. bold TNR	6 pt. line sp.
Body text	Full (left/right)	11 pt. TNR	12 pt line sp. (last)
Figures	Centre		6 pt. line sp.
Figure captions	Centre	11 pt. TNR	12 pt. line sp.
References	Left	10 pt. TNR (as shown)	6 pt. line sp

5. CONCLUSIONS

According to the teaching system of computer network experiment in CS Dept. Tsinghua University, we proposed the principles of designing computer network laboratory platform: multi-lever, remote, scalable and up-to-date. Then we design and implement three types of network experimental equipment: Virtual Network Experiment System based on commercial network equipment, Network Protocols Experiment System named NetRiver and IPv4/IPv6 Transition Technologies Experiment System based on 4over6 tunneling technology. These devices satisfy our designing principles. By proper combination of the experimental contents supported by them, we set up computer network laboratory education for the three experiment courses: Computer Network Experiment for non-computer major students, Computer Network Principles Experiment for computer major students and Specialty Practice in Network Security for network professional students.

REFERENCES

[1] MIT., Computer System Engineering, Spring 2016, http://web.mit.edu/6.033/www/

[2] Stanford University, Introduction to Computer Networking, Autumn 2011, http://www.scs.stanford.edu/11au-cs144/

[3] Andrew S. Tanenbaum, 2004, "Computer network", The 4th edition, Beijing: Tsinghua University Press

[4] S. Wei, Z. Qunyi,L. Xuefen, "Research of Educational Technology of Computer Networks", 2010 International Conference on E-Health Networking, Digital Ecosystems and Technologies, vol.4 pp. 192-195, 17-18 April 2010

[5] J. Janitor, F. Jakab, K. Kniewald, "Visual Learning Tools for Teaching/Learning Computer Networks", ICNS '10 Proceedings of the 2010 Sixth International Conference on Networking and Services, Pages 351-355

[6] Cisco Networking Academy, http://www.cisco.com/go/netacad

[7] N. Al-Holou, K. K. Booth, E. Yaprak, "Using computer network simulation tools as supplements to computer network curriculum", Frontiers in Education Conference, 2000. FIE 2000. 30th Annual, S2C/13 - S2C/16 vol.2, 18-21 Oct 2000

[8] N. Jovanovic, Z. Jovanovic, et al., " Computer Network Simulation and Visualization Tool for Educational Purpose", 11th International Conference on Telecommunication in Modern Satellite, Cable and Broadcasting Services (TELSIKS), 2013, vol.4 pp. 579-582, 16-19 Oct 2013

[9] M. W. Bazzaza, K. Salah,"Using the Cloud to Teach Computer Networks", 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC),pp. 310-314, 2015

[10] C. Yan, "Bulid A Laboratory Cloud for Computer Network Education", 2011 6th International Conference on Computer Science & Education (ICCSE), pp. 1013-1018, 2011

[11] L. Bellido, V. Mateos, et al., "Remote Access to Computer Networking Laboratories", 2012 9th International Conference on Remote Engineering and Virtual Instrumentation (REV), pp. 1-6, 2012

[12] Xu M., XU K., et al., "NetRiver, a computer network experiment system", SCE '08: Proceedings of the 1st ACM Summit on Computing Education in China on First ACM Summit on Computing Education in China,October 2008

[13] Ruijie Corporation, http://www.ruijie.com.cn/

[14] WireShark software, https://www.wireshark.org/

AUTHORS

Ninghan Zheng: Born in 1979. Master Degree of CS. Work in Department of Computer Science and Technology, Tsinghua University in Beijing, China. Engaged in experimental teaching research of Computer Network and Embedded System.

INTENTIONAL BLANK

BASIC EVALUATION OF ANTENNAS USED IN MICROWAVE IMAGING FOR BREAST CANCER DETECTION

Nouralhuda A. Hassan[1], Moustafa.M. Mohamed[1] and Mazher B. Tayel[2]

[1]Department of Medical Equipment Technology, Faculty of Allied Medical Sciences, Pharos University in Alexandria, Alexandria, Egypt.
[2]Electronic and Communication Engineering, Faculty of Engineering Alexandria, Egypt.
nouralhuda2006@yahoo.com

ABSTRACT

Microwave imaging is one of the most promising techniques in diagnosis and screening of breast cancer and in the medical field that currently under development. It is nonionizing, noninvasive, sensitive to tumors, specific to cancers, and low-cost. Microwave measurements can be carried out either in frequency domain or in time domain. In order to develop a clinically viable medical imaging system, it is important to understand the characteristics of the microwave antenna. In this paper we investigate some antenna characteristics and discuss limitations of existing and proposed systems.

KEYWORDS

Microwave imaging, breast cancer, UWB antenna

1. INTRODUCTION

Breast cancer is being one of the most frequent form of cancer that leading cause of cancer deaths in women worldwide. Around 50% of breast cancer cases death because the detection of the cancer is typically late. The early detection of tumor could save a lot of lives. Around 7% of women with breast cancer are diagnosed before the age of 40 years. Survival rates are worse when compared to those in older women [1,2].

Thus the role of cancer screening has become increasingly important leading to a demand in effective diagnostic measures; in particular non-invasive cancer diagnoses.
X-ray mammography is the most frequently used tool for breast cancer requires medical expertise to accurately diagnose the presence of tumor. The number of cancers found with mammography alone is very much less than that found with both mammography and physical examination. Other limitations include having high false negative and false positive rates [3].

Such large false negative which can be as large as 30% and false positive: on average, 75% of breast biopsies prompted by a “suspicious” mammographic abnormality prove benign which lead

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 55–63, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61005

to increased healthcare cost, unnecessary medical procedures and the distress and anxiety on the part of the patient [4]. Other important concerns ,is that screening mammography is less sensitive in women with radio-graphically dense breast tissue prevalent in younger women, where it has been shown that X-ray mammography has failed to detect up to 30% of cancers greater than 5 mm in diameter [5], due to its relatively poor soft-tissue contrast.

The other drawbacks include variability in radiological interpretation, and a slight risk of inducing cancer due to the ionizing radiation exposure. Frequent monitoring is difficult because of health concerns related to exposure to ionizing radiation.

2. Electrical Properties of The Tissues

A large scale studies determine the dielectric properties of a variety of normal, malignant and benign breast tissues, measured at the microwave range has been conducted in the USA [6]. The potential for using microwaves for detecting breast tumors is based on the concept of tissue-dependent microwave scattering and absorption in the breast to exploit the contrast in the dielectric properties of malignant and normal breast tissues.

It has been widely assumed that normal breast tissue is largely transparent to microwaves because they are featured with a low relative permittivity and conductivity at the microwave frequency bands, whereas lesions, which contain more water and blood are characterized by a high relative permittivity and conductivity at the microwave frequencies and hence they cause a significant backscatter [7]. Microwave breast cancer detection is based on the difference in the dielectric properties between healthy (ϵr=9 and σ=0.4 S/M) and malignant (ϵr=50 and σ=7 S/M) for example [8].

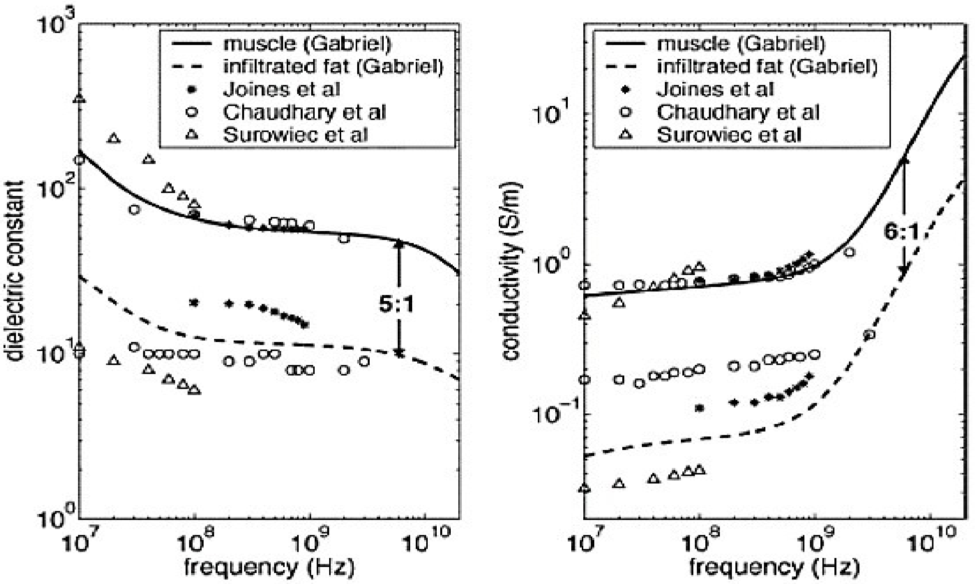


Figure 1: Electrical behavior of normal and malignant tissues at different microwave frequencies acquired by applying four-term Cole-Cole parametric dispersions models [9].

3. MICROWAVE IMAGING TECHNIQUES

There are three different methods in microwave imaging for breast cancer detection, passive, hybrid, and active.

3.1 Passive Microwave Imaging

In passive microwave imaging based on the fact that microwave pulses causes the cells temperature inside the breast increases, and due to different characteristics different tissues have different temperature. The produced image in this method shows measured temperature. In this method, the tumor is detected as points with higher temperature due to its characteristics [10–11].

3.2 Hybrid Method Microwave

In hybrid method, microwave and acoustic are combined together. Because tumors absorb higher energy compare with normal tissue they expand more. This expansion causes pressure waves. Using a focused ultrasound transducer these waves can be detected [12].

3.3 Active Microwave Imaging

In active microwave imaging, breast is illuminated by a microwave transmitter and the scattered waves are received. There are two approaches for image making, tomography and radar-based microwave imaging. In tomography, by analyzing the received scattered waves, we try to reconstruct the permittivity distribution of the breast tissues.

The higher permittivity shows tumor compare with low permittivity as normal tissue. In radar-based microwave imaging based on the strong scattered waves, tumor is detected. There are several researches using this method for breast cancer detection [13–15]. Based on the design, patient lies on supine or prone position and an antenna array locates on the breast. An UWB pulse is transmitted sequentially from each antenna. The scattered signal received and analyzed, and using these signals and space time beam forming method an image is created. Because tumors have a stronger backscattering compare with normal breast tissues, locations related to malignant tumors have higher energy level in the image.

4. MICROWAVE ANTENNAS EMPLOYED IN IMAGING

Antenna is the key element of the microwave imaging system that radiates and receives signals to or from nearby scattered objects. Many medical antennas have been designed and realized over the years. Some are patented, some are successful while others have never had the spread that was intended. Antenna technology for medical applications is a major research field.

The primary requirement for the UWB imaging antennas to be used in the 3D time domain non-linear super resolution inverse scattering microwave imaging techniques is low dispersive behavior (linear phase) over the operating bandwidth of 1-4 GHz.

There are many challenges for designing UWB antenna includes the ultra-wideband performance of the impedance matching, (i.e. good impedance matching “S11< -10 dB” in the band of 3.1 to 10.6 GHz), small size, minimum distortion, stable radiation pattern, and stable gain. Currently, there are many antenna designs that can achieve broad bandwidth to be used in UWB systems.

Various microwave antennas are used across the world by different microwave medical Imaging groups. This section details four such antennas which are either used in medical imaging applications or are identified as potential solutions to be used. In what follows, a discussion on each of these antennas will be made.

4.1 Dipole and Monopole Antenna

The dipole antenna is reasonably compact and lightweight, however, its bandwidth is not very large and, when excited by an impulse, the reflections of the impulse from the ends of the dipole are evident as a long, ringing, impulse response which, as discussed above, is undesirable.

The end-reflection problem can be eased by placing lossy material at the dipole ends in order to reduce the reflected wave (and hence the Q of the antenna), or by placing resistors at a quarter-wavelength distance from the ends of the antenna. The latter technique was used as far back as the early 1960s to create a travelling-wave antenna, although the varying electrical distance from the resistors to the ends of the dipole arms makes this a rather bandwidth-limited approach and it also requires the length of the dipole to be greater than a half-wavelength [16].

A resistively loaded monopole constructed by soldering together chains of 32 resistors has also been presented, with a quoted efficiency of 25% and very good impulse-radiation characteristics. In general, as the electrical size of the antenna reduces, the amount of resistive loading must be increased to maintain bandwidth, and efficiency reduces.

By using monopole antennas the entire imaging region can be illuminated by placing them close to the target, whereas in other antennas the distance has to be greater in order to provide sufficient illumination coverage. Space advantage offered by the monopole transmitters can prove to be very useful for systems using multiple transmit/receive channels. In a medium such as air or deionized water this type of antenna is notorious for producing exciting currents [17]. Demonstrate that the isotropic radiation pattern of the monopole does not serve to degrade imaging performance in the near field context, rather it actually increases the image quality obtained.

In order to realize a clinically viable system a fixed array data acquisition design may be desired. Planar monopole antennas are considered as promising candidates for microwave imaging. Many modern designs try to add more enhancements in terms of side lobe level (SLL) and the size.

4.2 Bow-Tie Antenna

The design of an efficient wideband coplanar strip line fed bow-tie antenna with improved bandwidth, low cross polarization and reduced back radiation [18]. The new antenna is constructed by structurally modifying the conventional micro strip bowtie antenna design; this is achieved by attaching an image plane. The antenna is designed as a patch on a single layered substrate with εr = 4.28 and thickness of 1.6mm.

The coplanar strip line is designed to have an input impedance of 50Ω in order to couple the antenna effectively with the measurement system. The parameters, such as the distance to the image plane, flare angle of the bow, and dimensions of the antenna, are found to affect the bandwidth. These parameters are optimized to enhance the performance. The antenna exhibits unidirectional radiation pattern with enhanced bandwidth reduced back radiation and low cross polarization in the operational band and thus making it suitable for Confocal Microwave Imaging (CMI). A typical wideband bow-tie antenna with coplanar strip line feed for CMI is shown in Figure 2. CMI employs back scattering to locate breast cancer tumors, so the antenna employed is required to focus the microwave signal towards the target and collect the back scattered energy . A 2:1 Standing Wave Ratio (SWR) bandwidth of 45.9% is obtained for the designed 4x4cm bow-tie antenna in air, which has a flare angle of 90°. The antenna operates in the band of 1850MHz - 3425 MHz with a return loss of -53dB. It is reported that in corn syrup the bandwidth is enhanced to 91% in the range of 1215 MHz – 3810 MHz with resonant frequency of 2855MHz and return loss of -41dB[19].

4.3 Vivaldi Antenna

The Vivaldi antenna that satisfies the requirements for imaging systems in terms of bandwidth, gain and impulse response, albeit at the expense of significant volumetric size. In addition to the bandwidth requirement, the antenna supports the sub nanosecond pulse transmission with negligible distortion to achieve precision imaging without ghost targets. Later in 2006 designed a Vivaldi antenna that reduced its physical dimensions such that it can be incorporated in a compact microwave imaging detection system whilst maintaining its ditortionless performance [20]. A typical Ultrawideband Antipodal Vivaldi antenna operates over an Ultrawideband (UWB) from 3.1GHz to 10.6GHz with a peak gain of 10.2dBi at 8GHz. These characteristics show that the Antipodal Vivaldi antenna has the potential to be used in medical imaging applications. The antenna is capable of radiating an impulse with little distortion, and their Superior directionality results in useful gain. This type of antenna is commonly used in commercial systems.

4.4 Pyramidal Horn Antenna

Horn Antennas are capable of radiating an impulse with little distortion, and their directionality results in useful gain. This type of antenna is commonly used in commercial systems, including those produced by Geophysical Survey Systems, The antennas are known for their higher aperture efficiencies but are constrained to certain applications due to their limited bandwidths. However, the bandwidth of the horn antennas can be increased significantly by adding metallic ridges to the waveguide and flared sections. Numerical and experimental investigations of pyramidal horn antennas with double ridges have been reported [21]. A designed a modified version of the ridged horn antenna in which the waveguide section is eliminated and one of the two ridges is replaced by a curve metallic plane terminated by resistors. Later in 2003 Susan C. Hagness and her team presented a complete numerical and experimental study of a specific realization of this design, wherein the antenna is customized to centimeter scale dimensions for operation in the microwave frequency range 1 to 11 GHz [22]. The pyramidal horn is connected to the outer conductor of the coaxial feed and serves as the ground plane, providing a current return path. Because of the coaxial feed, the ground plane configuration eliminated the need for a UWB Balun. The dimensions of the horn antenna are chosen according to the physical size required and operating frequency range.

This antenna yields VSWR of less than 1.5 over the frequency range and fidelity of approximately 0.96 in both the simulation and experiment [22]. The antenna has been tested under low loss immersion medium and achieved similar VSWR and fidelity. Overall it is evident that this type of antenna can be useful for biological sensing and imaging application.

4.5 Stacked-patch Antennas

While stacked-patch antennas are well known to have good operating bandwidths, the bandwidths achieved are usually of the order of 30 % [23]. The stacked patch antenna developed at the University of Bristol was designed from the outset to radiate directly into breast tissues, and furthermore achieves a bandwidth of approximately 77 %. It achieves this without resistive loading, and, in fact, FDTD models demonstrate that even if the losses in the surrounding tissues are removed, the bandwidth is practically unchanged.

4.6 Log periodic and spiral antennas

Although both log periodic and spiral antennas can operate in the UWB frequency band (3.1-10.6 GHz), they are not suitable for imaging applications because they have large physical dimensions as well as their dispersive characteristics and severe ringing effect [23].

5. SELECTED EXPERIMENTAL MICROWAVE IMAGING SYSTEMS

Investigated operational systems with the characteristics are described in Table 1.

Origin	Antenna configuration	Targets
Dartmouth College	2-8 antennas mechanical scanning, laser camera system	500+patients
University of Bristol	60 antennas ceramic fitting cups	95 patients
University of Calgary	1 Antenna mechanical scanning	Pilot clinical experiments 9 patients
McGill University	16 antennas two perpendicular arcs	13 healthy patients
ETRI, Korea	16 antennas Dartmouth inspired	Dogs Suitable for humans
Carolinas Medical	24 antennas 2-D dynamic phenomena	Anesthetised pig
University of Michigan	36 bow ties, 3 circular arrays	Acrylic spheres or hyperthermia
Supélec	2 large horns w. retinas	Tube with water in cylinder w. Triton X-100 mixture
University of Manitoba	1 Vivaldi antenna	Misc. dielectric cylinders

Politecnico	8 monopole antennas	Metal cylinder in glycerin/ water
Technical Univ. of Catalonia	1 fixed, 1 on linear positioner	Misc. rods and cylinders Clay balls in paraffin

6. CONCLUSION

The role of cancer screening and detection has become increasingly important leading to a demand in effective diagnostic measures; in particular non-invasive and non-ionizing cancer diagnostics. Microwaves, along with other methods, are actively pursued as an alternative to existing imaging modalities, which may help reduce the number of false• positive and false-negatives, especially in challenging cases radiographically dense breast tissue.

Microwave imaging is based on the significant dielectric contrast between normal and cancerous breast tissues at microwave frequencies. The technique is considered particularly promising due to the relatively short required penetration depth and the accessibility from different angles. Nevertheless, there are challenges associated with this modality. In terms of experimental implementation, the challenges include the design and fabrication of UWB antenna elements and arrays, the management of the aperture size and scan time, etc. Recent research suggests the use of microwaves for breast tumor detection, in particular the ultra-wideband (UWB) frequency region, offering a promising trade-off between imaging resolution and tissue penetration depth.

ACKNOWLEDGEMENTS

I would like to acknowledge Eng. Maged Eldaheebi for his advice and assistance. Many thanks to all reviewers for their time and effort.

REFERENCES

[1] National Cancer Institute Website: http://www.cancer.gov/types/breast/mammograms-fact-sheet.

[2] Anders CK, Johnson R, Litton J, Phillips M, Bleyer A. Breast Cancer Before Age 40 Years. Seminars in oncology. 2009;36(3):237-249. doi:10.1053/j.seminoncol.2009.03.001.

[3] S. Nass, I. Henderson, and J. Lashof, Mammography and beyond: Developing technologies for the early detection of breast cancer, National Academy Press, Washington D.C., 2001.

[4] E. A. Sickles, “Nonpalpable, circumscribed, noncalcified, solid breast masses: Likelihood of malignancy based on lesion size and age of patient,” Radiology, vol. 192, pp. 439–442, 1994.

[5] M. Lazebnik, et al, “A large-scale study of the ultrawideband microwave dielectric properties of normal breast tissue obtained from reduction surgeries,” Physics in Medicine and Biology, vol. 52, pp.2637–2656, 2007.

[6] Y. Xie, B. Guo, L. Xu, J. Li, and P. Stoica, “Multistatic adaptive microwave imaging for early breast cancer detection,” IEEE Trans. Biomed. Eng., vol. 53, pp. 1647–57, 2006.

[7] Surowiec, A. J., S. S. Stuchly, J. R. Barr, and A. Swarup, “Dielectric properties of breast carcinoma and the surrounding tissues,” IEEE Trans. Biomed. Eng., Vol. 35, No. 4, 257–263, 1988.

[8] Kösters JP, Gøtzsche PC (2003). "Regular self-examination or clinical examination for early detection of breast cancer". Cochrane Database Syst Rev (2): CD003373.

[9] B. Bocquet, J.C. van de Velde, A. Mamouni, Y. Leroy, G. Giaux, J. Delannoy, and D. Delvalee, Microwave radiometric imaging at 3 GHz for the exploration of breast tumors, IEEE Trans MicrowaveTheory Tech 38 (1990), 791–793.

[10] P.R. Stauffer, D.B. Rodriques, S. Salahi, E. Topsakal, T.R. Oliveira, A.Prakash, F. D'Isidoro, D. Reudink, B.W. Snow, and P.F. Maccarini,Stable microwave radiometry system for long term monitoring of deep tissue temperature, In: Proceeding SPIE 8584, Energy-based Treatmentof Tissue and Assessment VII, San Francisco, CA, 2013.

[11] D.B. Rodrigues, P.F. Maccarini, S. Salahi, T.R. Oliveira, P.J.S. Pereira, P. Limao-Vieira, B.W. Snow, D. Reudink, and P.R.Stauffer, Design and optimization of an ultra-wideband and compact microwave antenna for radiometric monitoring of brain temperature,IEEE Trans Biomed Eng 61 (2014), 2154–2160.

[12] X. Wang. Thermoacoustic applications in breast cancer detection and communications, Ph.D. Thesis, The University of Arizona,Arizona, 2014.

[13] W.C. Khor, M.E. Bialkowski, A. Abbosh, N. Seman, and S. Crozier, An ultra wideband microwave imaging system for breast cancer detection, IEICE Trans Commun 90 (2007), 2376–2381.

[14] R.K. Amineh, M. Ravan, A. Trehan, and N.K. Nikolova, Near-field microwave imaging based on aperture raster scanning with tem horn antennas, IEEE Trans Antennas Propag 59 (2011), 928–940.

[15] X. Li and S.C. Hagness, A confocal microwave imaging algorithm for breast cancer detection, IEEE Microwave Wireless Compon Lett 11 (2001), 130–132.

[16] H. Meinke and F.W. Gundlach, Taschenbuch der Hochfrequenztechnik, Berlin, Springer-Verlag, pp.531–5, 1968.

[17] Meaney, P.M., K.D. Paulsen, and J. Chang, Near-Field Microwave Imaging of Biologically based materials using a monopole system. IEEE Transactions on Microwave Theory and Techniques, 1998. 46(1).

[18] Bindu, G., et al., Wideband Bow-tie antenna with Coplanar Stripline Feed. Microwave and Optical Technology Letters, 2004. 42(3).

[19] Bindu, G., et al., Active Microwave Imaging For Breast Cancer Detection. Progress in Electro magnetics Research 2006. 58: p. 149-169.

[20] Abbosh A.M., H.K. Kan, and M.E. Bialkowski, Design of Compact Ultra Wideband Antipodal Antenna. Microwave and Optical Technology Letters, 2006. 48(12).

[21] Notras, B.M., C.D. McCarrick, and D.P. Kasilingam, Two Numerical techniques for analysis of pyramidal horn antenna with continous metallic ridges. Proceedings of IEEE Internationsl Symposium Antenna Propagation, Dig., 2001. 2: p. 560-563.

[22] Li, X., et al., Numerical and Experimental Investigation of an Ultrawideband Ridged Pyramidal Horn Antenna with Curved lauching Plane for Pulse radiation. IEEE Antennas and Wireless Propagation Letters, 2003. 2.

[23] R.J. Fontana, Recent system applications of short-pulse ultra-wideband (UWB) technology, IEEE Trans. Microwave Theory Techn., 52, 2087–104, 2004.

[24] A.F. Molisch, J.R. Foerster and M. Pendergrass, Channel models for ultra-wideband personal area networks, IEEE Wireless Communications Magazine, 10(6), 14–21, 2003.

INTENTIONAL BLANK

EXPLORE THE EFFECTS OF EMOTICONS ON TWITTER SENTIMENT ANALYSIS

Katarzyna Wegrzyn-Wolska[1], Lamine Bougueroua[1], Haichao Yu[2], Jing Zhong[2]

[1]Esigetel, Groupe Efrei Paris-Sud, Villejuif, France
katarzyna.wegrzyn@groupe-efrei.fr,
lamine.bougueroua@groupe-efrei.fr
[2]Allianstic, Groupe Efrei Paris-Sud, Villejuif, France
haichao.yu.20150767@efrei.net, jing.zhong.20150772@efrei.net

ABSTRACT

In recent years, Twitter Sentiment Analysis (TSA) has become a hot research topic. The target of this task is to analyse the sentiment polarity of the tweets. There are a lot of machine learning methods specifically developed to solve TSA problems, such as fully supervised method, distantly supervised method and combined method of these two. Considering the specialty of tweets that a limitation of 140 characters, emoticons have important effects on TSA. In this paper, we compare three emoticon pre-processing methods: emotion deletion (emoDel), emoticons 2-valued translation (emo2label) and emoticon explanation (emo2explanation). Then, we propose a method based on emoticon-weight lexicon, and conduct experiments based on Naive Bayes classifier, to validate the crucial role emoticons play on guiding emotion tendency in a tweet. Experiments on real data sets demonstrate that emoticons are vital to TSA.

KEYWORDS

Social Media, Social Network Analysis, Text Mining, Sentiment analysis, Tweets, Emoticon

1. INTRODUCTION

Sentiment Analysis (SA) [1] is a computational study of how opinions, attitudes, emoticons and perspectives are expressed in language. With the development of social network and dramatic development of big data, SA has been applied to a variety of domains to solve practical problems, such as understanding customer feedback, brand analysis, understanding public opinions, financial prediction, etc. Therefore, SA has become an important and hot research topic, which has attracted a large number of researchers from domains of machine learning, data mining and natural language processing (NLP). Theoretically, there are 3 classes of sentiment: positive, negative and neutral. However, most of the researchers usually focus on polarity classification: classifying sentence or document as positive or negative, which is two-way classification problem. Since SA has been formulated as machine learning based text classification problem by [2] [3] [4], machine learning methods have become the most important methods to solve SA problem.

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 65–77, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61006

Twitter is one of the most popular online social networking service today, which allow users to send and read short messages called tweets. With tweets, people can share with other people what they are doing and thinking [5]. According to recent statistical data[1], as of March 2016, there have been more than 310 million monthly active users and 330 million tweets are generated every day. The most important feature of Twitter is that every tweet is a message up to 140 characters. It is because of this character limitation that emoticon become very important in tweets, since emoticon can help people better express their emotion in a short message. However, most of the researchers have dismissed emoticons as noisy information and delete them in the pre-processing process. Nevertheless, we will explore the influence of emoticons on SA in this paper.

Very often SA is applied on movies review and news article [3] [4] [6]. Compared with movie reviews and news articles, tweets have a lot of difference [7]. On the one hand, tweets are shorter and more ambiguous than movie reviews and news articles because of the limitation of words. On the other hand, tweets contain much more misspelled words, slang, modal particles and acronyms because of the casual form. Considering these difference, the traditional SA methods for movie reviews and news articles are not appropriate for Twitter Sentiment Analysis (TSA) problem. Actually, many novel SA methods have been specifically developed for TSA, which include fully supervised method and distantly supervised method. With manually labelled data, fully supervised methods like Multinomial Naive Bayes (MNB) and support vector machine (SVM) are more accurate, but labelling data manually is more labour-intensive and time consuming. With data collected by Twitter API, distantly supervised methods are more efficient but less accurate. [8] even combined these two methods and developed the emoticon smoothed language models (ESLAM) for TSA.

In this study, we explore the effects of emoticons on TSA. At first, we compare three emoticon pre-processing methods: emotion deletion (*emoDel*), emoticons 2valued translation (*emo2label*) and emoticon explanation (*emo2explanation*). After that, we propose a method based on emoticon-weight lexicon to explore the influence of emotion on TSA. Experiments on real data sets demonstrate that emoticons are vital to TSA.

2. RELATED WORK

SA [1] has been a popular research topic over the past decades. Before [2], knowledge-based method dominated this domain. However, in [2], authors show that machine learning techniques like naive Bayes, maximum entropy and support vector machine can outperform the knowledge-based baselines on movie reviews. After that, machine learning based methods have become the most important methods for SA.

With the rapidly growth of Twitter, more and more researchers started to focus on TSA. Most of earlier works on TSA are fully supervised methods. In [9] [10], authors use traditional SA methods on normal text form to solve TSA problems. Authors propose target-independent SA based on SVM in [11]. In [12], authors present a dynamic artificial neural network to handle TSA.

Recently, different supervised methods are proposed. Authors in [13] utilize Twitter API to get training data which contain emoticons like :) and :(. They use these emoticons as noisy labels.

[1] https://about.twitter.com/company

Tweets with :) are thought to be positive training data and tweets with :(are thought to be negative training data.

In [8], authors present the ESLAM which combine fully supervised methods and distantly supervised methods. Although a lot of TSA methods have been presented, few of them explored the influence of emoticons on TSA, which motivates our work in this paper.

3. EXPLORE EFFECTS OF EMOTICONS

In this section, first we present our basic TSA classifier based on Naive Bayes (NB). Then, we introduce an emoticon lexicon which contain 50 most commonly used emoticons. After that, we present 3 emoticon pre-processing methods: *emoDeletion, emo2label* and *emo2explanation*.

Finally, we propose a method based on emoticon-weight lexicon and introduce a strategy to integrate emoticon-weight lexicon method with naive Bayes method.

3.1. Naive Bayes (NB) Model for SA

In this paper, we use a Twitter-aware tokenizer[2] combined with a Naïve Bayes model as our basic classifier. Refer to the Stanford Classifier[3], here is the basic idea for the Naive Bayes:

We assume that:

- n is the number of words appeared in training set T,
- n_c_j is the number of feature which belong to class j (c_j) in training set T (j can be positive or negative),
- n_f_i is the number of times feature i appeared in training set T,
- $n\_f_{i}\_c_i$ is the number of times feature i appeared in class j.

Then, we use the following equations to compute the probabilities p_c_j and $p\_f_{i}\_c_{j}$:

$$p\_c_j = \frac{n\_c_j + \varepsilon}{n + |classes| \times \varepsilon} \tag{1}$$

$$p\_f_{i}\_c_j = \frac{n\_f_{i}\_c_j + \sigma}{n\_f_i + |classes| \times \sigma} \tag{2}$$

While we have two classes (positive and negative), so |classes| = 2.

In (1) (2), the parameters ε and σ are smoothing parameters to avoid assigning zero weight to unseen feature. In our experiment, we choose $\varepsilon = 10^{-30}$ and $\sigma = 1.0$ (Laplacian smoothing).

With (1) (2), we can compute negative weight and positive weight of every feature:

$$W_{i,j} = \log\left(\frac{p\_f_{i}\_c_j}{p\_c_j}\right) \tag{3}$$

[2] http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py

[3] http://nlp.stanford.edu/software/classifier.shtml

After get weights of all features, we can compute the weights of sentences according to Naive Bayes assumption.

Assuming that tweet *t* consists of *n* features, then the weights of the tweet *t* will be:

$$W\_sentence_{t,j} = \sum_{i=1}^{n} W_{i,j} \tag{4}$$

Finally, we will compute the possibilities of the sentence belonging to negative class and positive class:

$$P(t \mid neg) = \frac{e^{W_{t,neg}}}{e^{W_{t,neg}} + e^{W_{t,pos}}} \tag{5}$$

$$P(t \mid pos) = \frac{e^{W_{t,pos}}}{e^{W_{t,neg}} + e^{W_{t,pos}}} \tag{6}$$

3.2. Emoticon Lexicon

Our emoticon lexicon is based on a Twitter emoticon analysis[4] which collected a large number of most commonly used emoticons. We choose the top 50 emoticons as our emoticon lexicon.

For every emoticon, we give a polarity value which can be negative or positive, a specific translation and a weight. This lexicon is showed in Table 1. We will use this emoticon lexicon in subsequent parts.

Table 1. Emoticon Lexicon

Emoticon	Value	Translation	Weight
:) :D :-) ;) XD :] =) (: ;-) =D =] :-D ^_^ (8 :o) (;=o 8) ;o) (= [: 8D :]	POSITIVE	happy	1
:o ;O o:	POSITIVE	surprise	1
=P :-P ;P =P	POSITIVE	playful	1
;D ;]	POSITIVE	wink	1
\m/	POSITIVE	salute	1
:(D: =( ): ;) :[;(=[	NEGATIVE	sad	-1
=/ :-/ :\ ;/ :-/ =\	NEGATIVE	annoyed	-1
:’(	NEGATIVE	crying	-1
:@	NEGATIVE	angry	-1
:|	NEGATIVE	indifferent	-1

3.3. Emoticon Pre-processing Methods

EmoDeletion: In this emoticon pre-processing method, we just delete all the emoticons defined in emoticon lexicon in TABLE 1 from the training data.

[4] http://www.datagenetics.com/blog/october52012/index.html

Emo2label: This emoticon pre-processing method is pretty simple and straightforward. We give all the emoticons a 2-valued label: NEGATIVE or POSITIVE. We give a label of NEGATIVE to those emoticons with negative meanings and give a label of POSITIVE to those emoticons with positive meanings. This kind of translation is not so close to natural language, but it is more intuitive and robust because it could avoid some translation errors. For both training data and test data, when we find any emoticon defined in emoticon lexicon, we replace it with its 2-valued labels in pre-processing.

Emo2explanation: When two people communicate face to face, they could notice the expression like "smile" or "frown" made by the other. For example, A is frowning and says to B "I'm fine". If C asks B the recent situation of A, B will not ignore A's expression but translate A's expression naturally. B will say: "I saw some days ago. She said she was fine but I noticed she was frowning. So I think maybe she met some trouble." Such like that, almost every emoticon can be described as a verbal word and it is much easier for a computer to recognize a word rather than an emoticon since most of the features extracted by classifier are words. Because of the similarity of some emoticons, we organize emoticons into emoticon synonymy sets, which we define as groups of emoticons with the same translation (see TABLE 1). From both training data and testing data, when we find any emoticon defined in emoticon lexicon, we replace it with its translation in pre-processing. For example, a tweet "This movie so cool!! :)" are translated into "This movie so cool!! happy" after pre-processing.

3.4. Emoticon-Wight Lexicon Model (EWLM) for SA

In polarity classification, we place a text into negative or positive class. Similarly, we use a polar weight to define an emoticon which is a character sequences. For an emoticon with positive meaning, we give it the value 1, otherwise, we give it the value -1 [14]. The format of an emoticon-weight lexicon is (emoticon, weight), for example, (:), 1), (:(, -1).

When classifying a text, we consider both emoticons and verbal cues, and combine the two factors to get an integrated assessment to the text. The framework is as below [Figure 1]: Firstly, we load a set of tweets for analysing sentiment. Then, the classifier split it into different tweets. For each tweet, we check if this tweet contains emoticon.

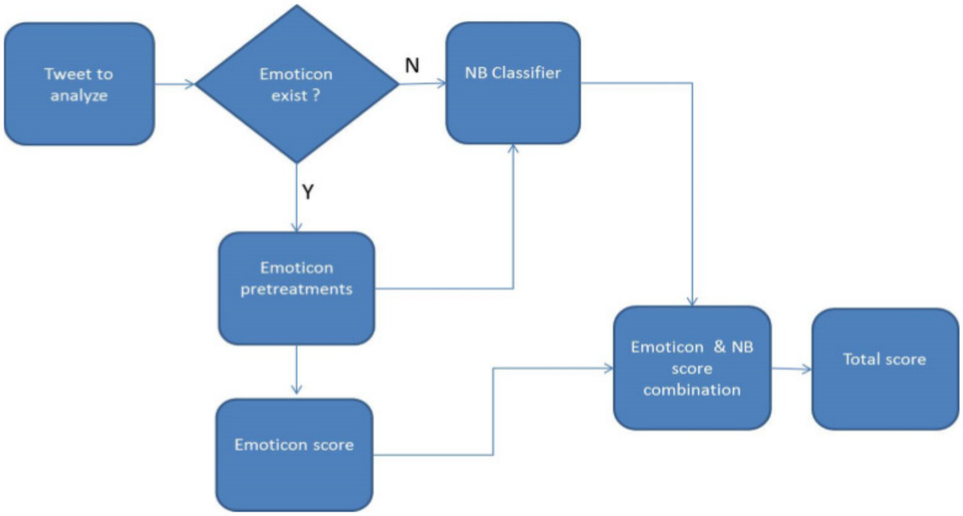


Figure 1. Framework architecture

We compare each word in the tweet with the emoticon lexicon entries. If there exist emoticons which match the emoticons in lexicon, we compute the emoticon score of this tweet and combine this score with words score. Otherwise, we just use the words score which is given by the NB classifier. When the tweet *i* contains emoticon, $e_i = 1$, otherwise $e_i = 0$. i.e.

$$e_i = \begin{cases} 0, & no\ emoticon\ in\ tweet\ i \\ 1, & exist\ emoticon\ in\ tweet\ i \end{cases} \tag{7}$$

For every tweet, the NB classifier gives us two probabilities $p_{iw}(neg)$ and $p_{iw}(pos)$ for classifying verbal cues. If $p_{iw}(neg) > p_{iw}(pos)$, the NB classifier places the tweet into negative class. Otherwise, the tweet is placed into positive class. When $e_i = 1$, the emoticon score of i^{th} tweet s_{ie} equals the sum of weights of each emoticon. Assuming that the number of emoticons in i^{th} tweet is $N_i\,(N_i > 0)$, and the weight of j^{th} emoticon is $W_{\_emoj}$, we have:

$$S_{ie} = \sum_{j=1}^{N_i} W_{\_emo_j} \tag{8}$$

The emoticon-weight lexicon helps us to deal with only emoticons. The NB classifier deals with verbal cues. Hence, we need a combination strategy to combine EWLM with NB classifier, to get a final classification result.

As above, s_{ie} is the sum of weight of emoticons in tweet $_i$, which is not in the range of (0, 1). We use the Sigmoid function to convert the range of s_{ie} into a new range which is between 0 and 1, because we need to combine this value with a probability value which is between 0 and 1 given by the NB classifier. With Sigmoid function, we can compute P_EWLM:

$$P_{EWLM}(t|pos) = Sigmoid(S_{ie}) \tag{9}$$

$$P_{EWLM}(t|neg) = 1 - Sigmoid(S_{ie}) \tag{10}$$

The sentiment of both emoticons and verbal cues can be computed as a probability of being negative or positive. We use α as a factor, which decides the importance of the emoticon in a tweet, to integrate these two probabilities and get the final probabilities. $p_i(pos)$ is the probability of the i^{th} tweet being positive, and $p_i(neg)$ is the probability of the i_{th} tweet being negative. If $\alpha \geq 0.5$, verbal cues play a more important role. Otherwise, the emoticon occupies a greater proportion on analysing sentiment.

$$P_i(neg) = \propto \times\, P_{NB(neg)} + (1 - \propto) \times P_{EWLM}(neg) \tag{11}$$

$$P_i(pos) = \propto \times\, P_{NB(pos)} + (1 - \propto) \times P_{EWLM}(pos) \tag{12}$$

The classification c_i of i^{th} tweet is defined as a function of its final probabilities $p_i(neg)$ and $p_i(pos)$:

$$c_i = \begin{cases} positive, & if\ P_i(pos) \geq P_i(neg) \\ negative, & if\ P_i(pos) < P_i(neg) \end{cases} \tag{13}$$

4. EXPERIMENT DESIGN

4.1. Data Set

We use the publicly available Sanders Corpus[5] as our experiment data, which consist of 5513 manually labelled tweets. These tweets involved with four different topics: Apple, Google, Microsoft, and Twitter. After removing the no English tweets, spam tweets, re-tweets and duplicate tweets, and setting the classes to be balanced, we get 952 tweets for polarity classification, including 476 negative tweets and 476 positive tweets. There are 200 tweets which contain emoticons in the whole data set (which means approximately 21% tweets contain emoticons).

We take the following measures to pre-process the data:

1. Replace the Twitter usernames which start with @ with USERNAME.
2. Replace urls in tweets with URL.
3. All words are changed to their lower cases. With these pre-processing measures, we can reduce the influence of meaningless strings and extract more representative features.

4.2. Experiment Setting

We assume that the total number of data, including training data and test data, is X (= 952). For every experiment, we randomly sample the same amount of tweets (say Y, Y = 16, 32, 64...) for both negative class and positive class as our training set, and use the rest X − 2Y tweets as our test set. In order to avoid the experiment contingency, every time we will conduct 60 times experiments independently and get the average performance, which is more accurate.

4.3. Evaluation

We evaluate the performance of our experiments by the values of accuracy and Macro-level F1-score. Accuracy is the percentage of correctly predicted data in all test data. The Macro-level F1-score is the average of the F1-scores of the positive and negative classifiers, where F1-score is the harmonic mean of precision and recall. F1-score is related with precision and recall calculated by the simplified formula [14]:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{14}$$

5. EXPERIMENT RESULTS

5.1. Effects of emoticon pre-processing methods

We conduct experiments based on NB model to compare with and without emoticon pre-processing methods and explore the influence of emoticons. In this experiment, we use different number of training data (i.e. 2Y = 32,64,128,256,512,768). The results are illustrated by Figure 2 with accuracy and Figure 3 with Macro-level F1-score.

[5] http://www.sananalytics.com/lab/twitter-sentiment/

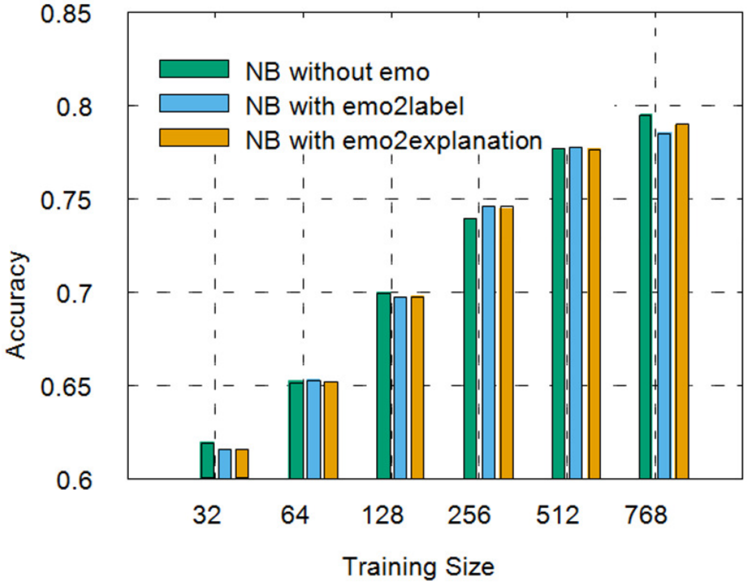


Figure 2. Effects of emoticon pre-processing methods measured by Accuracy

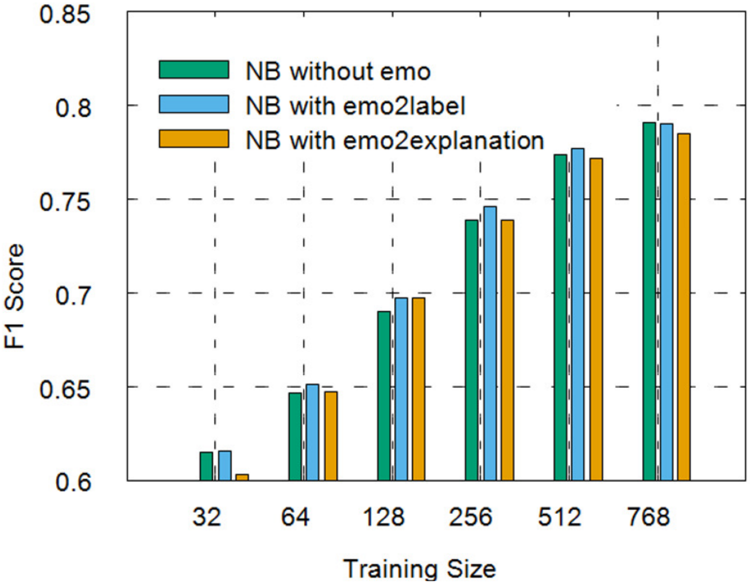


Figure 3. Effects of emoticon pre-processing methods measured by Macro level F1 score

From Figure 2 and Figure 3, we can easily see that emo2label has the best performance among the proposed emoticon pre-processing methods.

5.2. Effects of Emoticon-Weight Lexicon Model

We compare the performance of the NB model with and without EWLM to judge if EWLM can help the NB model to raise the performance on TSA. In this experiment, we also use different training size to train the classifier and utilizer accuracy and Micro-level *F1* score to evaluate the classifier.

The experiment result is showed in Figure 4 and Figure 5.

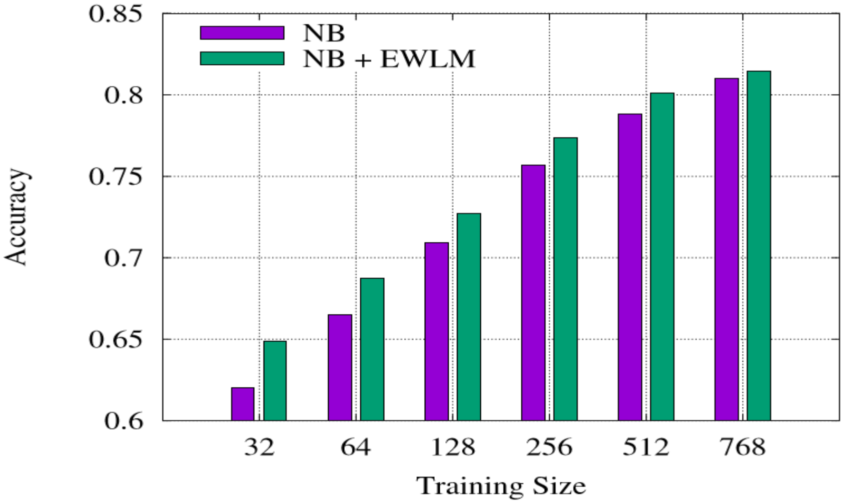


Figure 4. Effects of EWLM measured by accuracy

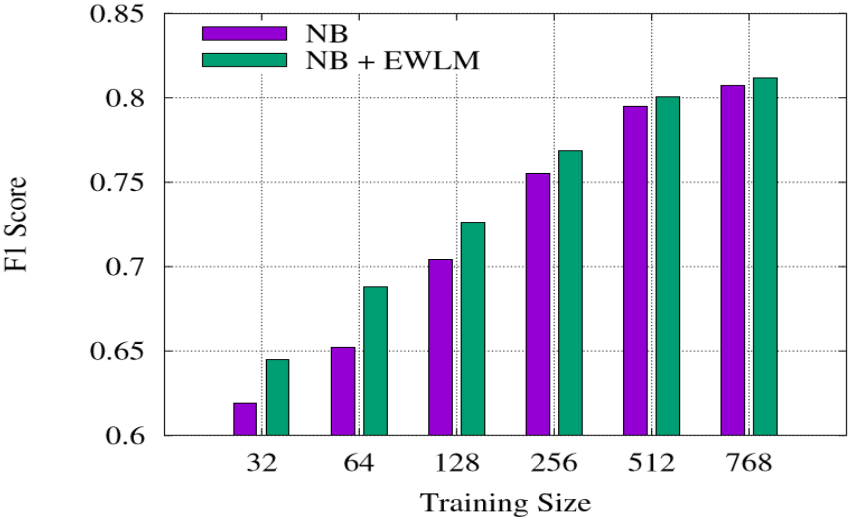


Figure 5. Effects of EWLM measured by Macro-level F1 score

From Figure 4 and Figure 5, it is obvious that EWLM can help the NB model to raise the performance on TSA, especially when the training size is small. When the training size is big enough, the data can provide more discriminating information for training the NB classifier, and the NB classifier could achieve a better performance. In this condition, the improvement brought by EWLM will become smaller. Anyway, the experiment results imply that the emoticons do have important information which could help the NB classifier to achieve better performance on TSA tasks.

5.3. Effects of the Combination Parameter Alpha

Alpha is a significant factor to combine NB model with EWLM. When alpha equals 1, there will be only NB model to conduct TSA task. When alpha is smaller, the EWLM will play a more important role in the combined classifier. In this experiment, we try different value of alpha to check which value of alpha is best. The experiment results can be seen in Figure 6 (training size equals 128) and Figure 7 (training size equals 512).

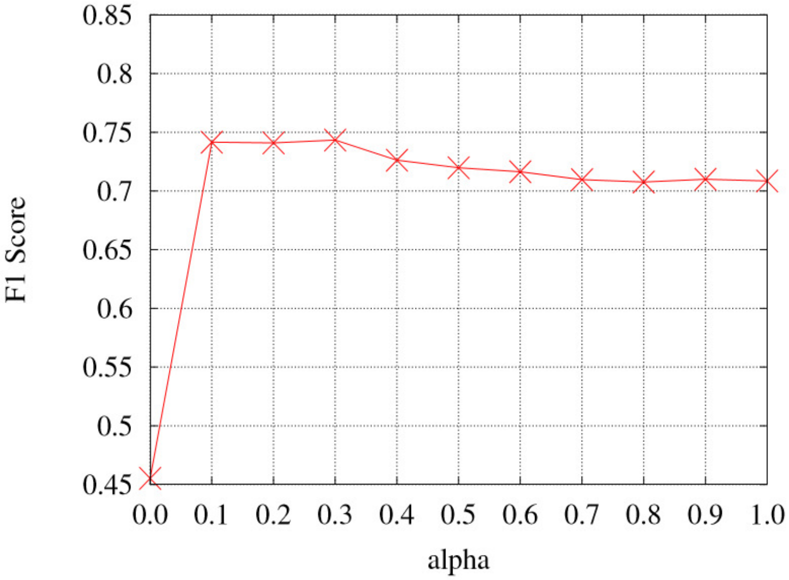


Figure 6. Effect of combination factor alpha with 128 training data

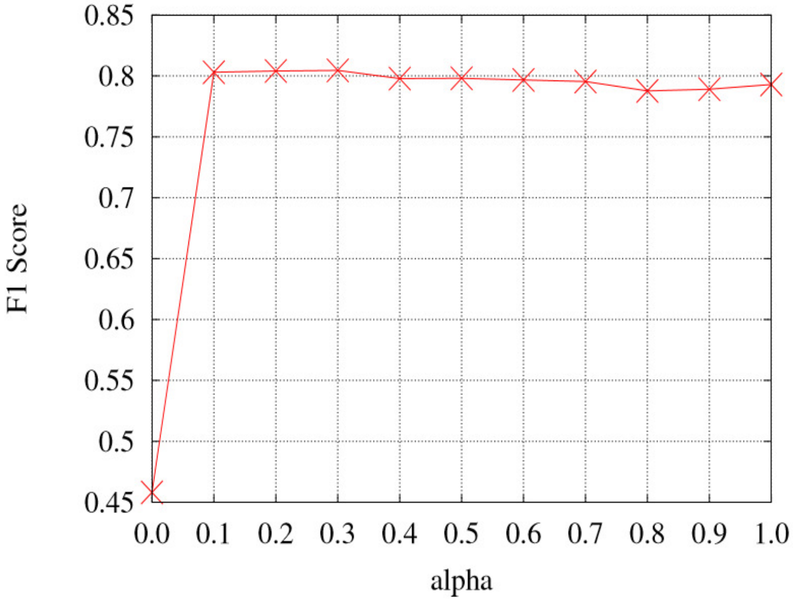


Figure 7. Effect of combination factor alpha with 512 training data

The experiments result in Figure 6 and Figure 7 clearly show that the combination strategy is better than the single NB model or single EWLM. Furthermore, we can see that when alpha take values from 0.1 to 0.3, the classifier can achieve the best performance. Also, in the experiment with 512 training example, we can notice that when alpha becomes larger, the performance of the classifier will not be influenced a lot. This is because a large manually labelled training data can provide enough discriminating information for the TSA classifier.

Our results could clearly indicate that considering emoticon into classifier is a necessary addition on sentiment analysis and whether a tweet contains emoticon or not, our methods will not weaken the performance. If no emoticon in data, the performance of our methods is same with NB classifier.

6. CONCLUSIONS

With the significance of sentiment analysis being recognized and the popularity rate of emoticon in social network getting higher and higher, the role of emoticon cannot be ignored on polarity classification. Our key contribution in this paper lies in validating the important role emoticon plays in conveying overall sentiment of a text in TSA though a series of experiments.

We compare 3 emoticon pre-processing methods and emoticon-weight lexicon method on the base of Twitter aware tokenizer and NB Model. We propose a combination strategy using factor alpha to integrate the Emoticon-Weight Lexicon with classifier. The result shows that the usage of emoticon-weight lexicon model improves the performance of NB model on TSA task. We can get the conclusion that some emoticons dominate the sentiment of a tweet and conquer the emotion of verbal cues.

As our results are very promising, we assume several directions for further work. First, we will look for some authoritative help to improve our emoticon dictionary and set more detailed score for emoticon weight to show its intensity of emotion. Second, we will study the impact of number of emoticons in experimental data on our emoticon weight lexicon.

REFERENCES

[1] Pang, Bo and Lee, Lillian, Opinion mining and sentiment analysis, Journal Foundations and trends in information retrieval, volume 2, number 1-2, pages 1–135, 2008.

[2] Pang, Bo and Lee, Lillian and Vaithyanathan, Shivakumar, thumbs up? sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language Processing-Volume 10, pages 79–86, 2002.

[3] Dziczkowski, G., & Wegrzyn-Wolska, K. 2007b. Rcss - rating critics support system purpose built for movies recommendation. In: Advances in Intelligent Web Mastering. Springer.

[4] Dziczkowski, G., & Wegrzyn-Wolska, K. 2008a. An autonomous system designed for automatic detection and rating of film. Extraction and linguistic analysis of sentiments. IN Proceedings of WIC, Sydney.

[5] Janik Lthi, Lamine Bougueroua and K. Wegrzyn-Wolska, Sentiment Polarity on Twitter messages with geolocation, in proceedings of the International Workshop on Computational Social Networks (IWCSN 2014) within the 15th International Conference on Web Information System Engineering WISE 2014, Thessalonique, Greece, October 2014, Springer Lecture Notes in Computer Science (LNCS).

[6] Dziczkowski, G., & Wegrzyn-Wolska, K. 2008b. Tool of the intelligence economic: Recognition function of reviews critics. In: ICSOFT 2008 Proceedings. INSTICC Press.

[7] Wegrzyn-Wolska, K., Bougueroua, L.: Tweets mining for French Presidential Election, In proceeding of the 4th IEEE/WIC International conference on computation aspects of social networks - CASoN 2012, SaO Carlos, Brazil, November (2012).

[8] Liu, Kun-Lin and Li, Wu-Jun and Guo, Minyi, Emoticon Smoothed Language Models for Twitter Sentiment Analysis., AAAI, 2012.

[9] Jansen, Bernard J and Zhang, Mimi and Sobel, Kate and Chowdury, Twitter power: Tweets as electronic word of mouth, Journal of the American society for information science and technology, volume 60, number11, pages 2169–2188, 2009, Wiley Online Library.

[10] Bermingham, Adam and Smeaton, Alan F, Classifying sentiment in microblogs: is brevity an advantage? Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1833–1836, 2010, ACM.

[11] Jiang, Long and Yu, Mo and Zhou, Ming and Liu, Xiaohua and Zhao, Tiejun, Target-dependent twitter sentiment classification, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 151–160, 2011.

[12] Ghiassi, M and Skinner, J and Zimbra, D, Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network, Journal Expert Systems with applications, volume 40/16, pages 6266–6282, 2013, Elsevier.

[13] Go, Alec and Bhayani, Richa and Huang, Lei, Twitter sentiment classification using distant supervision, journal CS224N Project Report, Stanford, volume 1, pages 12, 2009.

[14] Hogenboom, Alexander and Bal, Daniella and Frasincar, Flavius and Bal, Malissa and De Jong, Franciska and Kaymak, Uzay, Exploiting Emoticons in Polarity Classification of Text., J. Web Eng., volume 14, num1&2, pages22–40, 2015.

AUTHORS

Katarzyna Węgrzyn-Wolska received M.Sc. from the Silesian Technical University of Gliwice (Poland), and a further M.Sc. in Computer Science from the University of Val Essonne (France), her Ph.D. (2001) in Automatics, Real Time Computing and Computer Science from the Ecole Superieur des Mines de Paris, France and the habilitation (H.D.R), to become Full Professor in 2012. She is the Principal Professor and head of an SITR team at ESIGETEL, France. She is editor-in-chief of the International Journal on Social Informatics edited in ICST Transactions Series. She is involved in the organization of several International Conference as well as an expert for the group Information Society Technologies (IST) active in the European Community. Her main interests are Information Retrieval, Search Engines, Web Based Support Systems, Web Intelligence and Social Networks.

Lamine Bouguroua: He is a Research Associate Professor at school of Computer Science and Engineering ESIGETEL (Ecole Supérieure d'Informatique et Génie des Télécommunications), Villejuif, France. He received the Ph.D. degree in Sciences from the University of Paris XII, Paris, France, in March 2007. His research interests include programming languages, software architecture, object-oriented software systems, program analysis, scheduling, embedded systems, real time systems and fault tolerance. Today, his activity is concerned with scheduling and fault tolerance in Real-Time Systems, social network, multi-agent system.

INTENTIONAL BLANK

APPLYING A SYSTEMATIC REVIEW ON ADAPTIVE SECURITY FOR DSPL

Mohamed AMOUD, Ounsa ROUDIES

SIWeb Team - École Mohammadia d'Ingénieurs (EMI)
Mohammed V University in Rabat. Morocco
amoudmohamed@gmail.com , roudies@gmail.com

ABSTRACT

Providing security and privacy to Dynamic Software Product Lines (DSPL) is very challenging. DSPL is becoming the system with high vulnerability in which the security is a difficult task and critical for it to operate. Adaptive security is able to automatically select security mechanisms and their parameters at runtime in order to preserve the required security level in a changing environment. This paper presents a literature review of security adaptation approaches for DSPL, and evaluates them in terms of how well they support critical security services and what level of adaptation they achieve. This work will be done following the Systematic Review approach. Our results concluded that the research field of security approaches for DSPL is still poor of methods and metrics for evaluating and comparing different techniques.

KEYWORDS

Dynamic Software Product Lines; DSPL; Adaptive Security; Systematic Review

1. INTRODUCTION

Software product lines (SPL) have been used successfully in industry for building families of systems of related products, maximizing reuse, and exploiting their variable and configurable options.

DSPL extend the concept of conventional SPL by enabling software-variant generation at runtime and produce software capable of adapting to such fluctuations. In contrast with traditional SPLs, DSPL bind variation points at runtime, when software is launched to adapt to the current environment, as well as during operation to adapt to changes in the environment. Building a product line that dynamically adapts itself to changing requirements implies a deployment of the product configuration at runtime. It also means that the system requires monitoring capabilities for detecting changes in the environment. As a response to these changes, the system adapts by triggering a change in its configuration, providing context-relevant services or meeting quality requirements. Dynamic software reconfiguration is concerned with changing the application configuration at runtime after it has been deployed.

From the security point of view, dynamically changing DSPLs are a challenge, as static security mechanisms are not able to offer an optimal security level for the varying situations. Moreover, it

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 79–89, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61007

is impossible at design-time to anticipate all situations in which a DSPL application will be utilized. These challenges cause a need for self-adaptive security, which is able to select security mechanisms and tune their parameters at runtime.

Currently, several security adaptation approaches exist. On the one hand, approaches concentrate on adapting a particular security mechanism or supporting a specific security attribute. On the other hand, some approaches are generic; that is, they support different attributes and mechanisms. Hence, it is difficult to select the most suitable adaptation approach for different usages. Moreover, it is difficult to know what research steps are needed in the future.

The objective of this paper is to give an overview of the state of the art in the adaptive security issues for a DSPL by doing a Systematic Literature Review (SLR) on simple and clear question in this regard. In particular, we identify and compare different security adaptation approaches for DSPL, and evaluate them in terms of how well they support critical security services and what level of adaptation they achieve.

In the section 2 we describe our method for conducting the review. Results are presented in Section 3. Section 4 answers our questions. Finally, the Conclusion and future work Section close the paper.

2. METHOD

The aim of this study is identifying and comparing different security adaptation approaches for DSPL. We used guidelines proposed by Barbara Kitchenham [1] for performing our study. The main steps are explained in the next parts of this section.

2.1. Systematic Literature Reviews (SLR)

Systematic Literature Reviews (SLR) is a rigorous method for assessing, reviewing and aggregating research results. Unlike an ordinary literature review consisting of an annotated bibliography, SLR analyzes existing literature with reference to specific research questions on a topic of interest. Furthermore, it can be considered as much more effort prone than an ordinary literature survey.

2.2. Research Questions

RQ1: What is the focus of research in adaptive security of DSPL?

RQ2: What are the claimed benefits of self-adaptive security in DSPL and what are the tradeoffs implied by self-adaptive?

RQ3: how can DSPLs autonomously evaluate changes and threats in their environment in order to adaptively reconfigure themselves?

RQ4: What are the limitations of the existing approaches, and interesting areas for future research?

Regarding to RQ1, We were looking for researches and case studies need to get insight in the research trends in adaptive security of DSPL, providing context for the study.

Regarding to RQ2, it was important for us to know the claims associated with adaptive security, the evidence that exists for these claims and the tradeoffs implied by this adaptive security.

Regarding to RQ3, we want to know how applications with stringent safety requirements require security mechanisms that reduce human intervention when DSPL features change.

The goal of RQ4 is to help deriving conclusions from the study.

2.3. Research Process

Our search process for review was based on online searching in famous online databases which are addressed as table1. Since these databases cover almost all major journals and conference proceedings, manually review of journal was not required. Review has been carried on by mean of search facilities in these databases and using appropriate logical expressions. In first stage, our focus was on title and abstract of articles found in search process and select appropriate and relevant studies. If there was any doubt, our decision was based on reviewing it at one glance.

TABLE 1. STUDIES RESOURCE

Source	**Address**
Scopus	www.scopus.com
IEEE Xplore	ieeexplore.ieee.org
ACM Digital Library	Portal.acm.org
Springer Link	www.springerlink.com
Science Direct	www.sciencedirect.com

2.4. Inclusion and Exclusion Criteria

Our primary goal is to understand the claims and supporting evidence of adaptive security in DSPL, we excluded papers about theoretical aspects, as well as surveys and roadmap papers. We also excluded short papers of 1 or 2 pages.

There were some papers that were relevant to our study indirectly in our defined process. This will strengthen our review, because all relevant documents were included and our review covered sufficiently direct and indirect studies in this research.

All studies are assessed through a quality check, which is an inherent part of a thorough literature study. Checking the originality and quality of the studies is important for data synthesis and interpretation of results later on.

2.5. Quality Assessment

For assessing studies we defined the following questions:

QA1: Does study agree with existence of the focus of research in adaptive security for DSPL?

QA2: Does the security of DSPL recognise the need for adaptation?

QA3: Does study report any similar practices in the claimed benefits of adaptive security in DSPL and the tradeoffs implied by self-adaptive security?

QA4: Does study report show how to face securely unexpected risks and activate appropriate countermeasures to respond to new threats?

QA5: Is it possible to avoid specifying all the behaviours in advance for Autonomous and Adaptive Security?

We scored questions as bellow:

QA1. Y (Yes) study explicitly agrees with existence of any objectives; P (Partially) study implicitly agrees and N (No) study disagrees with existence of any objectives.

QA2. Y, study explicitly agrees with existence of any needs; P, study implicitly agrees and N, there is no need for adaptation.

QA3. Y, the authors address one or more similar practices; P, some of the ones practices could be tailored and customized in the second and N, there is no similar and adaptable practices in them.

QA4. Y, the authors report provide sufficient arguments; P, so not enough and N, there is no argument.

QA5. Y, study addresses possibility to avoid specifying all the behaviours in advance for Autonomous and Adaptive Security; P, study partly agrees (or implicitly) and N, there is no possibility.

We defined Y=1, P=0.5 and N=0 or Unknown where information is not clearly specified. All authors assessed every article and if there is no agreement in scoring, we discussed enough to reach agreement.

We defined Y=1, P=0.5 and N=0 or Unknown where information is not clearly specified. All authors assessed every article and if there is no agreement in scoring, we discussed enough to reach agreement.

2.6. Data Collection

These data were extracted from each article:

- The full source and references
- The author(s) information and details
- Research issues
- Main ideas

All articles were reviewed and data was extracted and checked. This idea was chosen for better consistency in reviewing all papers and improving quality of review.

2.7. Data Analysis

Our collect data was organized to address:

- Whether study agrees with existence of the objectives of adaptive security for DSPL or not? (Addressing RQ1)

- Whether study agrees with existence of any needs for adaptation or no? (Addressing RQ1, RQ2)

- Whether study mentions similar practice/concept in either methods or no? (Addressing RQ3)

- Whether study provides sufficient arguments to face securely unexpected risks and activate appropriate countermeasures or no? (Addressing RQ2 and RQ4)

- Whether study agrees with existence of possibility to avoid specifying all the behaviours in advance for Autonomous and Adaptive Security or not? (Addressing QR5)

- Whether authors believe that this area is promising or no? (Addressing QR5)

3. RESULTS

In this section we explain results of our review.

3.1 Search Results

Table 2 shows the results of our selection procedure. In this table, results of searching in all databases are provided, but, some of the studies were repeated in more than one online database, so, final number of unique studies selected for our review was distinguished after elimination of repeated articles. Final selected studies are listed in table 3.

3.2 Quality Evaluation of Studies

During this phase, we found that some of the selected articles discussing security in general or only the SPL, but, they do not provide any valuable information to our research, so, we decided to delete them from scope of our study. Assessment of each study was done by means of criteria explained in section 2.4 and the scores for each of them are shown in table 4.

3.3 Quality Factors

For assessing results of our quality questions, we use average of total scores. This average is useful for some questions, but it is not useful for some other. For instance, we cannot answer the question about possibility of integration with average of scores because of the nature of the question; instead, we use negative ideas for rejecting possibility.

TABLE 2. RESULTS OF STUDY SELECTION PROCEDURE

Source	Search Results	Selected Studies
Scopus	91	7
IEEE Xplore	85	10
ACM Digital Library	22	5
Springer Link	36	7
Science Direct	04	1
Total	**238**	**30**
Repeated articles		12
Finally selected articles		**18**

TABLE 3. SELECTED STUDIES FOR CONDUCTING REVIEW

ID	Title	Authors	Main Topic	Year
S1	Strategies for Variability Transformation at Run-time	O. Haugen and al. [2]	the security of DSPL recognises the need for adaptation	2009
S2	Self-Adaptive Software: Landscape and Research Challenges	M. Salehie and al. [3]	Self-protecting is the capability of recovering from their effects and detecting security breaches	2009
S3	Security Requirements Engineering Framework for Software Product Lines	D. Mellado and al. [4]	To describe a security requirements engineering in order to facilitate the development of secure SPLs	2010
S4	A Security Requirements Engineering Tool For Domain Engineering In SPL	J. Rodríguez and al. [5]	how to provide automated support through which to facilitate the application of the security quality requirements engineering process for SPL	2011
S5	Claims and Supporting Evidence for Self-Adaptive Systems: A Literature Study	D. Weyns and al. [6]	Claims versus the tradeoffs of adaptive security	2012
S6	Non-functional Properties in Software Product Lines - taxonomy for classification	M. Noorian [7]	The adaptive security attribute should be measured at runtime.	2012
S7	Automated Planning for Feature Model Configuration based on functional and non-functional requirements	S. Soltani and al. [8]	Adaptive security is a non-functional requirements in DSPL	2012

S8	A Systematic Review of Model-Driven Security	H. Nguyen [9]	How to improve the productivity of the development process and quality of the resulting secure systems	2013
S9	Runtime Monitoring and Auditing of self-adaptive systems	D. H. Carmo and al. [10]	How to avoid specifying all the behaviours in advance for Autonomous and Adaptive Security	2013
S10	Comparison of Adaptive Information Security Approaches	A. Evesti and al. [11]	Limitations and prospects of adaptive security	2013
S11	Architecture and Knowledge-Driven Self-Adaptive Security in smart space	A. Evesti and al. [12]	Self-adaptive security as an applicable solution to anticipate all the possible changes at design-time.	2013
S12	An overview of Dynamic Software Product Line architectures and techniques	R. Capilla and al. [13]	Challenges and solutions are necessary to support runtime variability and adaptive security mechanisms in DSPL models and software architectures.	2014
S13	A Systematic Survey of Self-Protecting Software Systems	E. Yuan and al. [14]	autonomic systems capable of detecting and mitigating security threats at runtime	2014
S14	Policy -Based Language for Autonomous and Adaptive Security	F. Cuppens [15]	how to simultaneously address both adaptive and autonomy in DSPL	2014
S15	Dynamic Reconfiguration of Security Policies in Wireless Sensor Networks	Mónica Pinto and al. [16]	self-protection solution based on the combination of dynamic adaptation and reconfiguration of security	2015
S16	Representing and Configuring Security Variability in Software Product Lines	V. Myllärniemi [17]	Security variability can be represented and distinguished as countermeasures	2015
S17	Security Systems Engineering Approach in Evaluating Commercial and Open Source Software Products	Jesus Abelarde [18]	The amount of security resources and time necessary to accommodate proper security evaluations is underestimated.	2016
S18	Trustworthy variant derivation with translation validation for safety critical product lines	J. Almendros-Jiménez and al. [19]	Propose a general technique of checking correctness through translation validation to automatically verify runs of a variant derivation tool.	2016

TABLE 4. QUALITY EVALUATION

Source	QA1	QA2	QA3	QA4	QA5
1	Y	P	N	Y	Y
2	Y	Y	P	Y	Y
3	P	Y	Y	P	Y
4	P	Y	P	Y	P
5	P	Y	Y	P	Y
6	Y	Y	Y	Y	Y
7	N	P	P	Y	Y
8	P	Y	Y	N	Y
9	Y	P	P	Y	Y
10	P	N	P	Y	P
11	N	P	P	Y	P
12	P	Y	N	Y	P
13	N	P	P	Y	Y
14	P	N	Y	Y	Y
15	P	Y	N	P	Y
16	N	P	Y	Y	P
17	Y	Y	P	P	Y
18	Y	Y	Y	Y	P
Average	**0,58**	**0,72**	**0,61**	**0,83**	**0,83**

4. DISCUSSION

In this part, the answers to our study questions will be discussed.

4.1. What is the Focus of Research in Adaptive Security of DSPL?

Most of the articles agree that there are the objectives of research in adaptive security for DSPL. By reviewing them, it seems that this research focus is derived from the following concerns: category of the study, subject, concrete focus and application domain. Overall, fifty eight percent of the studies focus on one or more activities of adaptive security (monitoring, analyzing, planning, execution), runtime models, multiple control loops and on reflection [10]- [19].

More than two thirds of the articles agree that the security of DSPL recognise the need for adaptation in order to achieve the required security level. On the one hand, a survey by D. Weyns et al. [6] reveals that the existing security approaches DSPL are not generic, but rather approaches focus on specific security objectives. Furthermore, a study of Yuan et al. [14] compares over 30 self-protection approaches and shows that most existing approaches focus on the part of the adaptive control loop, instead of covering the entire adaptation loop.In the adaptive security approaches, such as: Self-Adaptive Security in smart space [12], Self-Adaptive Software [3], Strategies for Variability Transformation at Run-time [2], A Security Requirements Engineering Tool For Domain Engineering In SPL [5] and Runtime Monitoring and Auditing of self-adaptive systems [10], Architectural Approach for Self-managing Security Services in [17]-[18], authors notice that any of these approaches support all security objectives but concentrate on specific and pre-selected objectives.

4.2. What are the Claimed Benefits of Self-Adaptive Security in DSPL and what are the Tradeoffs Implied by Self-Adaptive?

Eight studies examine the claims versus the tradeoffs of adaptive security and clearly demonstrate that the researches mainly report on claimed benefits, while little attention is given to the implications of adaptive security. It is remarkable that the efficiency/performance ratio is almost the only quality attribute with a negative effect due to the adaptive security [11]. We also evaluated the type of claims that have been made to the quality attributes and found that the dominant demand is improving quality attributes of the software. The main reported tradeoff implied by the adaptive security is the top performance [6]-[18].

4.3. How can DSPLs Autonomously Evaluate Changes and Threats in their Environment to Adaptively Reconfigure Themselves?

Most of the articles agree that it is possible to have an autonomous and adaptive security in DSPL [10]-[15]-[19]. Mónica Pinto and al. [16] present an approach to building adaptive security at runtime. They extend SPLs by adding the ability to automatically derive changed configurations by monitoring the context, and to automatically reconfigure the security application while it is running. The adaptation platform of this approach provides a conceptual model and reference architecture for adaptive system. A. Evesti and al. [12] address SPL that allow mobile devices in smart space to download software configurations on-demand. When a device enters a particular context, the application provider service must deduce and create a variant for the device. As devices enter a context, their unique capabilities must be discovered and dealt with efficiently and correctly. D. H. Carmo and al. [10] present a new approach where they meet challenges in adaptive security construction and execution by combining certain aspect oriented and model driven techniques in order to deal with complexity through abstractions used both to specify the dynamic variability at design time and to manage run time adaptations.

4.4. What are the Limitations of the Existing Approaches and Interesting Areas for Future Research?

The issue of security in SPL is long, but most solutions are based on the assumption that the SPL is a closed environment. Given current trends, where the SPL is dynamic and open system, these solutions are not sufficient to ensure the adaptive security [4]. Although there are researchers working in this field and solutions are provided to be better, but the mechanism of adaptive security for DSPL is not yet mature. In addition, the existing security solutions are based on the features of the current DSPL; since DSPL reveals more and more new features that may be supported in the future; the adaptive security mechanism has to be modernized and new security issues have to be identified [1]-[9]-[11].

The majority of articles are optimistic about the potential prospects of this promising area. The research should focus on the following areas: -1- Policy, model and design of the security architecture -2- Securing the management and sharing of knowledge...etc.

5. Conclusion and Future Work

The objective of this literature study was to summarize existing research on engineering self-adaptive software systems and shed light on the claimed benefits and provided evidence of adaptive security in DSPL. The study shows that the existing adaptive security approaches widely cover the information gathering. However, comparative approaches do not describe how to decide on a method for performing adaptive security DSPL or how to provide knowledge input for adapting security. Therefore, these areas of research are promising.

As an emerging topic, we expect that promising new research will bring better and integrated a self-adaptive security solution for Mobile Devices based on the combination of the MAPE- K reference model and DSPL approach.

REFERENCES

[1] B. Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey and Stephen Linkman,” Systematic literature reviews in software engineering- A Systematic literature reviews”, Keele University 2008

[2] O. Haugen C. Cetina, X. Zhang, F. Fleurey, V. Pelechano : ‘’ Strategies for variability transformation at run-time’’, SPLC '09 Proceedings of the 13th International Software Product Line Conference, Pages 61-70, Carnegie Mellon University Pittsburgh, PA, USA, 2009

[3] M. Salehie and L. Tahvildari, “Self-adaptive software: Landscape and research challenges,” ACM TAAS, vol. 4, 2009.

[4] D. Mellado, E. Fernández-Medina, M. Piattini: ‘’ Security Requirements Engineering Framework for Software Product Lines’’, Information and Software Technology, 2010. Volume 52: p. 1094-1117. Oct. 2010

[5] Jesús Rodríguez, Eduardo Fernández-Medina, Mario Piattini, Daniel Mellado:’’ A Security Requirements Engineering Tool for Domain Engineering in Software Product Lines ’’, part of the ESFINGE Project of the Ministry of Science and Innovation (Spain), 2011

[6] D. Weyns1, M. Usman Iftikhar, Sam Malek, J. Andersson, ‘’ Claims and Supporting Evidence for Self-Adaptive Systems: A Literature Study’’, 2012 ICSE Workshop on SEAMS, Zurich, pp.89-98, 4-5 June 2012.

[7] M. Noorian, E. Bagheri, W. Du:’’ Non-functional Properties in Software Product Lines: A Taxonomy for Classification’’, In Proceedings of "SEKE'12", Peges 663-667, 2012

[8] S. Soltani, M. Asadi, D. Gasevic, M. Hatala, E. Bagheri :’’ Automated planning for feature model configuration based on functional and non-functional requirements’’, SPLC '12 Proceedings of the 16th International Software Product Line Conference - Volume 1 Pages 56-65 ACM New York, NY, USA, 2012

[9] Phu H. Nguyen:’’ A Systematic Review of Model-Driven Security’’, Software Engineering Conference (APSEC, 2013 20th Asia-Pacific (Volume: 1) , IEEE, Univ. of Luxembourg, Dec. 2013

[10] D. H. Carmo, Sergio T. Carvalho, Leonardo G. P. Murta, Orlando Loques, ‘’ Runtime Monitoring and Auditing of Self-Adaptive Systems’’,8th IEEE International Conference on Global Software Engineering (ICGSE), Brazil 2013.

[11] A. Evesti; E. Ovaska, ‘’ Comparison of Adaptive Information Security Approaches’’, ISRN Artificial Intelligence, Article ID 482949, 18 pages. Volume 2013.

[12] A. Evesti. ‘’ Adaptive security in smart spaces’’. PhD's thesis, University of Oulu, on the 31st of January 2014.

[13] Capilla, R., et al., ‘’An overview of Dynamic Software Product Line architectures and techniques: Observations from research and industry’’. J. Syst. Software, 2014,

[14] E. Yuan, N. Esfahani, S. Malek:’’ A Systematic Survey of Self-Protecting Software Systems’’, ACM Transactions on Autonomous and Adaptive Systems (TAAS), Volume 8 Issue 4, New York, NY, USA . 2014.

[15] F. Cuppens:’’ Policy -Based Language for Autonomous and Adaptive Security’’, the Concordia Institute for Information Systems Engineering, Canada, Jan. 2014

[16] Mónica Pinto and al.,’’ Dynamic Reconfiguration of Security Policies in Wireless Sensor Networks’’, Sensors 2015, 15, 5251-5280; doi:10.3390/s150305251, 2015

[17] Varvana Myllärniemi : ‘’Representing and Configuring Security Variability in Software Product Lines’’, Proceedings of the 11th International ACM SIGSOFT Conference on Quality of Software Architectures, pp: 1-10, ACM New York, NY, USA, 2015

[18] Jesus Abelarde, ‘’ Security Systems Engineering Approach in Evaluating Commercial and Open Source Software Products’’, SANS Institute InfoSec Reading Room, January 25, 2016

[19] Jesús M. Almendros-Jiménez, Luis Iribarne, Jesús López-Fernández, Ángel Mora-Segura, ‘’ Trustworthy variant derivation with translation validation for safety critical product lines’’, Journal of Logical and Algebraic Methods in Programming, Vol. 85, Issue 2, February 2016

INTENTIONAL BLANK

KEY MANAGEMENT SCHEME FOR SECURE GROUP COMMUNICATION IN WSN WITH MULTIPLE GROUPS

H.S.Annapurna[1] and M.Siddappa[2]

[1]Dept. of Computer Science & Engg.,
Sri Siddhartha Academy of Higher Education, Tumakuru, India.
hsassit@gmail.com
[2]Dept. of Computer Science & Engg.,
Sri Siddhartha Institute of Technology, Tumakuru, India.
siddappa.p@gmail.com

ABSTRACT

Security is one of the inherent challenges in the area of Wireless Sensor Network (WSN). At present, majority of the security protocols involve massive iterations and complex steps of encryptions thereby giving rise to degradation of quality of service. Many WSN applications are based on secure group communication. In this paper, we have proposed a scheme for secure group key management with simultaneous multiple groups. The scheme uses a key-based approach for managing the groups and we show that membership change events can be handled with less storage, communication and computation cost. The scheme also offers authentication to the messages communicated within and among the groups.

KEYWORDS

Group key, Key management, Sensing Node, Secure Group Communication, Key Tree.

1. INTRODUCTION

Wireless Sensor Network is a collection of sensor nodes with limited capabilities in terms of battery, computation, storage etc. The data that flows in among the sensor nodes in WSN consists of physically captured data from the readings of sensors, a mobile code, security using key management techniques, and location information of the sensor nodes. Owing to the lesser amount of obtainable of computational origin in the miniature sensor nodes and wireless communication social, WSN endures from probable security threat aspects [1]. There are basically two types of attacks in sensor network, e.g., active and passive attacks [2]. The malicious nodes can enhance their attacking capabilities by intruding the private information from mobile codes as well as by accessing the information pertaining to the positioning of the nodes [3]. Using various eavesdropping techniques, it is possible for the malicious node to incorporate malicious programs on the mobile code and thereby spreading the malicious mobile code in the entire network. The malicious node can also use the position information to identify

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 91–101, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61008

the best node to invoke their attacks thereby potentially making security breach. Owing to the wireless medium of communication in WSN, it is very challenging task to identify the malicious nodes and design a security policy to deny the access in the network. The malicious nodes are quite capable enough to access the entire network using potential computers and sophisticated communication equipments. The malicious nodes can also seed themselves in the network environment without even getting caught [4]. It is said that sink is considered as the most reliable core of the wireless sensor network that stores significant information about the security protocols, readings of sensors, and routing information. These are very critical in group communication. In small scale sensor network, it is easier to capture the data, process it, and forward to sink. But in random and dynamic network of large size, it usually doesn't go by single hop communication. The nodes are formulated in groups, where each group member interacts with other group member to forward the processed data from one point to another. The process of data aggregation completely fails without group communication. Hence, it is very important that a robust security technique is to be developed to address the security issues in group communication system in WSN. Cryptography [5] is the most frequently adopted technique to incorporate security while performing group communication in WSN.

However, conventional cryptographic algorithms like SHA, AES, although have good security features, suffer from limitations too. Hence, keeping all these issues in mind, the paper introduces a scheme for secure group key communication with multiple groups. Remaining part of the paper is ordered as follows: Section 2 discusses background of research work followed by key management scheme in section 3. Authenticated group communication is presented in section 4 and section 5 summarizes the paper.

2. BACKGROUND

The study towards secure group communication is more than a decade old and there are various techniques that have been introduced by the various researchers. This section discusses some of the recent studies found in standard research manuscript that focuses on i) secure group communication and ii) key distribution mechanism.

Cheikhrouhou et al. [6] have discussed a protocol for ensuring secure group communication using elliptical curve cryptography over ring based topology of wireless sensor network. The authors have discussed their outcomes considering storage cost which was found to be efficient compared to existing techniques. However, the limitation of the scheme is the dependency of key storage of size 160 bits. Wang et al. [7] have proposed a predistribution policy considering hexagonal grids consisting of groups and keys. Miettinen et al. [8] have presented a security protocol by incorporating an authenticated pairing system based on key context. Furtak and Chudzikiewicz [9] have used asymmetric key pair as well as electronic signature to provide secure authentication in wireless sensor network. Xi et al. [10] have presented a key estimating process that is done in faster manner as compared to attacker. However, various attackers have various patterns of generating attacks, the authors have not discrete mentioned the names of the attack. Moreover the outcomes of the study were not found to be benchmarked.

Hence, it can be seen that there exists various security protocols in the research papers with advantages and limitations. The prime trade-off found in all the study is dependency of broadcasting the key. We comment that broadcasting of the key is very sensitive operation and is highly prone to capture if proper encryption scheme is not implemented. Another trade-off found is majority of the schemes are based on enhancement of conventional cryptographic scheme with less novelty in mathematical approaches. The third trade-off seen in all the studies is about the key sizes, which is 128, 216, 160, or 512 bits. Although the key sizes seem to be smaller but as majority of the existing approaches store this, grossly the sizes of the matrix holding the keys becomes eventually larger.

Many schemes for group key management have been proposed in the literature for WSN [11, 12, 13, 14]. But all these schemes consider a single group communication scenario. Aparna et al. [15] have discussed a scheme for secure group communication with multiple groups which is based on logical key trees. A combination of key-based and secret share-based approach is used for managing the group keys. Purushothama et al. [16] have proposed a group key management scheme for simultaneous multiple groups with overlapped membership. The scheme is based on key-user tree structure with substantial reduction in storage and rekeying cost. But both of these schemes are proposed for conventional networks. In this paper, we have proposed a scheme for secure group communication for WSN with multiple groups.

3. Key Management Scheme

We propose a scheme for group key management with multiple groups. A group consists of n sensing nodes and there are at most m simultaneous groups that need to be established. The nodes are numbered $s_1, s_2 \ldots s_n$ and groups are numbered $G_1, G_2 \ldots G_m$. A logical tree in constructed for each group G_i, for $i = 1,2 \ldots m$. The height of the tree for group G_i depends on the number of sensing nodes in G_i and it is $log_2 k$ if there are k $(k \leq n)$ nodes in the tree. The tree is maintained by the central node. It constructs a separate key tree for each group. Each sensing node shares a private key with the central node which is used for confidential communication. The group key (GK) is at the root of the tree and is used for confidential communication with the group members. An interior node with two child nodes forms a subgroup and keys associated with the subgroup are called secondary keys. These keys are named either k_{ij} for $j=1,2 \ldots m$ or k_{p-l} depending on whether they have two child nodes or one child node. The key is named k_{ij} if it is the root of the subtree with leftmost child s_i and rightmost child s_j and it is named k_{p-l} if it is the root of the subtree with one child node (left or right). k_p is the leftmost or rightmost child (whichever exists) of this subtree and l is the level number. Secondary keys (keys along the path excluding group key and private key) are used to encrypt new group key. Next we discuss group formation phase followed by computation and distribution of group key.

3.1 Group Formation Phase

The proposed scheme uses Logical Key Hierarchy (LKH) scheme [17] and a binary tree with two keys at each level. The central node is responsible for group formation and rekeying operations. It assigns each sensing node a unique id (*UID*) which is a binary string of length p where $p = \lceil log_2{}^n \rceil$ where n is the number of sensing nodes. A sensing node $s_{i,}$ which wishes to join the group G_j sends a join request of the form $JOIN\ (UID_i, G_j)$ to the central node where UID_i is the unique identification number of s_i. A node wishing to join more than one group sends individual join request to each group. A node can send a request to join more than one group in which case it will be a member of more than one key tree.

3.2 Rekeying Strategies and Protocols

We use key based approach for managing group keys and secondary keys. Whenever a node is compromised, it is evicted from the group(s) to which it belongs. Similarly, whenever a new node enters a monitoring area it is added to the group. In either case there is a membership change and hence the group key needs to be changed to prevent a new group member from reading past communications and old group member from reading current and future communications. Whenever there is a membership change, the central node updates the key tree, computes the new group key and distributes it to the existing nodes securely. In the following subsections we discuss the protocols for joining and leaving a group(s) represented by key tree(s).

3.2.1 Joining a key tree

A new node s_i $(1 \leq i \leq n)$ wanting to join a group G_j $(1 \leq j \leq m)$, sends a join request of the form *JOIN* (UID_i, G_j) to the central node (CN). Upon receiving this join request from s_i, the CN checks the node's identity and whether it is allowed to join the group G_j. If so, the CN updates the key tree by creating a node for s_i and ensures backward secrecy by changing the keys along the path from root till its parent and communicating them to appropriate users. The CN computes new group key GK_j' for group G_j and sends it to current members of G_j by encrypting it with old group key GK_j. For the new node, the CN sends the keys along the path by encrypting them with PK_i, private key of s_i. For example, consider an initial key tree with multiple groups shown Fig.1. In the figure s-nodes represent the sensing nodes and nodes labeled *PK* from PK_1 to PK_{11} represent the private keys of s_1 to s_{11}. The k-nodes represent the secondary keys and root nodes labeled GK_1, GK_2, GK_3 represent the group keys.

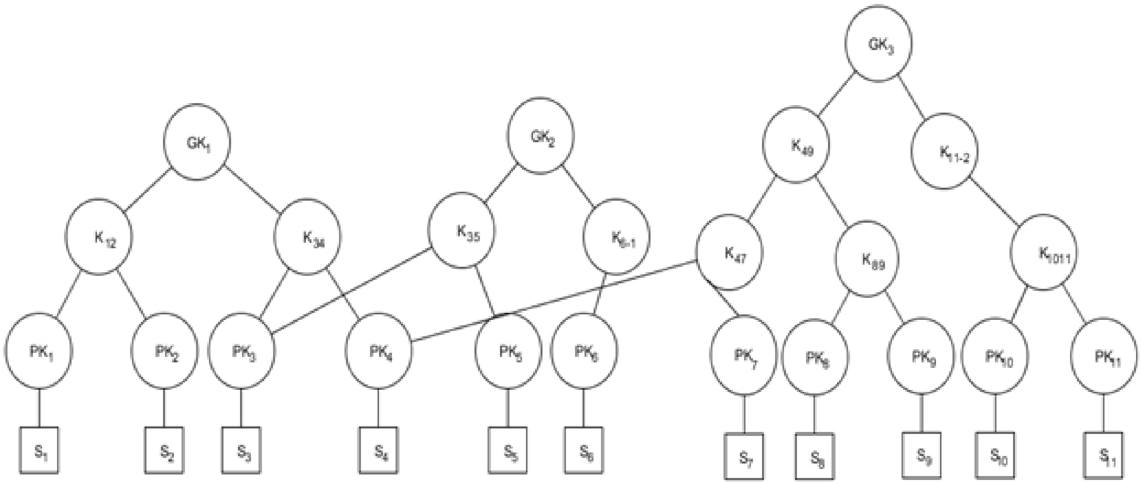


Fig 1 : Initial key tree with 3 groups

There are 3 simultaneous groups G_1, G_2, G_3 with four members, s_1, s_2, s_3, s_4 in G_1, 3 members s_3, s_5, s_6 in G_2 and 6 members s_4, s_7, s_8, s_9, s_{10}, s_{11} in G_3.

Now, suppose a new node s_{12} wants to join group G_3 in Fig.1, it sends a join request *JOIN* (UID_{12}, G_3) to the CN. If the requesting node is allowed to join, the CN updates the key tree as shown in Fig.2. The keys that must be changed are $K_{11\text{-}2}$ and GK_3. The CN changes $K_{11\text{-}2}$ to K_{1012}

and randomly selects a new group key $\mathrm{GK}_3^{'}$. The changed keys and the new group key are communicated to appropriate nodes by sending the following rekeying messages:

1. CN →$\{s_{12}\}$: $E_{PK_{12}}$ $(K_{12\text{-}1}, K_{1012}, G\mathrm{K}_3^{'})$
2. CN →$\{s_{10}, s_{11}\}$: $E_{K_{1011}}(K_{1012})$, $E_{GK_3}(G\mathrm{K}_3^{'})$
3. CN →$\{s_4, s_7, s_8, s_9\}$: $E_{GK_3}(G\mathrm{K}_3^{'})$

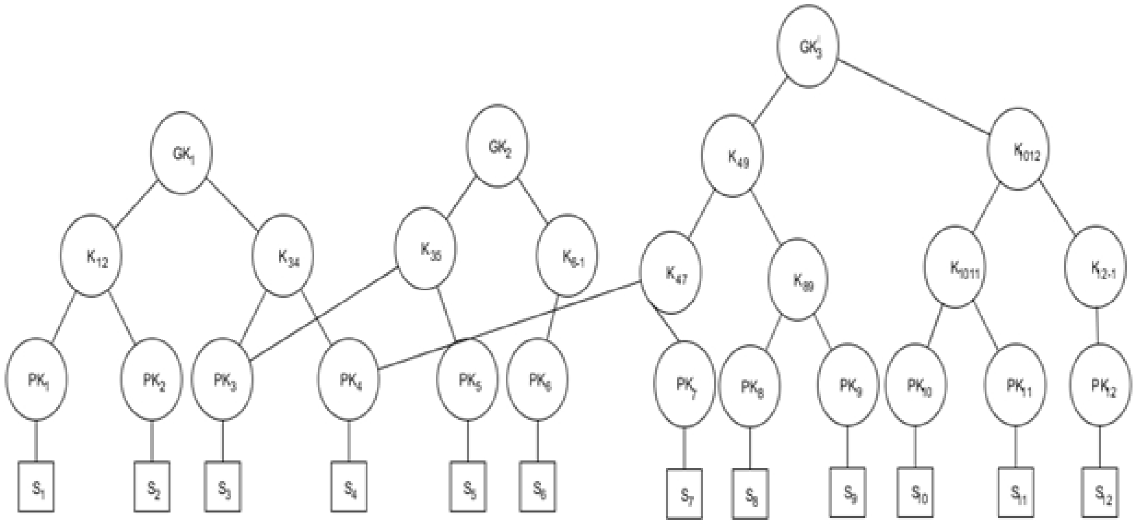


Fig.2 Key Tree after s_{12} joins G_3.

Suppose, now s_{13} wants to join both G_2 and G_3, it sends join requests to both the groups. It sends *JOIN*(UID_{13}, G_2) and *JOIN*(UID_{13}, G_3) to the CN. If the requested node is allowed to join G_2 and G_3, the CN updates the key tree as shown in Fig.3. The keys that must be changed are $K_{12\text{-}1}$, K_{1012}, $\mathrm{GK}_3^{'}$, $K_{6\text{-}1}$ and GK_2. The new keys are sent to appropriate nodes by generating the following rekeying messages:

1. CN →$\{s_{13}\}$: $E_{PK_{13}}(K_{613}$, $\mathrm{GK}_2^{'}$, K_{1213}, K_{1013}, $\mathrm{GK}_3^{''}$')
2. CN →$\{s_6\}$: E_{GK_2} ($\mathrm{GK}_2^{'}$, K_{613})
3. CN →$\{s_3, s_5\}$: E_{GK_2} ($\mathrm{GK}_2^{'}$)
4. CN →$\{s_{12}\}$: $\mathrm{E}_{\mathrm{PK}_{12}}$ (K_{1213}), $E_{K_{1213}}$ (K_{1013}), $\mathrm{E}_{\mathrm{GK}_3^{'}}$ ($\mathrm{GK}_3^{''}$)

5. CN →$\{s_{10}, s_{11}\}$: $E_{K_{1011}}(K_{1013})$, $E_{GK'_3}(GK''_3)$

6. CN →$\{s_4, s_7, s_8, s_9\}$: $E_{GK'_3}(GK''_3)$

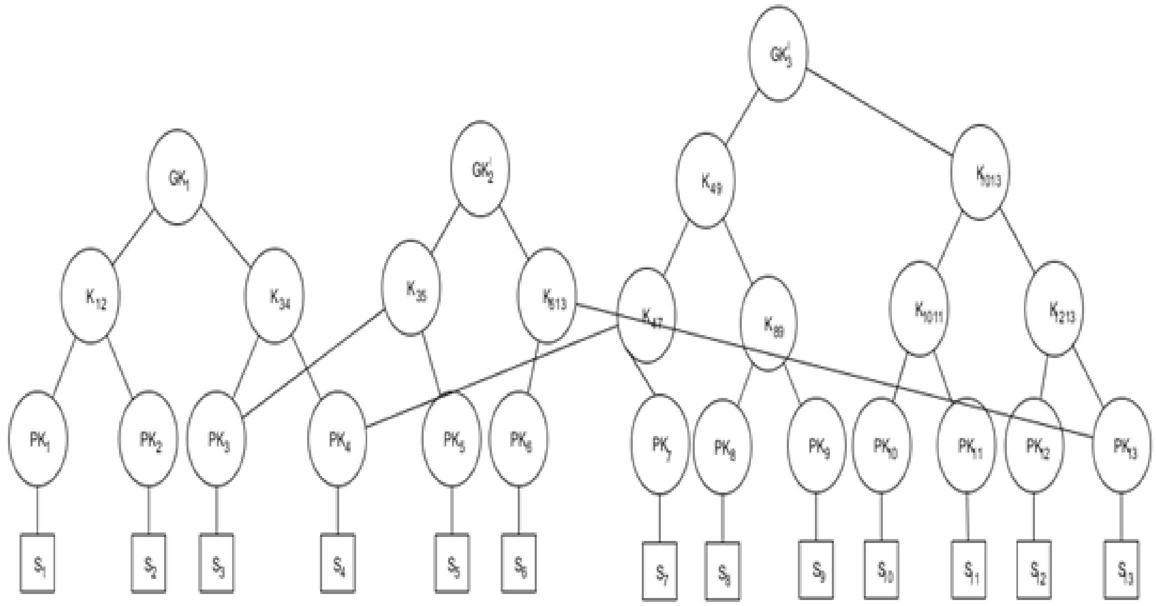


Fig. 3 : Key Tree after s_{13} joins G_2 and G_3.

When a member joins a group *Gj* with *k* members, then at most log_2k keys have to be changed, $\lceil 2log_2k \rceil$ encryptions are required and $\lceil log_2k \rceil$ rekey messages have to be built to communicate the changed keys to the appropriate members of the group. For a member joining *i* number of groups, the number of keys to be changed, number of encryptions required and number of rekey messages to be sent are$\sum_{j=1}^{i} log_2 n_j$, $2\sum_{j=1}^{i} log_2 n_j$,$\sum_{j=1}^{i} log_2 n_j$ respectively where n_j is the number of nodes in group *Gj*.

3.2.2 Leaving a key tree

After a node leaves a group, current group key can no longer be used for future communications and a new group key must be selected and distributed securely to the remaining group members. In addition, all other keys which are known to the leaving node must also be changed to ensure forward secrecy. A leaving node may be a member of a single group or more than one group. Depending on how many groups it belongs to and how many groups it wants to leave, the CN updates the key tree accordingly. Suppose a node s_4 which is a member of G_1 and G_3 wants to leave group G_3, it sends leave request of the form *LEAVE* (UID_4, G_3) to CN. Upon receiving this request CN removes it from G_3 and changes the keys along the path as shown in Fig.4. The keys that must be changed are GK''_3, K_{47}, K_{49}. GK''_3Changes to GK'''_3, K_{47} changes to $K_{7\text{-}1}$and K_{49}changes to K_{79}. The new keys are conveyed to the existing members of the group by sending the following rekey messages :

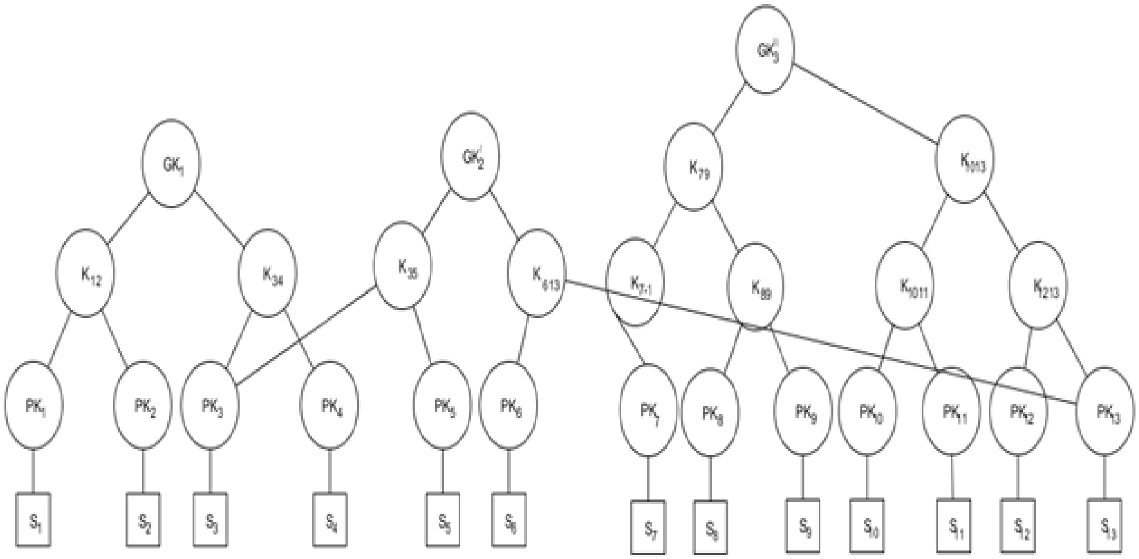


Fig. 4 : Key Tree after s_4 leaves G_3.

1. CN→$\{s_7\}$: E_{PK_7} $(K_{7\text{-}1})$, $E_{K_{7-1}}$ (K_{79}), $E_{K_{79}}$ $(\mathrm{GK}_3^{'''})$

2. CN→$\{s_8, s_9\}$: $E_{K_{89}}$ (K_{79}), $E_{K_{79}}$ $(\mathrm{GK}_3^{'''})$

3. CN→$\{s_{10}, s_{11}, s_{12}, s_{13}\}$: $\mathrm{E}_{\mathrm{K}_{1013}}$ $(\mathrm{GK}_3^{'''})$

Now, suppose s_3 wants to leave both G_1 and G_2, it sends leave request to both the groups. It sends *LEAVE* (UID_3, G_1) and *LEAVE* (UID_3, G_2) to CN. The CN removes s_3 from the trees representing G_1 and G_2. The key K_{34}, changes to $K_{4\text{-}1}$, K_{35} changes to $K_{5\text{-}1}$, GK_1 to $\mathrm{GK}_1^{'}$ and $\mathrm{GK}_2^{'}$ to $\mathrm{GK}_2^{''}$. The resulting key tree is shown in Fig. 5 below. The new keys are communicated to corresponding nodes by generating and sending the following rekeying messages :

1. CN →$\{s_4\}$: E_{PK_4} $(K_{4\text{-}1})$, $E_{K_{4-1}}$ $(\mathrm{GK}_1^{'})$

2. CN →$\{s_1, s_2\}$: $E_{K_{12}}$ $(\mathrm{GK}_1^{'})$

3. CN →$\{s_5\}$: E_{PK_5} $(K_{5\text{-}1})$, $E_{K_{5-1}}$ $(\mathrm{GK}_2^{''})$

4. CN →$\{s_6, s_{13}\}$: $E_{K_{613}}$ $(\mathrm{GK}_2^{''})$

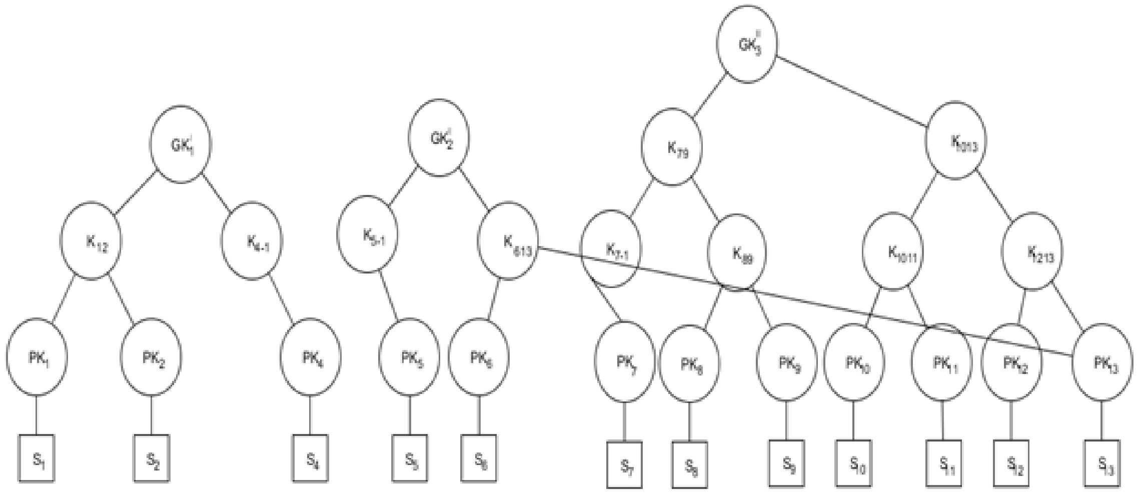


Fig.5 : Key Tree after s_3 leaves G_1 and G_2.

When a member leaves a group *Gj* with *k* members, then at most log_2k keys have to be changed, $\lceil 2log_2k \rceil$ encryptions are required and $\lceil log_2k \rceil$ rekey messages have to be sent to the appropriate members of the group to communicate the changed keys.

3.2.3 Changing the group membership

A node wishing to move from one group to another sends a move request to CN. Move request can be implemented as leave request followed by join request. For example, a node moving from G_i to $G_j$can be interpreted as leaving group G_i and joining group G_j. The CN must ensure forward secrecy for group G_i and backward secrecy for group G_j by changing the keys along the path in G_i and G_j. Consider the key tree in Fig.5. Now, suppose s_4 wants to move from G_1 to G_2, it sends a move request of the form *MOVE* (UID_4, G_1, G_2) to the CN. The CN now removes the node for s_4 from G_1 and inserts it to G_2. The resulting key tree is shown in Fig.6. CN constructs the following rekeying messages and sends to the appropriate nodes:

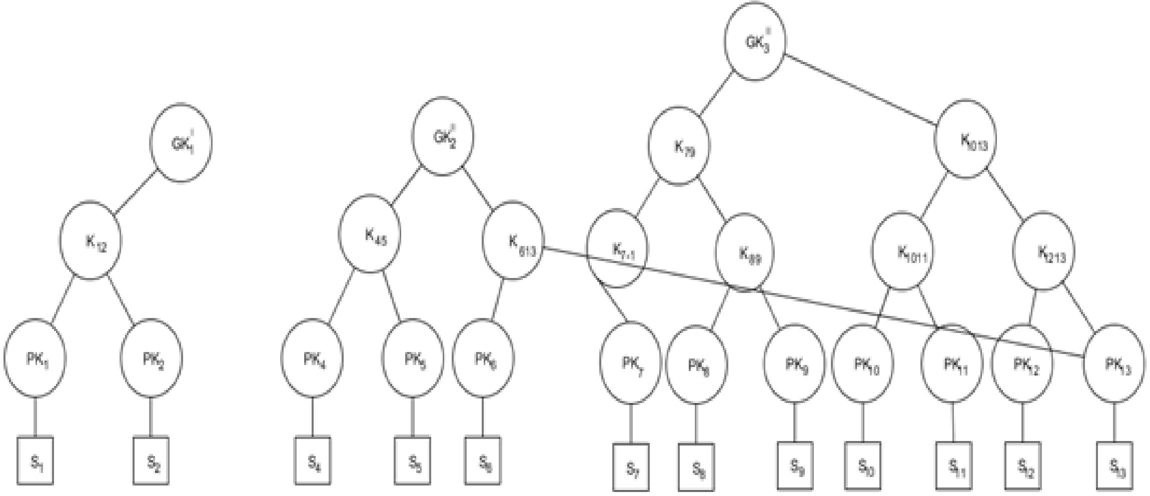


Fig.6 : Key Tree after s4 moves from G1 to G2.

1. CN $\rightarrow \{s_1, s_2\}$: $E_{K_{12}}$ ($GK_1^{''}$)

2. CN $\rightarrow \{s_4\}$: E_{PK_4} (K_{45}, $GK_2^{'''}$)

3. CN $\rightarrow \{s_5\}$: E_{PK_5} (K_{45}), $E_{K_{45}}$ ($GK_2^{'''}$)

4. CN $\rightarrow \{s_{6,} s_{13}\}$: $E_{K_{613}}$ ($GK_2^{'''}$)

For a member moving from group G_i to group G_j, the number of keys to be changed, number of encryptions required and number of rekey messages to be sent are $\lceil log_2 n_i + log_2 n_j \rceil$, $\lceil 2(log_2 n_i + log_2 n_j) \rceil$ and $\lceil log_2 n_i + log_2 n_j \rceil$ respectively where n_i and n_j are the number of nodes in groups G_i and G_j respectively.

Each member of the group needs to store $h_j - 1$ secondary keys and one group key where h_j is the height of the tree in group j for $j = 1$ to m. A node needs to store $\sum_{j=1}^{i}(h_j - 1)$ secondary keys and i number of group keys if it is a member of i number of groups.

4. VERIFYING AUTHENTICITY IN GROUP COMMUNICATION

Verifying authenticity of the sender is an important issue in secure group communication which provides protection against masquerade attack. For example, when a sender node s_i sends a message to group G_k, $(1 \leq k \leq m)$, the members of G_k must identify that the sender is s_i and it is not some other node s_j trying to impersonate s_i. In this section, we provide a protocol for authenticated group communication. When a node s_i wants to send a message to group G_k, it first sends a request to CN which includes the node's identity, group identity and a challenge C. The CN in turn sends a authentication key AK_i to s_i encrypted with private key PK_i and hash of AK_i to the members of G_k encrypted with group key GK_i. s_i, now computes hash of AK_i and sends the message M along with $H(AK_i)$ to group members encrypted with group key GK_i. Upon receiving this from s_i, group members decrypt it, compare the received $H(AK_i)$ with the one received from the CN. If both match the group members are sure of the sender and accept the message sent by s_i. Otherwise they discard the message. The use of challenge assures the group members that this is a fresh message and no old message has been replayed. Thus the protocol in Fig.7 provides authenticity as well as confidentiality in group communication. In the protocol we use the symbol || to denote concatenation operation.

1. $s_i \rightarrow$ CN : $[UID_i \,||\, G_i \,||\, C]$
2. CN$\rightarrow s_i$: E_{PK_i} $[AK_i \,||\, f(C)]$
3. CN $\rightarrow G_i$: E_{GK_i} $[H(AK_i) \,||\, f(C) \,||\, UID_i]$
4. $s_i \rightarrow G_i$: E_{GK_i} $[M \,||\, H(AK_i) \,||\, f(C) \,||\, UID_i]$

Fig.7: Protocol for authenticated group communication.

5. CONCLUSION

Secure group communication is an increasingly popular research area having received much attention in recent years. Group oriented applications in WSN demand for the security services to achieve the secure group communication. A common method is to encrypt messages with a group key so that entities outside the group cannot decode them. Therefore, key management is a fundamental building block for secure group communication systems. This paper introduces a key management scheme for WSN with multiple simultaneous groups. We have used a key-based approach for managing the groups and in case of membership change events the communication and computation costs are logarithmic in nature. The paper also provides a protocol for authenticated group communication.

REFERENCES

[1] N.Meghanathan, S. Boumerdassi, N. Chaki, D. Nagamalai, “Recent Trends in Network Security and Applications: Third International Conference”, The Third International Conference on Network Security and Applications, 2010.

[2] R.Shyamala, S. Valli, “Impact of Black hole and Rushing Attack on the Location-Based Routing Protocol for Wireless Sensor Networks”, Advance in Computer & Inform, Technology, pp. 349-359, 2012.

[3] T. Shimeall, J. Spring, “Introduction to Information Security: A Strategic-Based Approach”, Newnes Compute, pp. 382, 2013.

[4] J.Sen, “Security and privacy challenges in cognitive wireless sensor networks”, arXiv preprint arXiv: 1302.2253, 2013.

[5] G. Sharmaa, S. Balaa, A.K.Vermaa, “Security Frameworks for Wireless Sensor Networks-Review”, 2nd International Conference on Communication, Computing & Security, SciVerse Science Direct, 2012.

[6] C. Cheikhrouhou, A. Koubâa, G. Dini, and M. Abid, "RiSeG: a ring based secure group communication protocol for resource-constrained wireless sensor networks", Personal and Ubiquitous Computing, Vol. 15, No. 8, pp. 783-797, 2011.

[7] X. Wanga, P. Lia, Y. Suia, and H. Yanga, "A Hexagon-based Key Pre-distribution Scheme for Wireless Sensor Networks", Journal of Information & Computational Science, Vol. 11 (8), pp. 2479-2491, 2014.

[8] M. Miettinen, N. Asokan, T.D.Nguyen, A-R.Sadeghi, and M. Sobhani, “Context-Based Zero-Interaction Pairing and Key Evolution for Advanced Personal Devices”, In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, pp. 880-891, 2014.

[9] J. Furtak, and J. Chudzikiewicz, "The concept of authentication in WSNs using TPM", Computer Science and Information Systems, Vol. 3, pp. 183–190, 2014.

[10] W. Xi, X-Y. Li, C. Qian, J. Han, S. Tang, J. Zhao, and K. Zhao, "Keep: Fast Secret Key Extraction Protocol for D2D Communication", IEEE, 2014.

[11] Guorui Li; Ying Wang; Jingsha He, "Efficient Group Key Management Scheme in Wireless Sensor Networks", Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI), 2010

[12] Ju-Hyung, Jun-Sik Lee ; Seung-Woo Seo," Energy Efficient Group Key Management Scheme for Wireless Sensor Networks", 2nd International Conference on Communication System Software and Middleware, 2007.

[13] YuanZhang, Yongluo Shen ; SangKeun Lee, "A Cluster-Based Group Key Management Scheme for Wireless Sensor Networks", Web Conference (APWEB), 2010 12th International Asia-Pacific, 2010.

[14] A.S. Poornima, B.B. Amberker, "A Secure Group Key Management Scheme for Sensor Networks", Fifth International Conference on Conference: Information Technology: New Generations, 2008. ITNG 2008

[15] R. Aparna and B. B. Amberker, "Key management scheme for multiple simultaneous secure group communication," in Proceedings of the IEEE International Conference on Internet Multimedia Services Architecture and Applications (IMSAA '09), December 2009.

[16] B R Purushothama , B B Amberker, "Group key management scheme for simultaneous multiple groups with overlapped membership", Third International Conference on Communication Systems and Networks (COMSNETS 2011), 2011.

[17] C.K.Wong, M. Gouda, and S.S. Lam. Secure Group Communication Using key Graphs. IEEE/ACM Transactions on Networking, Volume 8,No.1, pp.16-30, Feb.2000.

AUTHORS

H.S Annapurna is currently working as Associate Professor in the department of Computer Science &Engg., Sri Siddhartha Institute of Technology, Tumkur. She has obtained her Bachelor of Engineering from University of Mysore, Mysore. She has received Masters degree in Software Systems from BITS, Pilani. She is currently pursuing Doctral degree in the area of cryptography and network security from Sri Siddhartha Academy of Higher Education, Tumakuru India.

M.Siddappa received B.E and M.Tech degree in Computer Science & Engineering from University of Mysore, Karnataka, India in 1989 and 1993 respectively. He has completed doctoral degree from Dr.MGR Educational Research Institute Chennai under supervision of Dr.A.S.Manjunatha, CEO, Manvish e-Tech Pvt. Ltd., Bangalore in 2010. He worked as project associate in IISc, Bangalore under Dr. M.P Srinivasan and Dr. V.Rajaraman from 1993 – 1995. He has teaching experience of 26 years and research of 10 years . He published 45 Technical Papers in National, International Conference and Journals. He has citation index of 113 till 2015 and h-index of 3 and i10-index of 1 to his credit. He is a member of IEEE and Life member of ISTE. He is working in the field of data structure and algorithms, Artificial Intelligence, Image processing and Computer networking. He worked as Assistant Professor in Department of Computer Science & Engineering from 1996 to 2003 in Sri Siddhartha Institute of Technology, Tumkur. Presently, he is working as Professor and Head, Department of Computer Science & Engineering from 1999 at Siddhartha Institute of Technology, Tumakuru. He has visited Louisiana university Baton rouge and California university.

INTENTIONAL BLANK

EXPLORING THE DYNAMIC INTEGRATION OF HETEROGENEOUS SERVICES

Makaziwe Makamba, Jabu Mtsweni and Ernest Ketcha Ngassam

[1]Department of Computer Engineering,
University of South Africa, Pretoria, South Africa
55423434@mylife.unisa.ac.za
[2]CSIR, Meiring Naude Road, Pretoria, South Africa
makamba.makaziwe93@gmail.com

ABSTRACT

The increase need for services to handle a plethora of business needs within the enterprise landscape has yielded to an increase in the development of heterogeneous services across the digital world. In today's digital economy, services are the key components for communication and collaboration amongst enterprises internally and externally. Since Internet has stimulated the use of services, different services have been developed for different purposes prompting those services to be heterogeneous due to incompatibles approaches relied upon at both conceptual and exploitation phases. The proliferation of developed heterogeneous services in the digital world therefore comes along with a range of challenges more precisely in the integration layer. Traditionally, integration is achieved by using gateways, which require considerable configuration effort. Many approaches and frameworks have been developed by different researchers to overcome these challenges, but up to date the challenges of integration heterogeneous services with minimal user-involvement still exist. In this paper, we are exploring the challenges of heterogeneous services and characteristics thereof with the aim of developing a seamless approach that will alleviate some of these challenges in near future. It is therefore of outmost importance to understand the challenges and characteristics of heterogeneous services before developing a mechanism that could eliminate these challenges.

KEYWORDS

Integration, heterogeneous, dynamic-integration, services, and heterogeneous-services.

1. INTRODUCTION

The rapid evolution of internet connectivity and service enterprises has raise potential interaction with many other service enterprises across the globe. In this context, the services are defined as an abstracted, logical view of actual programs that exchange messages between provider agent and requester [1]. A service is also defined as an abstract resource that represents a capability of performing tasks that form a coherent functionality [2]. The key to these services is loosely coupled nature, where the service interface is independent of the implementation. The increasing

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 103–110, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61009

number of heterogeneous services has made a significant drawback on the scalability, integration and the performance of the services [3]. Heterogeneous service in this context is defined as a uniqueness of services that are developed on different environments using different techniques and architectures to complete a specific purpose. These services have to interact with each other to achieve a certain business goal. As the development of services is increasing in enormous way across the domain, there is a need to find appropriate approach that will allow these heterogeneous services to be dynamically integrated without using hard-wiring approaches, which are tiresome [3]. There are traditional methods of integrating these heterogeneous services; however these traditional methods are not adequate as the require user-intervention when there is a change in service that need to be integrated. The traditional methods have some problems, such as the code complexity more difficult to maintain, lead the efficiency of the bottleneck easily [4]. It is therefore, essential to develop a method that will enable dynamic integration of the heterogeneous services without the need for user-intervention.

This paper is focusing on identifying the challenges of heterogeneous services integration; describe the characteristic of heterogeneous services and discussing the advantages and limitations of the existing integration models.

The remainder of the paper will discuss the heterogeneous services integration in Section 2. Section 3, will discuss the characteristics of heterogeneous services, the challenges of heterogeneous services which is the core of this paper will be discussed in Section 4 of this paper. The related work on what other researchers developed to alleviate these challenges will be discussed in section 5. Section 6 will conclude the paper with the recommendations of what needs to be developed to eliminate these challenges.

2. HETEROGENEOUS SERVICE INTEGRATION

Most of the services are heterogeneous, which means they are executed under different platforms, designed using different architectures and programming languages [5]. Heterogeneous services are services that are developed by diverse vendors for different purposes on different platforms, using different architectures; different protocols and different languages [6]. These heterogeneous services need to be integrated. Heterogeneous service integration is not a new phenomenal in the field of integration. Integrating heterogeneous services in a dynamic manner may improve if it could include prescriptive or solution-oriented knowledge where the result from scientific justification (predicting, understanding or explaining phenomena) can be used in designing solutions to this complex problems [7].

However, heterogeneous service integration has become major issue in the area of integration. This is motivated by complexity of these heterogeneous services and independent services when they need to communicate with one another to achieve a specific business goal. Service integration has to be performed for these services to communicate.

Service integration in this context is defined as the integration of discrete IT services components into a coherent set of end-to-end service bound by Operation Level Agreements (OLAs) and Services Level Agreements (SLAs) [9]. The service integration is a challenge when these heterogeneous services need to interact. There are certain aspects that make these services to be heterogeneous, which results to challenge when these services need to be integrated. The following section will discuss in details the characteristics of heterogeneous services.

3. HETEROGENEOUS SERVICES CHARACTERISTICS

Inter-enterprise integration is an essential requirement for today's successful business [10]. With the aim of overcoming heterogeneity, various technologies and standards for the definition of languages, vocabularies and integration patterns are being developed. However, before discussing the pattern and the technologies that have been developed, there is a need to define the heterogeneous services and the characteristic of these services. As the above section has defined the heterogeneous services, this section will discuss the characteristics of these heterogeneous services. After an extensive literature review and the empirical study that was conducted. We have discovered that services are heterogeneous based on the following aspects; environment, technology, architecture and programming language [11]. As Hajiji mentioned in 2012, there are many fundamental aspects that synthesize the heterogeneity in these services [12]. The heterogeneity of services is also characterized by 4 fundamental aspects.

1. Environment- Due to execution environment, as some services are developed to be executed on different environments such as Mobile Services, Cloud Services, Desktop services, Web Services, ATM Services and many more. These services are explicitly developed to run on these environments for a specific purpose.

2. Systems and Technology- There are several services that run on specific system. Some services are developed to run on Android, some on IOS, while some run on Linux, Windows and Mac OS. Systems and technology are making these services heterogeneous.

3. Architecture- Architecture is another fundamental aspect that differentiates services. Some services are developed for Service Oriented Architecture (SOA) and some on Enterprise application Integration (EAI) while some other architecture allows services to return a JSON and while some are allowed to return Extensible Mark-up Language (XML).

4. Programming Language- This is another component that makes services to be heterogeneous. Some services are developed using php while others are developed using Java, C++ and C#. This makes services to be heterogeneous as they are difficult to integrate due to different programming languages [9].

These above characteristics are granting a challenge on service integration environment. As the figure 1 below depicts, services can be heterogeneous because of the deployment environment, could be mobile, desktop or laptop, and the architecture that is used to develop the services. However these services can be heterogeneous but they need to talk to each other to share information.

As it shown in the picture, these services are deployed in different environment; they are using different architecture and different technology. There are services that are deployed for mobile phone, there are services that are created for desktops and there are services are designed for laptops, however these service needs to interact with each other. It is therefore critical to integrate these services in a dynamic manner. A manner that will not require user-intervention, as it delays the progress of integration task. In the following section, this paper will discuss in details the key challenges of heterogeneous service integration.

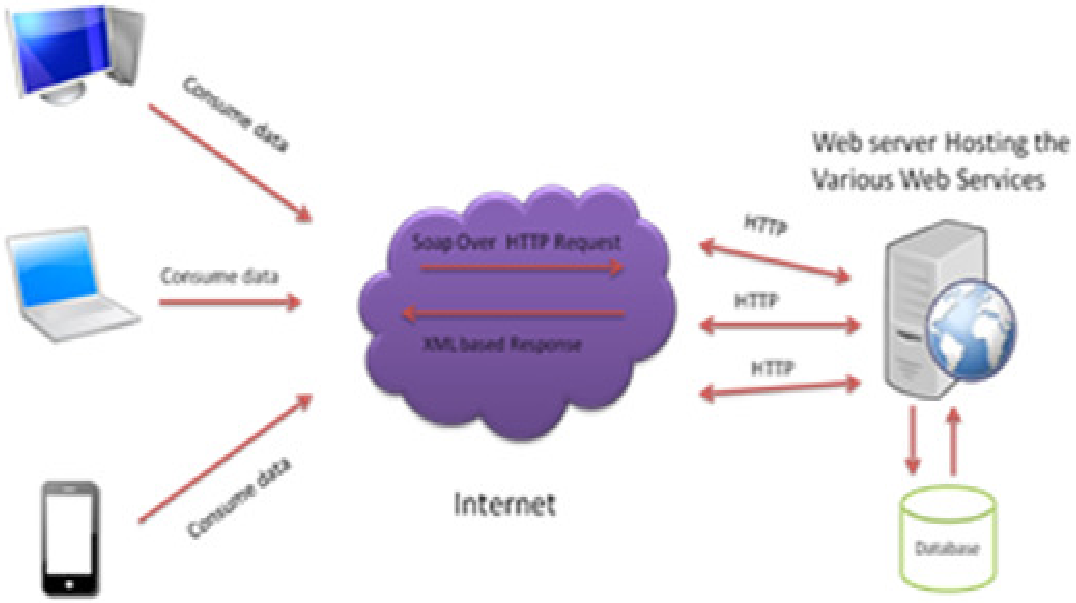


Figure 1: Heterogeneous Service Integration [9]

4. CHALLENGES OF HETEROGENEOUS SERVICE INTEGRATION

In today's world, many enterprises use services as a main core of communication. The challenge that most organizations are facing is to provide a method that could allow dynamic integration of heterogeneous service, a method that can allow these services to work together to address business goals that constantly evolve [13]. There are several heterogeneous services that have been developed by different vendors, such as mobile services, cloud services, web services and others [14]. Most of these services are integrated manually by using Application programming interface (APIs), Adapters, ESBs and many more. This led the companies to lose profit in their business as the integration task becomes cumbersome for the developer, as he/she needs to be involved when integrating these services. These services has varied and multiple integration points. This Increases the challenge of integrating heterogeneous services on the fly. Each service is independent and performs its own task, which might be valuable to a cohesive system. The other challenge that makes it difficult for integration is multiple data models and multiple instances of the services. Achieving a seamless flow of information requires a significant integration method. A method that can scale with business needs as new applications and platforms appear. The current methods of integrating heterogeneous services are very tedious, inflexible and time consuming when the services change requirements [14].

One of the aspects that make these services to be complex to integrate is the 4 mentioned fundamental characteristics that are mentioned above. As these services run on different platform, environment, and system and developed using different programming languages. It makes it difficult to integrate them as they are heterogeneous. Most of the services are currently integrated using manual approach which is to develop an API for each service that needs to be integrated. When there is a change on a service an API has to be modified which make it unmanageable when there is increase in number of services. Hence this manual method is regarded as an inadequate method [14].

Integrating heterogeneous services in order to connect their functionality is utmost importance for promoting deployment dynamic integration of heterogeneous service. It is therefore important to understand the characteristic of these heterogeneous services and explore the challenges of integrating them in order to develop a mechanism to that will enable dynamic integration of these services. A method that will allow heterogeneous services to integrated with a minimal user-involvement. Hence we are exploring the challenges so that we can develop a method that will alleviate these challenges in near future.

Many studies have been conducted to resolve the issue of heterogeneous service integration. The following section will discuss more in details about the related work.

5. RELATED WORK

The challenges of heterogeneous service integration is not a new phenomenal in the field of integration. Many researchers have explored the different paradigm to develop solutions that can eliminate these challenges of heterogeneous service integration. However, up to today, the challenges have not been resolve completely. A new service integration system was developed by Huiyang (2006) to resolve the existing challenges. This system only allows the integration of services at a component level [7]. In this system a user can insert common service logic to their business logic [7]. The system establishes an agent between service components; this was done to allow the request and orchestration to identify the components in a correct manner [7]. However, this system requires a developer's intervention to integrate, modify and change certain service requirements.

Another agent-based web service integration model was developed by Yu, et al (2008) to integrate the services and ensuring the Quality of Services (QOS) of the integration system [8] This QOS-based integration was developed to improve services integration efficiency and minimize the integration cost through the selection mechanism. However, this mechanism worked for integrating services but it does not eliminate the user-involvement challenge when the change is required on a services. Another integrated model was proposed and the integration framework of component-based was constructed by Shaoba et al (2010) to solve the problems of heterogeneous service integration [4]. This integration model was done for user management and user access control and register service component. This framework was mainly used to standardize enterprise business processes, to coordinate the data processing, and to build configurable enterprise service bus; data management modules [4].This framework has achieved data synchronization between heterogeneous systems but has not solved the problem of heterogeneous service integration with a minimal user-involvement. Hence, there is a need to develop a mechanism that will dynamically integrate heterogeneous services without a need for user-intervention. The following table will give more advantages and limitations of the current methods of integration.

Table1: Comparisons of integration models

Integration model	**Advantages**	**Limitations**
ESB	Services are connected to this logical bus through smart connectors, which encapsulate system functionality and provide a layer of abstraction between bus and application. ESB use of open communication standards, connectivity between bus and applications is established. ESB uses several integration points. It's easy to onboard new services and new nodes when using ESB.	Requires ongoing management of message versions to ensure the intended benefit of loose coupling. Though ESB systems can require a significant effort to implement, they produce no commercial value without the subsequent development of SOA services for the ESB. ESB have lot of upfront overheads. When using ESB there is a need to define canonical message format. The developmental complexity. It normally requires more hardware than simple point-to-point messaging. ESB becomes a single point of failure. It requires analysis skills to configure, manage and operate ESB.
EAI	EAI solutions provide process management functionality to orchestrate inter-application message exchanges, and an administration console to monitor and track the workings of the hub. Maintains information integrity across multiple systems. EAI focuses on sharing both business data and business process. EAI focus on integrating services on the data level, application level and the method level.	EAI is more on point-to-point effort, which becomes unmanageable as number of services increases. EAI consume considerable time and effort to deploy, and are subject to high project failure rates. Requires expect knowledge to integrate services using EAI. Require a fair amount of up front design, which many managers are not able to envision.

Middleware	Middleware enable services to exchange messages, even though we do not know the platform of other service. It guarantees throughput, reliability and efficiency. It flexible to modify and separate interfaces from the applications. Independence of layers and database. Transparency and it easily customize all components.	It uses API to integrate services. User-intervention is required whenever there is a change in services that need to be integrated. It is limited in terms of scalability, each client adds overhead. It uses low-level languages. A small change is the environment development can lead to major downtime of real-time applications. It is expensive.

Due to these series of limitations on the existing models of integration, this is a reason why there is a need to develop a method that will enable dynamic integration of heterogeneous service. Due to advantages of dynamic integration model, it would be easy to integrate heterogeneous service without a need of user-intervention. This could eliminate the challenges that many business enterprise are facing. As we can identify the gap that needs to be closed, dynamic integration could be a solution to this existing challenge. Dynamic integration would be a proper solution because it does not need user-intervention. It does not need a developer to modify API as the services will automatically update them. It would be scalable since services are integrated on the fly. In these days is hard to live normally without having to rely on some kind of technology. Technology is everywhere and used in different ways by all kinds of people in different work areas. Enterprises are now exposing their business as services for global visibility and for automation of business processes [15]. It is important to understand the current challenges and what has been developed to eradicate this challenge. In future we will develop a dynamic method of integrating heterogeneous service without a need of user-intervention.

6. CONCLUSION

The challenges of heterogeneous service integration are affecting multiple organizations. The technology must be flexible and scalable in reconciling semantic differences amongst these information exchange entities. It is therefore essential to have mechanisms that will dynamic integrate heterogeneous services, to make integration as seamless as possible. Hence in this paper, we described the challenges, characteristics of integration heterogeneous services, and comparison of existing models of integration so that it can be easy to develop a method that will close the existing gaps. In future we will develop a method that will allow dynamic integration of these heterogeneous services.

ACKNOWLEDGMENT

This work is undertaken within the Council for Scientific and Industrial Research (CSIR) in Defense, Peace Safety and Security (DPSS). The support from University of South Africa (UNISA) is also acknowledged.

REFERENCES

[1] A. Reshamwala, G. Khetan and K. Gupta, (2009) “AIDS Helpline based on Service Oriented Architecture”

[2] H. Zhu, E. S, Madnick, (2004) “Context interchange as a Scalabe Solution to Interoperating Amongst Heterogeneous Dynamic Services”

[3] W. He, and L.H. Xu. (2011) “Integration of Distributed Enterprise Applications. IEEE procedengs.

[4] L. Shaobo, H.Yao and X. Qingsheng,(2010) “Heterogeneous system integration based on service component” Applied Mechanics and Materials Vols. 20-23 (2010) pp1305-1310.

[5] M.Makamba, (2016) “Preliminary Investigation of Challenges in Dynamic Integration of Heterogeneous Services” systemic, cybernetics and informatics volume 14-number1-year 2016, ISSN: 1690-4524.

[6] Capgemini, (2013) “Service Integration; A blueprint for regaining control of complex IT vendor landscape” capgemini.

[7] X. Huiyang, S. Meina and S. Junde, (2006) “A new service integration system for modern service industry based on SOA” Proceedings of International conferences of technology and system in Beijing.

[8] W. Yu, (2011) “An integrated middleware based solution for supporting secured dynamic coating applications in Heterogeneous environment. IEEE, 2011.

[9] F. Hajaji, (2012) “Five Aspects of Application Integration Requirements. ARPN journal of system and software, 2012. Vol.2 no3

[10] Athanasopoulos.G, T. a. P., (2010) “Interoperability and Heterogeneous Services”. China, MCSIN

[11] WebLogic, (2012) “Determining Integration Solution Requirements”, USA: BEA systems

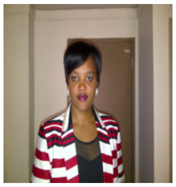
[12] Makamba, M., Mtsweni, J. & Ngassam, E. K., (2015) “Dynamic Integration of heterogeneous Services”. Cape Town, South Africa, SAICSIT.

[13] MagnetyzingSolution, 2007. Middleware, India: Magnetyzing solutions

[14] Kaviani, N., Mohabbati, B. & Lea, R., 2012. ReCoIn: A Framework for Dynamic Integration of Remote Services into a Service-Oriented Component Model. Colombia, ACM.

[15] Laliwala, Z., Kholsa, R., Majumdar, P. & Chaudhary, S., (2006) “Semantic and Rules Based Event-Driven Dynamic Web Services Composition for Automation of Business Processes”. Chicago, USA, IEEE

AUTHOR



Makaziwe Makamba is a PhD candidate at the University of South Africa. She completed her Masters in 2012 at the University of Fort Hare in Alice, South Africa. Her area of interest is service integration, e-services and the dynamic integration.

OBJECTIVE EVALUATION OF A DEEP NEURAL NETWORK APPROACH FOR SINGLE-CHANNEL SPEECH INTELLIGIBILITY ENHANCEMENT

Dongfu Li and Martin Bouchard

School of Electrical Engineering and Computer Science,
University of Ottawa, Ottawa, Canada
Lidongfu1983@gmail.com, martin.bouchard@uottawa.ca

ABSTRACT

Single-channel speech intelligibility enhancement is much more difficult than multi-channel intelligibility enhancement. It has recently been reported that machine learning training-based single-channel speech intelligibility enhancement algorithms perform better than traditional algorithms. In this paper, the performance of a deep neural network method using a multi-resolution cochlea-gram feature set recently proposed to perform single-channel speech intelligibility enhancement processing is evaluated. Various conditions such as different speakers for training and testing as well as different noise conditions are tested. Simulations and objective test results show that the method performs better than another deep neural networks setup recently proposed for the same task, and leads to a more robust convergence compared to a recently proposed Gaussian mixture model approach.

KEYWORDS

Single-channel speech intelligibility enhancement processing, Deep Neural Networks (DNN), Multi-Resolution CochleaGram (MRCG), Gaussian Mixture Models (GMM)

1. INTRODUCTION

Single-channel speech intelligibility enhancement is more challenging than multi-channel speech intelligibility enhancement, because information about the spatial sound propagation is not available. In most cases, it is hard for a single-channel noise-reduction algorithm to know how and to what extent to modify a specific parameter to improve speech intelligibility [1].

A lot of previous work has been designed on the prior knowledge or estimation of noise, such as the a priori Signal-to-Noise Ratio (SNR) algorithm, the Minimum Mean-Square Error (MMSE) approach, the log-MMSE approach, the Wiener filter, and so on. They have all been shown to be ineffective for intelligibility enhancement [1].

Two machine learning training-based methods have recently been proposed for intelligibility improvement: the Gaussian Mixture Models (GMM)-based approach [2] and the Deep Neural

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 111–123, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61010

Networks (DNN)-based approach [3]. In 2006, an efficient way to train a multilayer neural network was proposed [4] and a new area of machine learning emerged, which is called deep learning, deep hierarchical learning or DNN [5],[6]. Key aspects of machine learning and artificial intelligence have been widened by the techniques developed from deep learning [7]-[9], and there are several active researchers in this area [10]. For speech intelligibility processing using DNNs, in 2013 a DNN using 85 features as inputs was proposed for a speech recognition task with hearing-impaired listeners [11]. In 2014, it was reported that the Multi-Resolution CochleaGram (MRCG) feature set produced a better result in a multilayer perceptron neural network, which is a simpler type of neural network [3].

Under some conditions, it has been reported that the DNN approach, which better represents the state of the art in machine learning, generalizes better and better processes previously unseen data patterns compared to the GMM method. The main goal of this paper is to evaluate the performance of the DNN method proposed in [3] using objective measures under different conditions (different types and levels of noise, mismatch between training set and testing set, etc.) and to compare the performance with the GMM approach previously proposed for the same task [2].

2. Deep Neural Networks

2.1 Pretraining DNN with a Restricted Boltzmann Machine

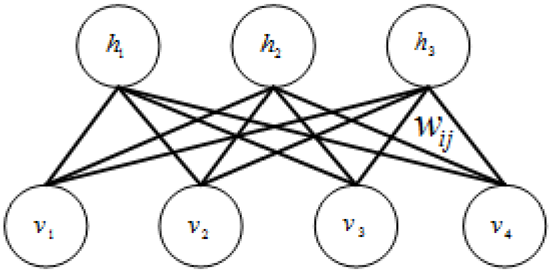


Figure 1. RBM with 4 visible units and 3 hidden units.

Training deep neural networks is challenging, because training can easily get stuck in undesired local optima which prevent the deeper layers from learning useful features. This problem can be partially circumvented by pretraining, i.e., performing a step of unsupervised training before the supervised learning step. The Restricted Boltzmann Machine (RBM) method is a useful way to conduct DNN training. A RBM is a simplified kind of a Boltzmann Machine with no visible-visible units connections and hidden-hidden units connections. In a Restricted Boltzmann Machine, a visible unit only has connections to hidden units, and reversely a hidden unit only has connections to visible units. This special kind of structure provides the advantage that when a visible unit is learning its optimal weights corresponding to a set of hidden units, the learning is independent to other visible units. This advantage also applies to the training of hidden units. Then the whole network can be trained in parallel. With the great progress made in graphics processing unit processors (GPU), i.e., in parallel computing, this can become a great advantage. Figure 1 shows a simple Restricted Boltzmann Machine with 4 visible units and 3 hidden units.

The "energy" of a Restricted Boltzmann Machine can be written as below:

$$E(\mathbf{s}) = -\sum_{i} s_i b_i - \sum_{i<j} s_i s_j w_{ij} \tag{1}$$

where $\boldsymbol{s}$ is the state vector $\boldsymbol{s} = \{ v_1, v_2, v_3, \ldots, v_n, h_1, h_2, \ldots, h_m \}$, n is the number of visible units, m is the number of hidden units, and s_i is a component of $\mathbf{s}$ that can be a visible unit or a hidden unit. If s_j is the unit connected to s_i, then w_{ij} is the weight between s_i and s_j

2.2 Restricted Boltzmann Machine Learning

Let unit i be a unit to update its binary state. The total input z_i for this unit i is the sum of its bias b_i and the weighted products from connections to other units:

$$z_i = b_i + \sum_{j} s_j w_{ij} \tag{2}$$

The probability for this unit to turn on or off is given by a logistic function:

$$prob(s_i = 1) = \frac{1}{1 + e^{-z_i}} \tag{3}$$

As mentioned earlier, the visible units only connect to hidden units. For a given state vector $\boldsymbol{s}$, no matter if the network is updated in any order, the network will eventually reach a stationary distribution (equilibrium). Then for all possible binary state vectors $\boldsymbol{u}$, the probability of vector $\boldsymbol{s}$ can be given by the energy:

$$P(\mathbf{s}) = e^{E(s)} \Big/ \sum_{\mathbf{u}} e^{E(\mathbf{u})} \tag{4}$$

Given a training set of state vectors (data), the goal of the learning is to find the optimal weights and biases to make the state vectors maximize the product of the probabilities that the Boltzmann machine assigns to the binary vectors in the training set. By differentiating (4) using $\partial E(\mathbf{s}) / \partial w_{ij} = -s_i s_j$, it can be shown that:

$$\sum_{\mathbf{s}} \frac{\partial \log P(\mathbf{s})}{\partial w_{ij}} = \left\langle s_i s_j \right\rangle_{data} - \left\langle s_i s_j \right\rangle_{model} \tag{5}$$

where $\left\langle s_i s_j \right\rangle_{data}$ is the state value of the data distribution and $\left\langle s_i s_j \right\rangle_{model}$ is the state value when the Boltzmann machine is sampling state vectors from its equilibrium distribution. Then the gradient ascent is surprisingly simple, because the differentiation $\sum_{\mathbf{s}} \frac{\partial \log P(\mathbf{s})}{\partial w_{ij}}$ only depends on the states:

$$\Delta w_{ij} \quad \propto \quad \left\langle s_i s_j \right\rangle_{data} - \left\langle s_i s_j \right\rangle_{model} \tag{6}$$

The update value Δw_{ij} is the product of $\left\langle s_i s_j \right\rangle_{data} - \left\langle s_i s_j \right\rangle_{model}$ and the learning rate. The learning rate can be constant or it can vary during the training steps to satisfy different situations. To get the Boltzmann equilibrium distribution, we can follow the steps below:

1) Starting with a data vector on visible units, update all of the hidden units in parallel;

2) Update all of the visible units in parallel to get a "reconstruction" of the visible units;

3) Update all of the hidden units again until it is the equilibrium distribution.

This algorithm may take a long time to get to the equilibrium. It can be stopped at step 3) or continue iteratively to update 1) and 2), this is called "contrastive divergence" and has been found to work well in practice [12].

3. Multi-Resolution Cochleagram Feature

The multi-resolution cochleagram (MRCG) feature was originally proposed in [3]. The MRCG feature is a multi-resolution power distribution of an acoustic signal in the time-frequency representation. The cochleagram represents the excitation pattern on the basilar membrane in the inner ear as a function of time. Four cochleagrams at different frequency resolutions are combined to form the MRCG feature, including one high resolution cochleagram and three low resolution cochleagrams.

The cochleagram is calculated in two steps. The input signal is first filtered by a gammatone filter bank:

$$g_{f_c}(t) = t^{N-1} \exp[-2\pi t\, b(f_c)] \cos(2\pi f_c t) u(t) \tag{7}$$

where f_c , the center frequencies, are uniformly spaced on the equivalent rectangular bandwidth (ERB) scale, N is the order of the filter, $u(t)$ is the step function, and $b(f_c)$ is the bandwidth related to f_c:

$$b(f_c) = 1.019 * ERB(f_c) = 1.019 * 21.4 * \log(1 + 0.00437 * f_c) \tag{8}$$

$b(f_c)$ increases as f_c increases, which means that the frequency resolution decreases as the frequency increases. The signal is then divided into frames. A cochleagram is the power of each time frame and in each gammatone bank channel.

The MRCG feature can then be described as below:

1) Given an input, compute a 64-channel cochleagram (CG1) using 20 ms frames with 10 ms frame shifts, and apply a log operation on the output in each time-frequency (T-F) unit;

2) CG2 is similar as CG1, but with 200 ms frames and 10 ms frame shifts;

3) CG3 is derived by averaging CG1 across a square window of 11 frequency channels and 11 time frames centered at a given T-F unit (zero padding is applied at the edges of CG1 when units outside CG1 are needed in the averaging process);

4) CG4 is similar as CG3, but with a 23*23 square window;

5) Concatenate CG1-CG4 to obtain the MRCG feature.

Delta and double-delta features (i.e., computing the differences of feature values at consecutive times, and computing the differences of those differences) are widely used in speech processing to capture temporal dynamics. Delta and double-delta features were thus also used on the MRCG features generated above. Then the final MRCG feature set is obtained. Each time frame becomes an MRCG feature set of dimension 64 by 12.

4. MRCG-DNN ALGORITHM

The MRCG-DNN algorithm is a kind of channel-selection algorithm. A channel-selection based algorithm will retain the T-F bins where speech is dominant, and discard the T-F bins where noise is dominant. The algorithm makes the retain/discard decision based on the SNR of each T-F bin, compared to a threshold called the local SNR criterion (LC). In an ideal case, where separate access to clean speech and noise-only signals is possible and the exact SNR can be computed, using a channel-selection algorithm we can get the ideal binary mask (IDBM):

$$B(k,t)=\begin{cases}1 & SNR>LC\\ 0 & otherwise\end{cases} \tag{9}$$

where $B(k,t)$ is the IDBM mask at channel (frequency) k and time t.

The IDBM has been shown to improve speech intelligibility at any input SNR level [1] (even as low as -40 dB) . The IDBM is an unrealistic condition that can never occur in real life, but for several reasons this result is important. In particular, the outcome of the IDBM can provide an upper bound, so that the IDBM can be a criterion to estimate the performance of a practical algorithm.

The MRCG-DNN algorithm uses a relatively large corpus of speech and noise sources, together with the calculated IDBM as the input of a system to be trained. An overview of MRCG-DNN algorithm is shown in Figure 2.

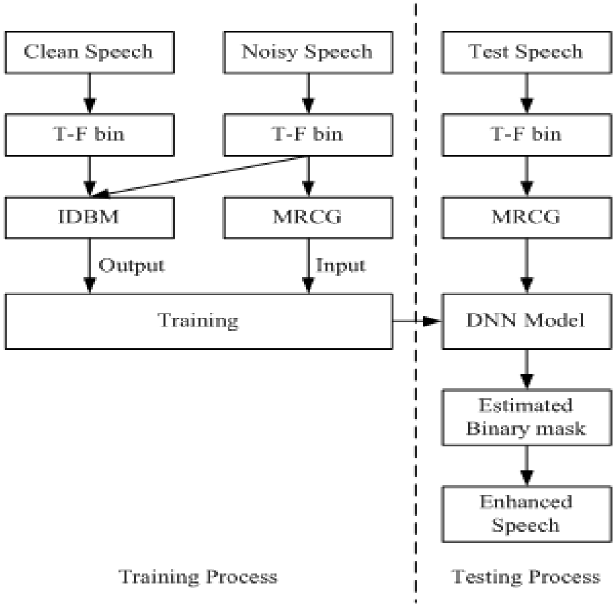


Figure 2. Overview of the MRCG-DNN algorithm

5. RESULTS

To test how the MRCG-DNN algorithm performs in different situations, we did some tests using the HINT speech database [13], the TIMIT speech database (LDC93S1) [14], and noise from the Aurora 2 database [15]. The HINT speech files have a 2 to 3 seconds length for each file, all from the same male speaker. The TIMIT is a corpus of phonemically and lexically transcribed speech of American English speakers of different genders and dialects. Each sentence is about 3 to 5 seconds long. The Aurora noise files contain different situations of noise, including babble, airport, restaurant, and street. In the tests we have used these four kinds of noise.

In this paper, the Matlab™ code is based on two Deep Learning toolboxes: DeepLearnToolBox [16] and DeepNeuralNetwork [17]. The bone structure of the first toolbox, a few active functions of the second toolbox, and some improvements were combined together to perform the simulations of this work.

In the following tests, 200 clean speech files are used for training (approx. 5-6 minutes of recordings), and 70 clean speech files are used for testing (approx. 2-3 minutes of recordings). If the training data is too short (i.e., less than 1 min) the result was found to be poor. If the training data is too long, it can either cause the computer to run out of memory or to take too much processing time. Using 5-6 minutes of recordings for training was found to produce a good result with a relatively fast processing speed. As a comparison, in [18] 390 sentences were used for training and in [3] 100 sentences were used. The sampling rate was set to 16000 Hz. The noisy files were produced by randomly selecting a noise segment that has the same length as the clean speech files and adding them together with the desired SNR. The tests were operated using Matlab™ on a Windows 7™ (64-bit) system, with an Intel Core™ i5-4310M CPU, and 8 GB memory.

To evaluate the estimated mask compared to the IDBM, a method called the HIT - FA (HIT minus FA) metric is used, which has been shown to correlate well with human intelligibility [1]. The HIT is the probability of correct detection (the percentage of target-dominant T-F units correctly classified), while the FA is the probability of false alarm (the percentage of noise-dominant units incorrectly classified). Both the HIT and FA need to be taken into account when evaluating the performance of the binary mask estimate, and the HIT-FA is therefore a simple difference metric that can be used to quantify the performance of the estimate.

5.1 Tests for simple noises

These first tests used the HINT database, which is a single speaker database. White, purple and pink noises were used. The model was trained with -5dB, 0 dB and 5dB input SNR, and tests were performed with -5dB SNR. For each noise that was tested, the DNN model was trained with speech from the training set and additive noise consisting of only one noise type. The testing phase was performed with speech from the test set and again additive noise consisting of only the same noise type. Table 1 shows the HIT-FA results as well as URL links to the clean test male speech sound file, the noisy sound files and the processed sound files.

Table 1. Results for simple noises

Clean test speech					
Noise type	**HIT**	**FA**	**HIT-FA**	**noisy**	**processed**
white	84.2%	2.6%	81.6%	link	link
purple	92.3%	3.1%	89.2%	link	link
pink	78.5%	20.2%	58.3%	link	link

From Table 1 we can see that the DNN model performs very well for white noise, and even better for purple noise. The HIT rate is very high, while the FA rate is very low. The processed sound files are also quite clear. On the other hand, the DNN model produces a much higher FA for pink noise, and the processed file is highly distorted. The reason for this is that purple noise energy is mainly in high frequency bands, while pink noise is mainly in low frequency bands, and human voice is also mostly distributed in low frequencies. In the purple noise case, the DNN model can easily separate them, it behaves as a kind of high-end low pass filter. In the pink noise condition, the separation is more difficult for the DNN because of the increased frequency overlap between speech and noise.

5.2 Same speaker for training and testing, same noise type for training and testing

In this test we used the HINT database which has a single speaker, with different sentences for training and testing. We used four kinds of realistic noise: airport, babble, restaurant, and street. We trained the DNN model with -5dB, 0 dB and 5dB input SNR, and then performed the testing with -5dB SNR. For each type of noise we trained the model and then used the model to test different sentences corrupted by the same kind of noise as the one used for training (but of course using different noise segments). Table 2 shows the HIT-FA results and URL links to the clean test male speech sound file, the noisy sound files and the processed sound files.

Table 2. Results with same speaker for training and testing, same noise type for training and testing.

Clean test speech					
Noise type	**HIT**	**FA**	**HIT-FA**	**noisy**	**processed**
babble	76.2%	7.3%	68.9%	link	link
airport	80.3%	11.0%	69.3%	link	link
restaurant	77.5%	9.9%	67.6%	link	link
street	78.2%	7.5%	70.7%	link	link

From Table 2 we can see that the HIT rate of the estimated mask is relatively high and the FA rate is low, regardless of the noise type. The resulting HIT-FA rate are fairly good and are comparable with previous results from the literature [18]. When listening to the processed output files, it is debatable if the intelligibility is really improved or not compared to the original noisy files, i.e., subjective listening tests would be required to fully determine this. In some cases (e.g. babble noise) the intelligibility of the processed files appears to be better than for other cases (e.g. street noise), so this indicates that the performance can be noise dependent. It should be noted that the noises used for testing were challenging noises, but they are more likely to correspond to real noise conditions in practice.

5.3 Different speakers for training and testing, same noise type for training and testing

In this section, we used the TIMIT dataset with several different speakers for training and another speaker was used for testing (but for the same female gender as the ones used in training). Four kinds of noise were used: airport, babble, restaurant, and street. We trained the DNN model with -5dB, 0 dB and 5dB SNR, and did testing at -5dB SNR. For each type of noise we trained the model and performed the testing using the same type of noise (but different noise segments). Table 3 shows the HIT-FA results and URL links to the clean test female speech sound file, the noisy sound files and the processed sound files.

Table 3. Results for different speakers for training and testing, same noise type for training and testing.

Clean test speech					
Noise type	**HIT**	**FA**	**HIT-FA**	**noisy**	**processed**
airport	83.8%	8.2%	75.6%	link	link
babble	80.6%	6.9%	73.7%	link	link
restaurant	82.2%	13.2%	69.0%	link	link
street	84.4%	6.0%	78.4%	link	link

FromTable 3 we can see that the HIT rate of the estimated mask is relatively high and the FA rate is low, regardless of the noise type. This indicates that in terms of HIT-FA rate the performance of the MRCG-DNN method can be robust to different speakers, i.e., when the speakers used for training differ from the speakers used in testing. Although it is possible to train simultaneously for both male and female speech, our experience has been that better results are obtained when the same gender is used for training and testing. In terms of intelligibility, as in the results of Table 2, it is debatable if the intelligibility is improved in the processed sound files of Table 3, and the performance seems to be better for some noise types (e.g. airport and street).

5.4 Training with 3 types of noise and testing with a fourth type of noise

For this test, we have used either the HINT dataset (single speaker) or the TIMIT dataset (different speakers), as well as airport, babble, restaurant, and street noise. We trained the DNN model with airport, babble and restaurant noise, then we tested the model with street noise. For the HINT database the training and testing sentences were different but from the same speaker. For the TIMIT database, training and testing were made with different speakers (of the same gender) and different sentences. Table 4 shows the HIT-FA results and URL links to the clean test speech sound file, the noisy sound files and the processed sound files.

Table 4. Results for training with 3 types of noise and testing with a fourth type of noise.

Clean test speech					
Database	**HIT**	**FA**	**HIT-FA**	**noisy**	**processed**
HINT database	76.2%	11.4%	64.8%	link	link
Clean test speech					
Database	**HIT**	**FA**	**HIT-FA**	**noisy**	**processed**
TIMIT database	84.6%	18.9%	65.7%	link	link

From Table 4, we can see that the HIT rate of the estimated mask is relatively high and the FA rate is low. This indicates that in terms of HIT-FA rate the performance of the MRCG-DNN method can in principle be robust to conditions where we have both different speakers and different noise types between training and testing, as long as training is done with appropriate data (i.e., the training speech should have some similarity with the test speech, the training noise should have some similarity with the test noise). But as observed in the previous results of Table 2 and Table 3, for the results of Table 4 it is highly debatable if the intelligibility is improved.

5.5 Training with different levels of noise, testing with the same noise type at a specific level

Here we used the HINT dataset (same speaker for training and testing, but different sentences), and babble noise for training and testing. We trained the DNN model with -5dB, 0 dB and 5dB SNR, then tested with -5dB SNR. We trained the DNN model with -7dB, -5 dB and 0dB SNR, then tested with -5dB SNR. Finally, we trained the DNN model with -10dB, -7 dB and -5dB SNR, and tested with -5dB SNR. **Error! Reference source not found.** shows the HIT-FA results and URL links to the clean test speech sound file, the noisy sound files and the processed sound files.

Table 5. Results for training with different levels of noise, testing with the same noise type at a specific level.

Clean test speech					
SNR	**HIT**	**FA**	**HIT-FA**	**noisy**	**processed**
-5dB, 0dB, 5dB	76.2%	7.3%	68.9%	link	link
-7dB,-5dB, 0dB	74.8%	6.6%	68.1%	link	link
-10dB,-7dB,-5dB	70.3%	5.3%	64.9%	link	link

From Table 5 we can see that the HIT rate as well as the FA and HIT-FA rates of the estimated mask all slightly drop from the top row to the bottom row. This indicates that training with data having the same and higher SNR than the testing data may lead to a slightly higher HIT-FA rate. Informally, the intelligibility in the sound files also seems to follow that trend.

5.6 Training with different local SNR criterion (i.e., thresholds used for binary mask decision)

For this test, we used the HINT dataset (same speaker for training and testing, but different sentences), and babble noise for training and testing. We trained the DNN model with -5dB, 0 dB and 5dB SNR, and tested with -5dB SNR. We did this procedure 3 times, and the difference between the 3 cases is that different local SNR criterion values were used (thresholds used for the binary mask decision): 0dB, -5dB, and -10dB. Table 6 shows the HIT-FA results and URL links to the clean test speech sound file, the noisy sound files and the processed sound files.

Table 6 Results of training with different local SNR criterion (thresholds)

Clean test speech					
local SNR criterion	**HIT**	**FA**	**HIT-FA**	**noisy**	**processed**
0dB	74.3%	8.8%	65.5%	link	link
-5dB	76.2%	7.3%	68.9%	link	link
-10dB	79.9%	11.7%	68.3%	link	link

From Table 6 we can see that the HIT rate of the estimated mask increases as the local SNR criterion (threshold) decreases, but the HIT-FA rate doesn’t change too much. Listening to the resulting processed files, we note (informally) that the intelligibility improves from the top row to the bottom row. A higher HIT rate means that a higher percentage of target-dominant T-F bins are kept, so based on these results we adopted a dynamic adjustment method during training in the software program, to make sure that the HIT rate is at least 75%. If the condition is not fulfilled then we dynamically lower the local SNR criterion used for training. But overall, as in previous tables, it is debatable if the processed files provide some intelligibility improvement over the original noisy files.

5.7 MRCG-DNN compared with other approaches

In this section we use two other types of speech intelligibility enhancement approaches and compare them with the MRCG-DNN approach: an DNN using 85 input features proposed in [11] and an Amplitude Modulation Spectrogram (AMS)-GMM proposed in [18]. We used our own implementation of the 85 input features DNN and the original code from [18] for the AMS-GMM. The database and noise conditions were the same as the ones originally used in the 85 input features DNN and the AMS-GMM. The IEEE speech database is also used as clean speech data in this section [19].

For the DNN with 85 input features, the input data is passed through a 64 band gammatone filter. Each subband is divided to 20 ms frames with 10 ms overlap. Each T-F unit extracts 85 features: 15 AMS features, 13 Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) features, 31 Mel-Frequency Cepstral Coefficients (MFCC) features, and 13 delta features for the RASTA-PLP features. This method uses a DNN model for each subband, for a total of 64 DNN models. The MRCG-DNN on the other hand only uses one DNN model for all subbands, so the MRCG-DNN training takes much less time than the 85 input features DNN.

The results for the 85 input features DNN and the MRCG-DNN are compared in Table 7. They both use the same DNN structure, and were trained for 100 iterations. “n6 noise” is a babble noise produced by adding several TIMIT sentences together. We can see from Table 7 that the HIT-FA for the MRCG-DNN is always significantly higher than for the 85 input features DNN, indicating a better performance in terms of HIT-FA rate.

Table 7. Results for 85 input features DNN and MRCG-DNN.

		85 input features DNN			**MRCG-DNN**		
database	**noise**	**HIT**	**FA**	**HIT-FA**	**HIT**	**FA**	**HIT-FA**
HINT	n6 (babble)	62.9%	7.4%	55.5%	72.3%	6.4%	65.9%
IEEE	factory	46.2%	3.4%	42.8%	66.5%	3.8%	62.7%

The other method used for comparison is the AMS-GMM [18]. More specifically, in this approach 4 sub-GMMs are used in order to make the decision to set the TF binary mask to 1 or 0. The results for the AMS-GMM and the MRCG-DNN are compared in Table 8. The HIT-FA of the MRCG-DNN is slightly less (worse) than for the AMS-GMM method, but with a much lower FA (better). Overall it can be said that the results of these two methods are fairly comparable for the considered setup, in terms of HIT-FA rate. However, we also found that the AMS-GMM method fails to find a clustering solution in some conditions (e.g., IEEE database speech with factory noise), while the MRCG-DNN method always converges to a solution, so it was found to be much more robust.

Table 8. Results for AMS-GMM and MRCG-DNN.

		AMS-GMM			**MRCG-DNN**		
database	**noise**	**HIT**	**FA**	**HIT-FA**	**HIT**	**FA**	**HIT-FA**
IEEE	speech shaped-noise	93.4%	12.3%	81.12%	80.5%	3%	77.4%

In conclusion, the MRCG-DNN has a better HIT-FA than the DNN with other features (85 input features DNN), and it is more robust than the AMS-GMM method, with a similar performance. In addtion, the MRCG-DNN method is faster for training than the 85 input features DNN and the AMS-GMM methods: the MRCG-DNN takes about half an hour of training time, while the 85 input features DNN and AMS-GMM both need more than three hours of training time. Therefore, compared to other available methods, the MRCG-DNN was confirmed as a good choice for attempting to improve speech intelligibility through maximizing the HIT-FA rate.

6. CONCLUSIONS

The MRCG-DNN approach was found to outperform other approaches that have appeared in the literature for single-channel speech intelligibility enhancement processing, either in terms of objective measures (HIT-FA rate) or in terms of robustness to converge to a solution. Using the DNN method with “easier” noises (white noise, purple noise) lead to a better HIT-FA rate performance because of the stationarity of the noise and the reduced overlap with the speech content. For cases with more challenging noise conditions, the HIT-FA rate performance of the different approaches was not as good and the true intelligibility improvement was more debatable. Training with more data (clean speech data, noise data, or both) could be an option to attempt to improve the performance, although that option may not always be feasible in real-life scenarios.

ACKNOWLEDGEMENTS

Financial support for this work was provided by the Correctional Service Canada and by a NSERC Discovery grant.

REFERENCES

[1] Loizou, P. C. (2013) Speech enhancement: theory and practice. CRC press, 2nd ed.

[2] Kim, G., and Loizou, P. C. (2010) "Improving speech intelligibility in noise using environment-optimized algorithms." IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No.8, pp.2080-2090.

[3] Chen, J., Wang, Y., and Wang, D. (2014) "A feature study for classification-based speech separation at very low signal-to-noise ratio." Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.7039-7043.

[4] Hinton, G. E., and Salakhutdinov, R. R. (2006) "Reducing the dimensionality of data with neural networks." Science, Vol. 313, No. 5786, pp.504-507.

[5] Bengio, Y. (2009) "Learning deep architectures for AI." Foundations and trends in machine learning, Vol. 2, No.1, pp.1-127.

[6] Hinton, G. E., Osindero, S., and Teh, Y. W. (2006) "A fast learning algorithm for deep belief nets." Neural computation, Vol.18, No.7, pp.1527-1554.

[7] Arel, I., Rose, D. C., and Karnowski, T. P. (2010) "Deep machine learning-a new frontier in artificial intelligence research." IEEE Computational Intelligence Magazine, Vol. 5, No. 4, pp.13-18.

[8] Deng, L. (2011) "An overview of deep-structured learning for information processing" Proceedings of Asian-Pacific Signal & Information Processing Annual Summit and Conference (APSIPA-ASC).

[9] Yu, D., and Deng, L. (2011) "Deep learning and its applications to signal and information processing." Signal Processing Magazine, Vol.28, No.1, pp.145-154.

[10] Deng, L., and Yu, D. (2014) "Deep learning: methods and applications." Foundations and Trends in Signal Processing, Vol. 7, No.3–4, pp.197-387.

[11] Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (2013) "An algorithm to improve speech recognition in noise for hearing-impaired listeners." The Journal of the Acoustical Society of America, Vol. 134, No. 4, pp.3029-3038.

[12] Carreira-Perpinan, M. A., and Hinton, G. E. (2005) "On contrastive divergence learning." Proceedings of the 10th international workshop on artificial intelligence and statistics, pp.33-40.

[13] Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994) "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise." The Journal of the Acoustical Society of America, Vol. 95, No.2, pp.1085-1099.

[14] TIMIT Acoustic-Phonetic Continuous Speech Corpus https://catalog.ldc.upenn.edu/LDC93S1

[15] AURORA Project Database 2.0 - Evaluation Package, http://catalog.elra.info/product_info.php?products_id=693

[16] DeepLearnToolbox, https://github.com/rasmusbergpalm/DeepLearnToolbox

[17] Deep Neural Network toolbox, http://www.mathworks.com/matlabcentral/fileexchange/42853-deep-neural-network

[18] Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (2009) "An algorithm that improves speech intelligibility in noise for normal-hearing listeners." The Journal of the Acoustical Society of America, Vol.126, No. 3, pp.1486-1494.

[19] Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (1969) "IEEE recommended practice for speech quality measurements." IEEE Trans. Audio Electroacoust., Vol. 17, No.3, pp.225-246.

AUTHORS

Dongfu Li received the M.A.Sc. degree in Electrical and Computer Engineering from the University of Ottawa in 2016, a M. Eng. degree from Chongqing University of Posts and Telecommunications in 2010, and a B. Eng. from Chongqing University of Posts and Telecommunications in 2007. He has completed internships at Chongqing Jiuzhou Starnav System Co. in 2010-2011, at Chongqing University of Posts and Telecommunications in 2008-2009, and at Chongqing Aerospace New Century Satellite Application Technology Co. in 2007-2008.

Martin Bouchard received the B.Eng., M.App.Sc. and Ph.D. degrees in Electrical Engineering from Université de Sherbrooke (Sherbrooke, Qc., Canada) in 1993, 1995 and 1997 respectively. In January 1998, he joined the School of Electrical Engineering and Computer Science (formerly known as the School of Information Technology and Engineering) at the University of Ottawa (Ottawa, Canada). Over the years Professor Bouchard has conducted research activities and consulting activities with several private and governmental partners. He has served as a member of the Speech and Language Technical Committee (SLTC, IEEE Signal Processing Society, 2009-2011), as an Associate Editor for the EURASIP Journal on Audio, Speech and Music Processing (2006-2011), and an Associate Editor for the IEEE Transactions on Neural Networks (2008-2009). His current research interests are signal processing methods in general, with an emphasis on speech, audio, acoustics and hearing aids applications.

INTENTIONAL BLANK

ROBUST VISUAL TRACKING BASED ON SPARSE PCA-L1

Yuanyuan Zhang[1] and Fuxiang Wang[2]

[1]National Key Lab of CNS/ATM, School of Electronic and Information Engineering, Beihang University, Beijing, China
buaa_zyy@163.com
[2]National Key Lab of CNS/ATM, School of Electronic and Information Engineering, Beihang University, Beijing, China
wangfx@buaa.edu.cn

ABSTRACT

Recently, visual tracking based on sparse principle component analysis has drawn much research attention. As we all know, principle component analysis (PCA) is widely used in data processing and dimensionality reduction. But PCA is difficult to interpret in practical application and all those principal components are linear combinations of all variables. In our paper, a novel visual tracking method based on sparse principal component analysis and L1 tracking is introduced, which we named the method SPCA-L1 tracking. We firstly introduce trivial templates of L1 tracking method, which are used to describe noise, into PCA appearance model. Then we use lasso model to achieve sparse coefficients. Then we update the eigenbasis and mean incrementally to make the method robust when solving different kinds of changes of the target. Numerous experiments, where the targets undergo large changes in pose, scale and illumination, demonstrate the effectiveness and robustness of the proposed method.

KEYWORDS

Visual tracking, sparse principal component analysis, particle filter

1. INTRODUCTION

Visual tracking is a very important part in computer vision field. The applications of surveillance, vehicle navigation, medical diagnostic, virtual reality and human computer interface are all based on visual tracking. And it is a hard problem because there are many different and varying circumstances that have to be considered in tracking algorithm, such as illumination variation, occlusion, pose changes, and background clutters.

There are two major categories of tracking methods now, discriminative and generative methods, used in current tracking techniques. Discriminative online learning methods [1] treat object tracking as a classification problem. Generative online learning methods are adopted to track an object by searching for region most similar to the target model. There are three main tracking strategies in generative online learning area. The first strategy is about tracking method based on templates, in which dynamic and multi-feature templates are very important to the tracking. The

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 125–135, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61011

second one is method based on sparse representation, in which most of the appearance information about the target is represented by a linear combination of only a few basis vectors. And the third one is method based on subspace analysis, which we are studying about. It includes the methods based on Non-negative matrix factorization (NMF), methods based on Kernel, and the ones based on principal component analysis (PCA).

Recently, the PCA draws more and more attention. There are many works following PCA. Ross et al. [4] proposed an adaptive probabilistic tracking approach to update the models of a target by means of incremental eigenbasis updates. Kwon and Lee [10] apply sparse PCA to formulate tracking. Under the subspace assumption, Sui and Zhang[11] propose a locally structured Gaussian Process and cast tracking as a regression problem. These methods based on PCA provide a compact representation of the target, which is efficient in feature extraction and removing redundant information. At the same time, the probabilistic model facilitates efficient computation.

But subspaces learning is essentially sensitive to occlusion. For this reason, some methods for occlusion handling need to be used with subspace learning together. Wang et al. [12] use the subspace model to represent the target and impose sparsity on the residual errors to deal with occlusions.

In this paper, under the framework of particle filter, we proposed a new tracking method based on sparse principal component analysis, which can work more robustly, especially with the occlusion. Our contributions are as follows:

1) Reconstruct new target models with PCA subspace learning and trivial templates, so we can, at the same time, describe the candidates and noise, the trivial templates can deal with the noise, like occlusion, during the tracking;

2) Use lasso model to get the sparse coefficient of principle components;

3) Incrementally learn and update the low-dimensional subspace representation, including correctly update the eigenbasis.

The remaining part of this paper is organized as follows. In Section 2, we review some relevant approaches that motivated our work. The details of our method are described in Section 3. Experimental results are reported in Section 4. We conclude this paper in Section 5.

2. RELATED WORK

In this section, we will briefly introduce the tracking method IVT (Incremental Visual Tracking) [4, 5] and L1 minimization visual tracking [7]. Both of the methods are in particle framework. The IVT is a traditional tracking method based on incremental subspace learning. It learns and updates the low-dimensional subspace representation of the targets. In order to estimate the locations of the target in the new coming frames, IVT predicts the candidates by using a sampling algorithm with likelihood estimates instead of gradient descent. It performs well with the variation of the target and surrounding illumination. But when the target is occluded by other things, it may drift away and miss the target in the end. And L1 minimization is a method by casting tracking as a sparse approximation problem. It introduces a set of trivial templates to deal with the occlusion problem. The trivial templates are used to capture the noise and occlusion. The L1 tracking is robust with the noise and partly occlusion. But the tracking speed is slowly and it is not real-time.

Our work is motivated by the IVT method and L1 minimization method. We use PCA to reconstruct the target templates, so the templates are orthometric with each other and need smaller storage space. At the same time, we introduce the trivial templates to handle the noise. And we choose the candidate with smallest residual as the new target. Then we update the subspace by incrementally updating the eigenbasis.

3. OUR WORK

In this section, the details of our tracking method based on sparse PCA-L1 will be introduced. Our tracking method is conducted within the particle filtering framework. The main parts of our method are as follows.

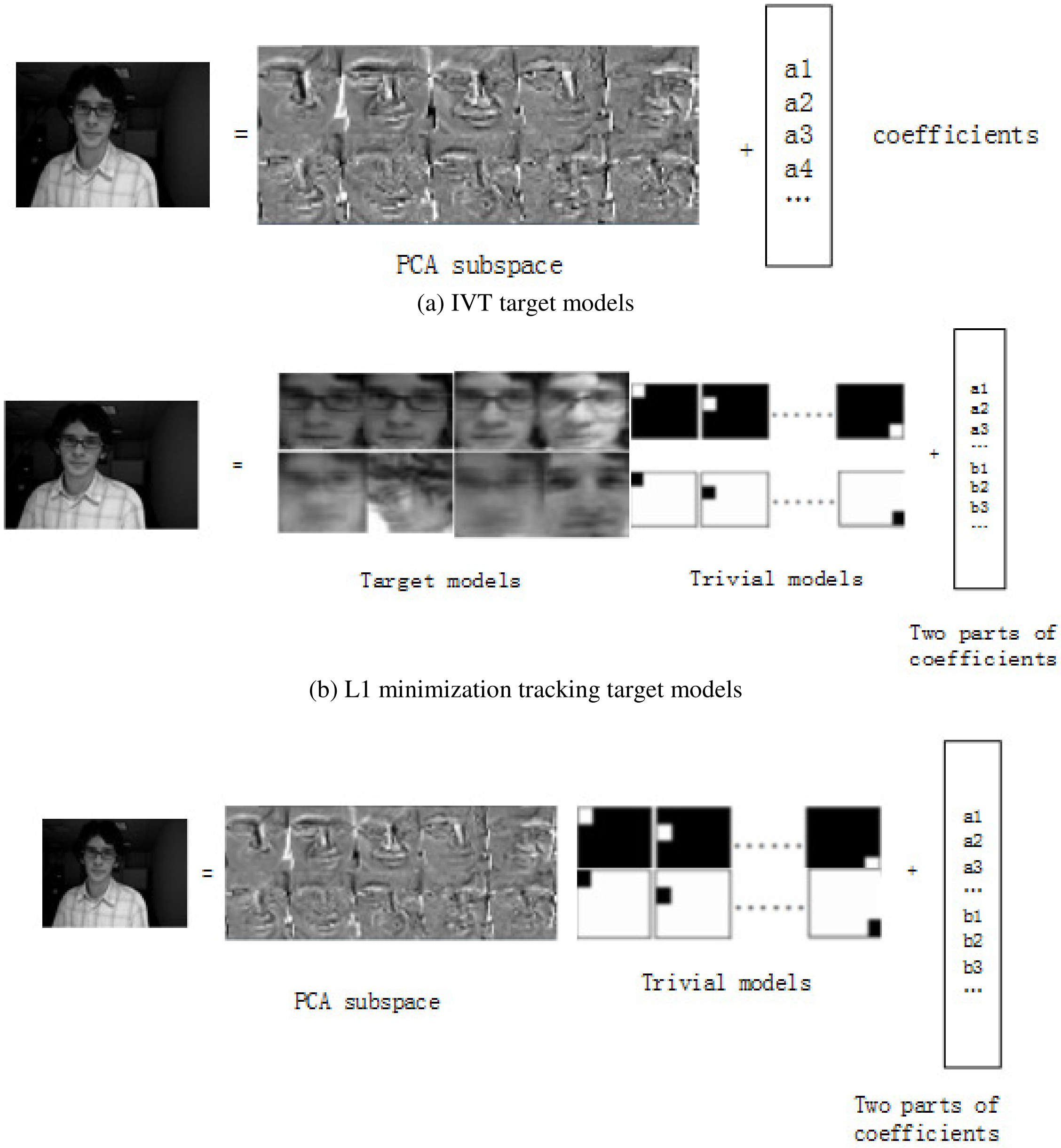
A. New target models subspace and trivial templates

We want to reconstruct new target models involving PCA subspace and trivial templates, so we can describe the candidates and noise at the same time. The trivial templates can deal with the noise, like occlusion, during the tracking at the same time. The Fig.1 shows the difference between the models of L1 tracking, IVT tracking and our tracking method, Sparse PCA-L1. In Fig.1 (a), we can see the subspace representation of the target object in IVT tracking. Fig.1 (b) shows the models of L1 minimization tracking, which includes trivial templates to capture the occlusion. Our tracking models are represented in Fig.1 (c). At initialization, we manually select the first target template from the first frame; then apply zero mean unit norm normalization, it is the first template. The rest templates are created by moving one pixel in four possible directions at the corner points of the first template in the first frame. These are the basic models. Then we use PCA to learn those basic models, making them orthotropic to each other. Meanwhile we introduce trivial templates to represent the occlusion. Thus our models need less space than L1 tracking models and more robust than IVT when the target is occluded.

We use $\boldsymbol{T = [t_1, t_2..t_n] \in R^{d\times n}(d \gg n)}$ to represent the new template set, and use $\boldsymbol{I = [i_1, ...i_d] \in R^{d\times d}}$ to represent the trivial template set. So the candidate can be describe by the following formulation:

$$\boldsymbol{y = (T, I)\left(\begin{matrix} a_T \\ a_I \end{matrix}\right) + e} \qquad (1)$$

Where $\boldsymbol{y}$ is the observed candidate, $\boldsymbol{a_T}$ is the coefficient of the main templates, $\boldsymbol{a_x}$ is the coefficient of the trivial templates and $\boldsymbol{e}$ is the representation residual.



(a) IVT target models

(b) L1 minimization tracking target models

(c) SPCA-L1 target models

By using the new models, we can collect the most information of the target and candidates. Meanwhile when the target is covered, the trivial templates can represent the noise and make our tracking robust. It will be proved in the experiment part following.

B. Lasso model for sparse representation

As we introduced in the former part, after learning the basic templates, we use subspace learning method PCA to reduce the dimension of template set. Then we need to get the sparse coefficients a in formation (1).

We know that the formation has more than one solution. But we want to get the sparse solution, so we exploit the compressibility in the transform domain by solving the problem as an l_1 regularized least squares problem, which is known to typically yield sparse solutions. As the coefficients include the ones of trivial templates, so the problem in our tracking method is transformed to the following formation:

$$\hat{a} = \arg\min_{a}\{\frac{1}{2}\| y - Ba \|_2^2 + \lambda \| a \|_1 + \frac{d}{2}\| a_I \|_2^2 \} \qquad s.t. \qquad a \geq 0 \tag{2}$$

Where $B = [T, I]$ in formation (1). $\| . \|_1$ and $\| . \|_2$ denote the l_1 and l_2 norms respectively.

At last, we choose the candidate which has the smallest residual as the new target.

$$identity(y) = argmin\{\| e \|_2^2 \} = argmin\{\| y - Ba \|_2^2 \} \tag{3}$$

C. Online Tracking models and incrementally update of the mean and eigenbasis

The appearance of a target object often change drastically due to different kinds of factors. Therefore, we need to adapt the appearance online to reflect these changes. The appearance model we have chosen, a eigenbasis, is typically learned off-line from a set of training images, $d \times n$ matrix, $T = \{t_1, ..., t_n\}$, by taking computing the eigenvectors U . Then we present our approach for drawing particles in the motion parameter space and predicting the most likely candidate with the help of the learned appearance model.

The sample mean of the training images is $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$, and the sample covariance matrix is $\frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{t})(t_i - \bar{t})^T$,and its eigenvector is describe as U . Equivalently one can obtain U by computing the singular value decomposition $T = U\Sigma V^T$ of the centered data matrix $[(t_1 - \bar{t})...(t_n - \bar{t})]$, with columns equal to the respective training images minus their mean. When the additional m images, $d \times m$ matrix, $W = \{t_{n+1}, ... t_{n+m}\}$ comes, we retrain the eigenbasis. This update could be performed by computing the singular value decomposition $[T \quad W] = U'\Sigma'V'^T$ of the centered data matrix $[(t_1 - \overline{t'})...(t_{n+m} - \overline{t'})]$, where $\overline{t}'$ is the average of the entire $n + m$ training images.

We describe a new method based on the Sequential Karhunen Loeve (SKL) algorithm. Letting $\tilde{W}$ be the component of W orthogonal to U , we can express the concatenation of T and W in a partitioned form as follows:

$$[T \quad W] = [U \quad \tilde{W}][\begin{matrix} \Sigma & U^T W \\ 0 & \tilde{W}^T W \end{matrix}][\begin{matrix} V^T & 0 \\ 0 & I \end{matrix}] \tag{4}$$

Let $R = \begin{bmatrix} \Sigma & U^T W \\ 0 & \tilde{W}^T W \end{bmatrix}$, which is a square matrix of size $k+m$,where k is the number of singular values in Σ . The SVD of R, $R = \tilde{U}\tilde{\Sigma}\tilde{V}^T$, can be computed in constant time regardless of n , the initial number of data. Then the SVD of $[T \quad W]$ can be expressed as follows:

$$[T \quad W] = ([U \quad \tilde{W}]\tilde{U})\tilde{\Sigma}(\tilde{V}^T \begin{bmatrix} V^T & 0 \\ 0 & I \end{bmatrix}) \qquad (5)$$

By using this algorithm, we can calculate the updating process more efficiently. The following table (1) shows the storage space and computational complexity change before and after the optimization. It is clear that both storage space and computational complexity are reduced, and it is helpful to track more quickly.

Table1: Comparison before and after the optimization

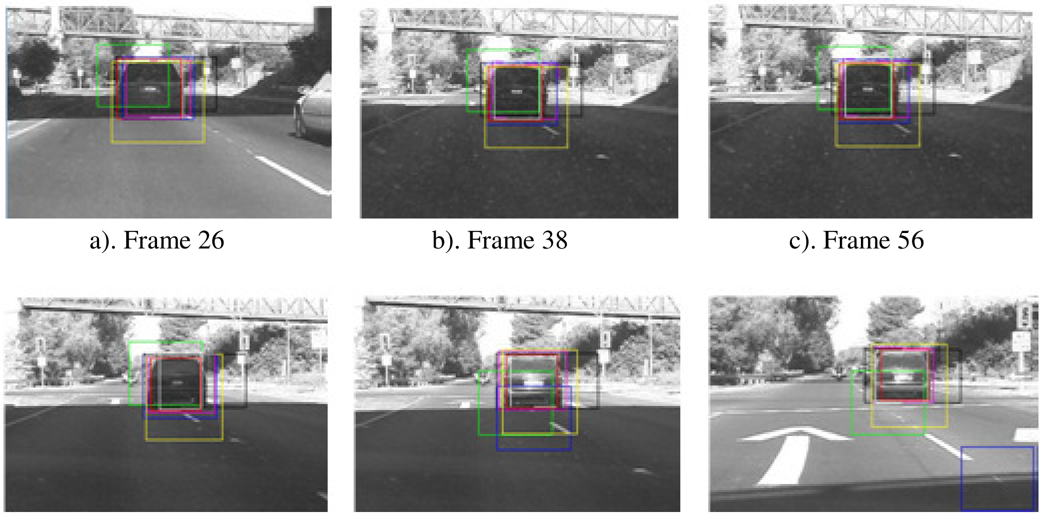
	Update object	Storage space	Computational complexity
Before	$[T \quad W] = U'\Sigma'V'^T$	$O(m(n+q)^2)$	$O(m(n+q)^2)$
After	$[T \quad W] = [U \quad \tilde{W}]\begin{bmatrix} \Sigma & U^T W \\ 0 & \tilde{W}^T W \end{bmatrix}\begin{bmatrix} V^T & 0 \\ 0 & I \end{bmatrix}$	$O(m(k+q))$	$O(mq^2)$

4. RESULTS

In this section, we present the experiment results achieved by applying our method. To demonstrate the performance of our method, we track the classical videos in the visual tracking, involving 1) target pose change 2) different light conditions 3) scale change and so on. We do the experiment on the computer: Inter(R) Core(TM) i5-4590 CPU, 3.30GHz, 4GB. And we use Matlab to do the simulation. All the results are as follow: the red color stands for our method, the pink is IVT, the yellow one represents L1 tracking, the black is APG, the blue is CT, and the green is MIL. We pay especially attention to the comparison of our method with IVT, which based on subspace.

A. Intuitive comparison

As the Fig.2 shows, the first car4 video shows a car passing beneath a bridge with sudden illumination change. Our algorithm can track the car efficiently. While the MIL drifts a little when the car undergoes illumination, and CT drifts and at last loses the target.

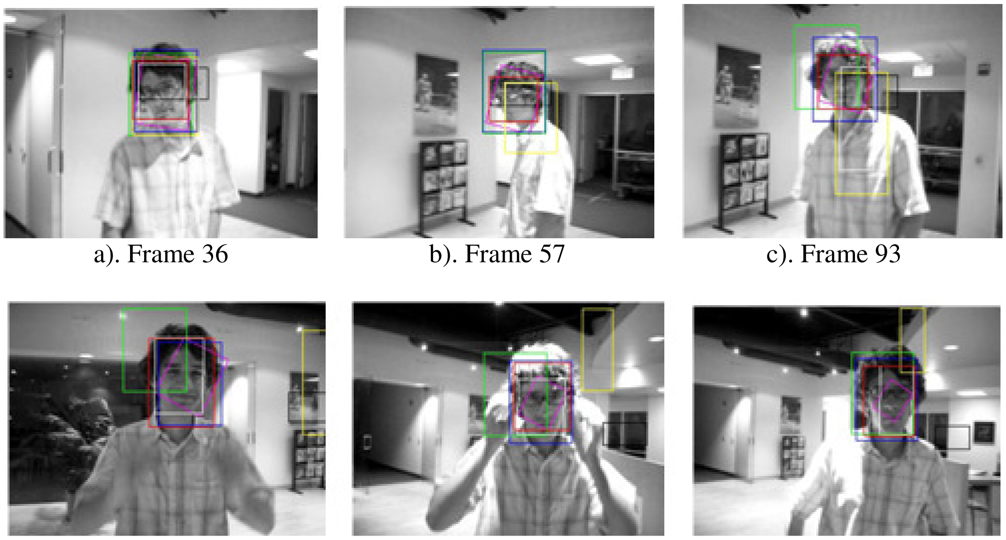


a). Frame 26 b). Frame 38 c). Frame 56

d). Frame 71 e). Frame 76 f). Frame 96

Fig.2 Car4 tracking results

In Fig.3, the second David indoor video involves light change and target pose change, recorded at 15 frames per second with a moving digital camera. The man moves from a dark area to a bright area, the back to the dark area, who undergoes lighting and pose changes. Notice that there is a large scale variation in the target. It’s clear that our algorithm is able to track the target throughout the sequences, no matter of light change or scale variation. The MIL method and L1 tracking experience drift, and the L1 even loses the target at #311.

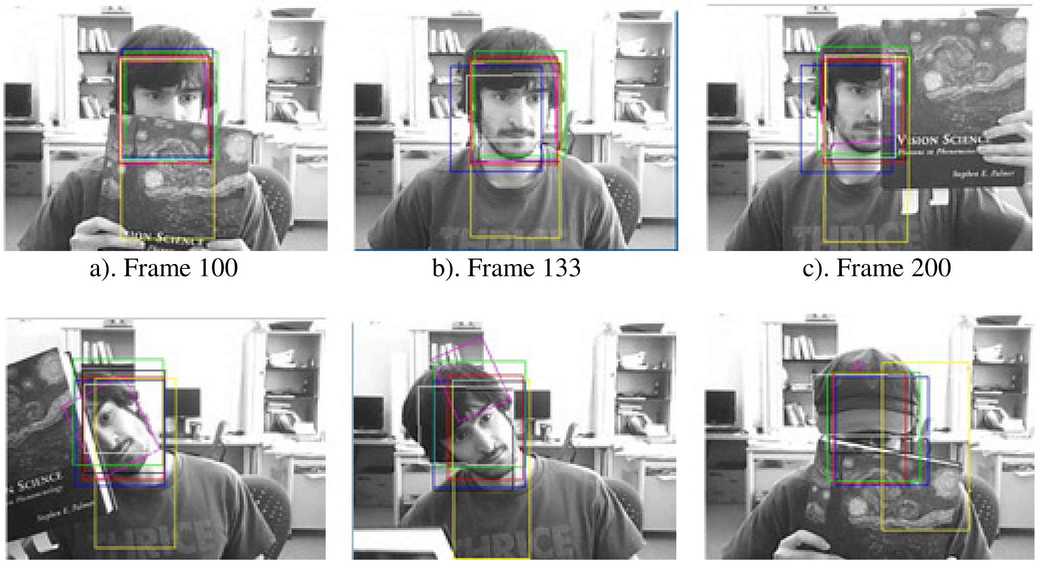


a). Frame 36 b). Frame 57 c). Frame 93

d). Frame 189 e). Frame 300 f). Frame 311

Fig.3 David indoor tracking results

About Fig.4, the third video occluded face2 involves occlusion. In the video, the target is occluded by a book. We can find that when the man’s face is covered by a book and a hat, L1 tracking and CT drifts away from the target. And when the man tilted his head, the IVT is not robust. It is the same condition in Fig.3 when David turns left at #189, and since then the tracking rectangle is not identical to the target. During the process, our method perform well and deal with occlusion problem perfectly.

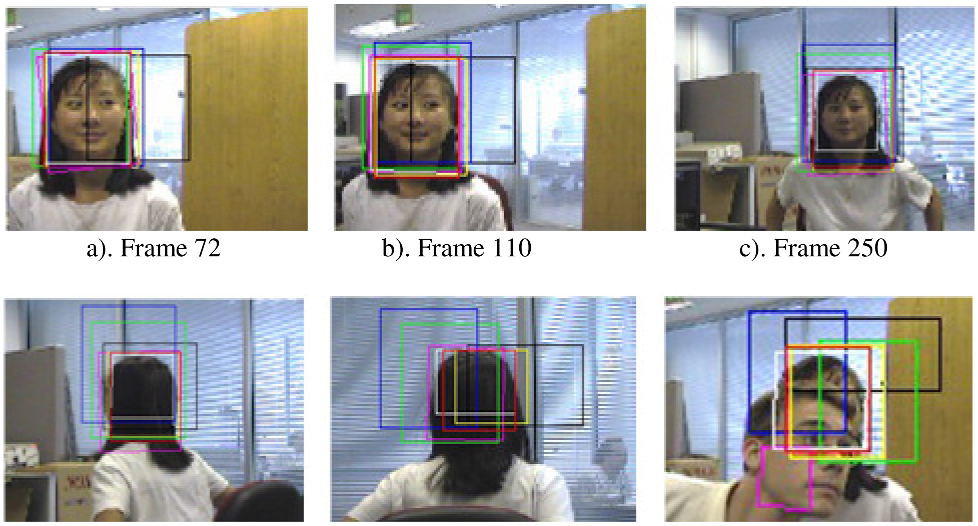


a). Frame 100 b). Frame 133 c). Frame 200

d). Frame 345 e). Frame 450 f). Frame 638

Fig.4 Occluded face2 tracking results

In Fig.5, It is a video of a girl, who turns her head around and later is covered by a man's face. MIL and APG perform a little bad, especially when the man covers the girl's face. And the IVT mistakes the man as the target instead of the girl because of occlusion. And our method locks the target even she turns around. And just drifts a little when the man appears.



a). Frame 72 b). Frame 110 c). Frame 250

d). Frame 320 e). Frame 390 f). Frame 455

Fig.5 Girl tracking results

B. Accuracy comparison

To evaluate the accuracy of our tracking algorithm, we consider the center position errors of each tracking method. As we can see, the center position errors of CT and MIL in sequences ‘Car4, David indoor and Girl’ increase at or after 100 frames. The L1 tracking method does a bad job in David indoor. Meanwhile, the IVT performs badly when the target is occluded because the center position errors increase in Occluded face2. But in these 4 videos, our tracking method does better than the other 5 methods.

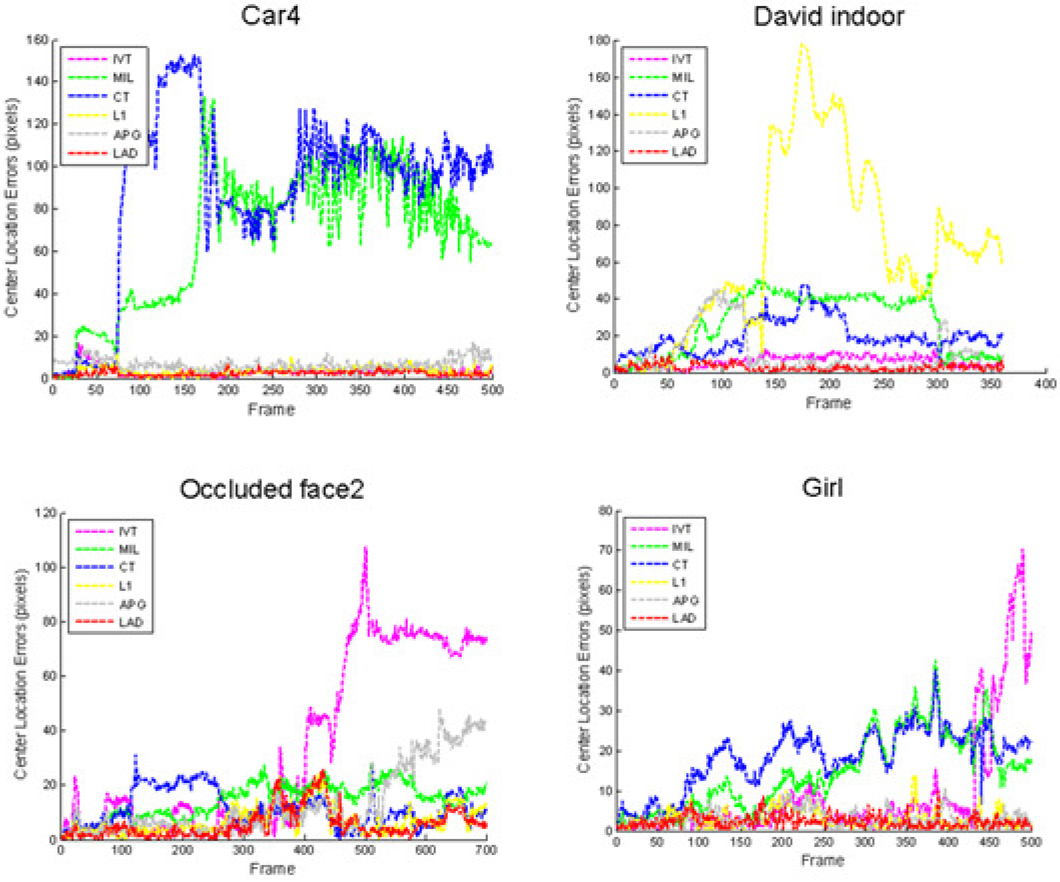


Fig.6 Center position errors of four challenging sequences

Besides, we calculate the average center position errors of the above results. As showing in Table 2, the red one represents the smallest average center position errors, the blue is the second best. So our tracking enhances a little in accuracy than IVT and L1 tracking methods.

Table 2: Average center position errors

Trackers Sequences	LAD (pixels)	IVT	MIL	CT	L1	APG
Car4	2.8	3.4	67.5	90.2	3.6	6.7
David indoor	3.3	6.1	27.0	7.5	66.7	10.2
Occluded face2	5.9	33.7	14.3	11.1	6.9	13.0
Girl	2.9	9.3	14.6	18.3	3.0	3.2

4. CONCLUSION

The tracking method sparse PCA-L1 is a robust tracking method. Because we add the trivial template and PCA subspace together, the tracking method does a good job when the target undergo pose, illumination and appearance change and occlusion. We mix the advantages that PCA subspace can represent most information of the target and trivial template can describe the noise like occlusion and so on. At the same time, we update the eigenbasis and mean accurately and efficiently. So when the target changes, we can gradually form the new models according to the eigenbasis.

REFERENCES

[1] Z. Kalal, J. Matas, and K. Mikolajczyk. P-N learning: Bootstrapping binary classifiers by structural constraints. In CVPR, 2010.3,7.

[2] Grabner,H., Leistner,C., Bischof,H.: ‘Semi-supervised online boosting for robust tracking’. Proc. European Conf. Computer Vision (ECCV), Marseille, France, October 2008,pp.234-247

[3] Sankarayanan, K., Davis, J.W.: ‘One class multiple instance learning and applications to target tracking’. Proc. IEEE Conf. Computer Vision(ACCV), Daejeon, Korea, November 2012, pp.1-14

[4] Ross,D. , Lim, J. , Yang, M.H : ‘Adaptive probabilistic visual tracking with incremental subspace update’. Proc. European Conf. Computer Vision(ECCV), Prague, Czech Republic, May 2004, pp.470-482

[5] Lim, J., Ross, D., Lin, R.S., Yang, M.H.: ‘Incremental learning for visual tracking’, in Weiss, Y., Bottou, L., (Eds.), ‘Advances in neural information processing systems’ (MTI Press, 2005), pp.793-800

[6] Mei, X., Ling, H.B.: ‘Robust visual tracking and vehicle classification via sparse representation’, IEEE Trans. Pattern Anal. Mach. Intell., 2011,33,(11), pp.2259-2272

[7] Mei, X., Ling, H.B.: ‘Robust visual tracking using L1 minimization’. Proc. IEEE Int. Conf. Computer Vision(ICCV), Kyoto, Japan, September 2009, pp.1436-1443w

[8] Zou,H., and Hastie, T.(2005),'Regularization and Variable selection via the Elastic Net' Journal of the Royal Statistical Society, Series B,67,301-320.

[9] Jolliffe,I.T., Tredafilov, N.T., and Uddin, M.(2003),'A Modified Principle Component Technique Based on the Lasso' Journal of Computational and Graphical Statistics, 12,531-547.

[10] J. Kwon and K. Lee. Visual tracking decomposition. In CVPR,2010.2,3,7.

[11] Y. Sui and L. Zhang.' Visual Tracking via Locally Structured Gaussian Process Regression'. IEEE Signal Processing Letters,22(9):1331-1335,2015.3,7

[12] D. Wang, H. Lu and M. Yang. 'Online object tracking with sparse prototypes'. IEEE Transactions on Image Processing(TIP), 22(I):314-325,2013.2,3

[13] T. Zhang, B. Ghanem, S. Liu, and N. Ahuji, 'Low rank sparse learning for robust visual tracking'. In ECCV,2012.3,6,7.

[14] S. Hare, A. Saffari, and P. Torr. Struck:' Structured output tracking with kernels'. In ICCV,2011.3.

[15] Yao Sui, Yafei Tang, Li Zhang. 'Discriminative Low Rank Tracking'. ICCV 2015.3002-3010.

[16] A. d' Aspremont, L. EI Ghaoui, M.Jordan, and G.Lanckriet.' A direct formulation for sparse PCA using semidefinite programming'. SIAM Review ,46(3),2007.

AUTHORS

Yuanyuan Zhang received B.S. degree in Electronic science and technology from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2014. Now she is studying the M.S. degree in School of Electronics and Information Engineering of Beihang University, Beijing, China. Her current research interests lie in the areas of image processing and computer visual tracking.

Fuxiang Wang received the B.S. degree in 1999 and the Ph.D. degree in 2007, all from Beihang University, Beijing, China. He is currently a lecture with the School of Electronics and Information Engineering, Beihang University, Beijing, China. His current research interests lie in the areas of blind separation and their applications.

INTENTIONAL BLANK

USING CISCO NETWORK COMPONENTS TO IMPROVE NIDPS PERFORMANCE

Waleed Bul'ajoul[1], Anne James[1], Siraj Shaikh[1] and Mandeep Pannu[2]

[1]Faculty of Engineering and Computing, Coventry University, Coventry, UK
bulajouw@coventry.ac.uk
a.james@coventry.ac.uk
aa8135@coventry.ac.uk
[2]Department of Computer Science, Kwantlen Polytechnic University, Surrey, British Columbia, Canada
mandeep.pannu@kpu.ca

ABSTRACT

Network Intrusion Detection and Prevention Systems (NIDPSs) are used to detect, prevent and report evidence of attacks and malicious traffic. Our paper presents a study where we used open source NIDPS software. We show that NIDPS detection performance can be weak in the face of high-speed and high-load traffic in terms of missed alerts and missed logs. To counteract this problem, we have proposed and evaluated a solution that utilizes QoS, queues and parallel technologies in a multi-layer Cisco Catalyst Switch to increase NIDPSs detection performance. Our approach designs a novel QoS architecture to organise and improve throughput-forward-plan traffic in a layer 3 switch in order to improve NIDPS performance.

KEYWORDS

Network security, intrusion detection system, network intrusion detection system, open source, Cisco switch configuration & Quality of Service.

1. INTRODUCTION

Despite the existence of a variety of security protection measures, attackers often attempt to render services unavailable to the intended, legitimate users [1, 2, 3, 4]. In general, there are three types of network security techniques: prevention, detection and correction techniques. The prevention technique actively works to block intrusions, but it can also be used to battle a successful intrusion. A number of successful attacks can be controlled using the prevention technique if an attack is detected at the interim stage of prevention systems. Unfortunately, some successful attacks can get through the prevention system [2, 5]. In this instance, depending on preventive techniques is unlikely to resolve the issue, especially when an attacker has successfully obtained vulnerable information from the network; however, prevention can successfully and effectively maintain a network before an attack is launched. The correction technique is adopted to protect computer systems. It is used when the prevention technique has failed. In these cases, the system is attacked and compromised; consequently, it malfunctions. The correction technique restores the system to a stable state when an attack has been detected. Clearly, both prevention and correction require a detection phase, which should be constantly active to combat intrusions.

David C. Wyld et al. (Eds) : CSEN, SIPR, NCWC - 2016
pp. 137–157, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61012

Network intrusion detection and prevention systems (NIDPS) are commonly used to detect and prevent attacks. A popular system is the open-source Snort NIDPS. Our previous studies [4, 5] have been carried out on the use of a Snort NIDPS in high speed networks. We found that in high speed and high volume environments, the Snort NIDPS drops packets. Our studies focused on improving the performance of the NIDPS in analysis mode. Improving the analysis phase for any security production is important, because it is difficult to detect or prevent threats or malicious traffic without analysing the traffic. This paper however focuses on NIDPS detection rather than analysis. Thus we have now considered rule-based actions as well as the passive analysis of the packets received. Our paper uses Snort NIDPS. We conducted experiments to test Snort's detection-mode performance reaction to ICMP, UDP and TCP headers and malicious packets under high-load and high-speed traffic. We further demonstrate that Snort's performance can be improved by using additional technologies such as a Quality of Service (QoS) configuration and parallel technologies.

The remainder of this paper is organised as follows. Section 2 gives an overview of previous related work. Section 3 provides a background on: intrusion detection and prevention systems (IDPSs); network intrusion detection and prevention systems (NIDPSs); and information about Snort NIDPS. Section 4 explains our experimental design. Section 5 presents a first set of experiments which demonstrated some NIDPS weaknesses. Section 6 presents our solution for combatting such weaknesses and also provides an experimental evaluation of our solution. This is followed by section 7 which provides more information about the technical aspects of the approach. Finally Section 8 concludes the paper and suggests further work.

2. RELATED WORK

Chen, et al. [17] proposed an application-specific integrated circuit (ASIC) design with parallel exact matching (PEM) architecture to accelerate throughput. The ASIC hardware has been designed to operate at 435MHz to perform up to 13.9 Gbps throughput. The aim is to manage the requirements of high speed and high accuracy for Snort IDS (Intrusion Detection System) and overcome the complexity of managing data received from the 10Gbps core network. They proposed the SRA (Snort Rule Accelerator) which processes rules in parallel to increase the performance of the Snort IDS. The SRA has a stateless parallel-matching scheme to perform high throughput packet filtering as an accelerator of the Snort detection engine. The ASIC is combined of five major modules, including the inspector, counter, parallel matching, conformity and compare modules. The functionality of the parallel matching scheme is to compare payloads of packets with the stored rules. When an entry packet is matched with Snort rules, the ASIC is in an idle state and sends a compare and signal to the conformity module, which integrates all signals and determines whether an abnormal payload is presented. Here the authors designed half mesh architecture in the parallel matching rules module, which allows the traffic to be compared with several rules at the same time. Our work addresses performance in a different way. Instead of processing rules in parallel, it processes traffic in parallel. It explores the use of hardware Layer-3 network switches and Cisco configuration with parallel queue technologies to improve QoS and, hence, NIDPS performance.

Jiang et al. [18] proposed a new NIDS architecture based on multi-parallel core processors. They exploited many-core computational power by adopting a hybrid parallel architecture combining data and pipeline parallelism. They designed a system for parallel network traffic processing by implementing an NIDS on the TILERAGX36 (a 36 core processor) [19]. The system was designed according to two strategies: first a hybrid parallel architecture was used, combining data and pipeline parallelism; and secondly a hybrid load-balancing scheme was used. They took advantage of the parallelism offered by combining data, pipeline parallelism and multiple cores, using both rule-set and flow space partitioning. They showed that processing speeds can handle and reach up to 7.2Gbit/S with 100-byte packets. Our approach differs from theirs in that we have shown how can exploit QoS and queues technologies in a multi-layer switch to improve packets processing throughput. Further enhancements can occur if we combine queuing together with parallel technologies. Our approach requires less specialised equipment.

In previous work [4], we presented experiments that demonstrated that Snort dropped packets when it was deployed in high speed traffic in analysis mode. Packets were dropped or left outstanding when the in-coming speed of traffic is higher than NIDPS processing speed limit. We used QoS technology in a Cisco Catalyst switch to improve NIDPS performance and proposed a solution which reduces the packet arrival speed to the NIDPS node processor limit. This approach reduces dropped packets or the number of packets left outstanding by using queues to reduce the arrival rate to a speed that can be processed successfully. We then used parallel technology to increase the throughput rate.

This paper presents similar technology to that applied in our previous work [4] but this time it is applied to Snort in detection mode. The QoS configuration in layer 3 Cisco switches provides the capability to differentiate among different classes of traffic and to prioritize the traffic in times of network congestion, according to relative importance. The research described in this paper uses QoS queue technology to increase monitoring speed rather than reducing or slowing down traffic speed [4]. We used parallel technology to speed up the throughput of NIDPS packets processing to the level of the arrival traffic speeds. Our novel architecture addresses the weakness of NIDPS performance detection caused by increasing network traffic speed. The strength of this work over previous work is that in previous work [4, 5], it was shown that analysis can fail in high speed and high volume traffic in terms of the fact that many packets are dropped or left outstanding and not analysed. But in analysis mode, dropped packets refer to packets that eventually do not get through to the destination and outstanding packets are those left unprocessed in the system. In this paper we show that detection can also fail. This is more significant because failing to detect, i.e. failing to recognize malicious packets, means that attacks can penetrate the system. The contribution of this work is to offer a solution to the problem of detection in high speed traffic based on network switch, QoS and parallel technologies. In the next section the background to our research is provided.

3. BACKGROUND

Security products, such as firewalls and antivirus programs, are less efficient than intrusion detection and prevention systems (IDPSs) and have different functionalities. IDPSs analyse collected information and infer more useful results than other security products [4, 5]. However, some researchers have indicated that whilst IDPSs have significantly improved with the passage of time, they still often produce an unacceptable quantity of false positives and false negatives [4, 10, 11, 12]. In addition, it is difficult to detect suspicious activities in the midst of high traffic and other such adverse circumstances in the network, consequently resulting in an inaccurate detection mechanism.

IDPSs consist of either software applications or hardware that listen to and detect malicious activities at the gateway (incoming and outgoing) of individual or network systems. IDPSs are capable of monitoring, identifying and reporting evidence of malicious activities and attacks, such as flood attacks, unauthorised log-ins, privilege escalation, illegitimate access, modification of data and data-driven attacks, [4, 5]. Therefore, an IDPS sniffing mechanism is effectively applied at the network gateway, which provides useful information about packets and traffic to security professionals [4].

The specialised IDPS mechanism is based on how, where and what it detects/prevents, along with mandatory requirements. In particular, IDPSs should be based on flexible and scalable network components to accommodate the drastic increase in today’s network environments. They should also provide straightforward management and operational procedures and steps, instead of procedures that complicate the underlying tasks. Lastly, they should provide user-friendly IDP mechanisms [4, 5].

3.1 Network Intrusion Detection and Prevention System (NIDPS)

NIDPSs are used to analyse the traffic in all Open Systems Interconnection (OSI) layers, to enable differentiation of normal traffic as opposed to suspicious activities. They are also used to detect and react to the unauthorised access to network systems, [4, 12]. There are three modes of NIDPS: analysis mode (sniffer mode); detection mode (passive mode); and prevention mode (inline mode).

Analysis mode is used to recognise and display the type of packets coming into the network. Various levels of detail can be displayed on the console, for instance application data which is attached to the packet additionally to TCP, UDP and ICMP header information [4, 7]. The detection system is capable of detecting suspicious activity and generating alerts based on recognised signatures and rules [4, 7]. Signature analysis is generally based on patterns inside the data packet. This technique aims to detect multiple kinds of attacks such as the presence of scripts in packets destined for web services [4, 7]. Alternatively, anomaly-based NIDPSs notice packet anomalies available in the header parts of the protocol [7]. Logging and alerts depend on the nature of what is detected inside the packets. If any suspicious activity is found inside a packet, the packet usually logs the malicious activity and/or generates an alert. Logs are usually stored in simple text-based files [4, 7, 8]. Output modules (plug-ins) are capable of performing multiple operations depending on the results generated by the logging and alerting system. In general, output modules control the form of outcome produced by the logging and alerting system [7, 8].Network intrusion prevention systems (NIDPSs) are active, inline devices in a network that can drop, block or reject packets and or stop malicious connections before these reach the targeted system [4, 12]. NIDPSs are further classified into software and hardware based. Hardware based NIDPSs are effective and can overcome some performance issues of software-based NIDPS but high cost is an issue. One of the most popular software-based NIDPSs is Snort [4].

3.2 Snort Network Intrusion Detection and Prevention System (Snort-NIDPS)

Snort is released as an open-source, rule-centred NIDPS, which stores information in text files that can be modified by a text editor [4, 8]. Snort rules activate on the network IP layer and TCP/UDP layer protocols. Rules are grouped into categories, and the rules belonging to each category are stored as information in separate files; these files are then integrated into the main configuration file, named “snort.conf”. The data is captured in terms based on described rules, which are read at the initialisation of Snort and are used to construct the internal data structure [4, 7, 8]. Furthermore, Snort is a combination of both basic signature code analysis and content-driven rules [7, 8]. Snort can execute a protocol analysis and a search and match of the content. It can be utilised for the detection of various attacks and probes, such as those regarding stealth port scans, buffer overflow, SMB probes, CGI attacks, fingerprinting attempts of OS and many more [4, 5]. Snort uses the rules, which have been written by using a flexible language that can be managed by developers and the executer. These rules identify the traffic types that can be passed or collected, and they can function as a detection engine, as well [4, 5, 7].

There are several features available for Snort; the most common feature is its real-time alerting mechanism. Alerts can also be collected by using a mechanism for syslog, which allows the reporting of suspicious activities in logs for additional investigation, a UNIX socket, a specified file of user, or a WinPopup message to the window client [7, 8]. Hence, Snort is different from other packet sniffers, due to the tcpdump sniffer, which has the capability to be run by different operating systems, and the use of the hexdump payload dump that tcpdump has employed during recent years. Snort also has the capability to display packets via different networks through the same method. Snort is a multi-mode IDPS and NIDPS providing analysis, detection and prevention [4, 7, 8]. It is capable of reading chains (internal data structures), which have to be matched against all packets. If a packet does not match any rule, it will be passed; otherwise appropriate action is taken. The default detection method of Snort NIDPS is the signature-based detection system, which can utilise rules to search or match any errant packets on the controlled network. In turn, the alerts are activated and sent to a receiver such as system log, database, management team or even a trap. Many studies have used Snort NIDPS to detect attacks such as DoS and DDoS by developing and designing new rules [13, 14, 15, 16].

4. EXPERIMENT DESIGN

This research carried out two sets of experiments. The first set was carried out to show the weakness of Snort NIDPS in detection mode in high speed and high volume traffic (results are given in section 5). The second set was carried out to show the effectiveness of our novel architecture (results given in section 6.2). A set of facilities was needed in order to demonstrate Snort NIDPS weakness and later show how such weakness can be overcome through the use of our novel architecture.

The experimental set up supports acquisition, analysis of the data, and detection of various types of traffic. Tools included: the Snort 2.9.7.2, which was issued in October 2014; the WinPcap tool, to capture packets on Windows and Linux OSs; the NetScanPro tool, to manage traffic in different time scales; the Packets Generator tool, to generate (ICMP, UDP and TCP) traffic at different speeds and values; and the Flooder Packets tool to generate flood traffic and malicious UDP packets (threads) in high-load traffic and high speed.

Figure 1 shows our experimental set up. Snort NIDPS was implemented in parallel on four work stations while a fifth workstation was used to generate traffic. All workstations were connected through a CISCO SW 3560 switch. Performance metrics were evaluated in the experiments to measure the ability of Snort NIDPS detection mode. The Snort breakdown analysis includes the following metrics: the number of packets analysed of the total packets received; the number of Eth (Ethernet) packets received of the total packets analysed; the number of IP packets received of the total Eth packets received; and the number of TCP, UDP and ICMP packets analysed. The Snort action statistics provides: the number of packet alerts of the total TCP/IP packets analysed; and the total packets logged of the total TCP/IP packets analysed. These parameters indicate NIDPS performance.

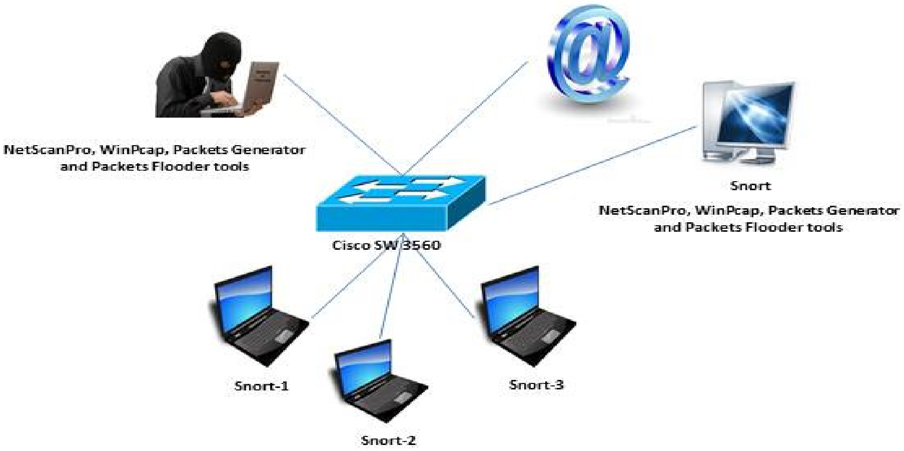

NetScanPro, WinPcap, Packets Generator and Packets Flooder tools
Snort
NetScanPro, WinPcap, Packets Generator and Packets Flooder tools
Cisco SW 3560
Snort-1
Snort-2
Snort-3

Figure 1. Simple network design

5. EXPERIMENTS, IMPLEMENTATION AND RESULTS

We ran (4) experiments to test Snort NIDPS reactions to (1) detect ICMP header under different speed traffic; (2) detect UDP header under different speed traffic; (3) detect TCP header under different speed traffic; and (4) detect malicious packets under different speed traffic. For each experiment, we ran three (3) consecutive tests; for each test the speed at which the packets were sent was increased each time. The packets were sent at interval times of various millisecond (ms). Each packet carries1KByte. Snort was run in detection mode. We created some rules to alert and log unwanted traffic. We also used the Packet Generator, WinPcap and Flooder packets tools to generate packets and threads (malicious UDP packets) through the network and hosts at different speeds.

5.1 Experiment 1: Detecting (alert and log) ICMP Header

In this experiment, more than 1 million IP/ICMP packets have been sent at different speeds (10ms, 5ms and 1ms intervals). In fact, this first common rule does a good job of testing if Snort is working well and if it is able to generate all alert actions:

Alert icmp any any ->any any (msg: ”Detect ICMP Packets”; sid:100001;).

Snort will alert and detect any ICMP packets from any sources to any destinations address from and to any ports.

Table 1. Snort reaction to ICMP header

Traffic speed per milliseconds	Machine packets received	% packet analysed	Eth packets received of the total packets analysed	ICMP packets analysed	TCP packet analysed	UDP packets analysed	Number of packets alert	Number of packets logged	% packet alerts	% packet logs
10ms	100%	4.300%	100%	1874	0	44459	1874	1874	100%	100%
5ms	100%	1.120%	100%	345	0	13463	231	231	66.96%	66.96
1ms	100%	0.141%	100%	730	0	1144	405	405	55.47%	55.47

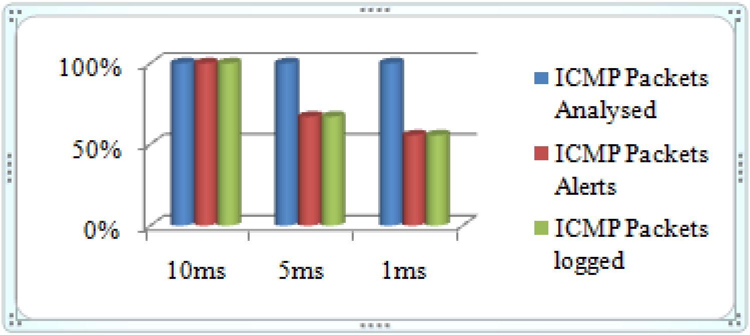


Figure 2. ICMP packets detection.

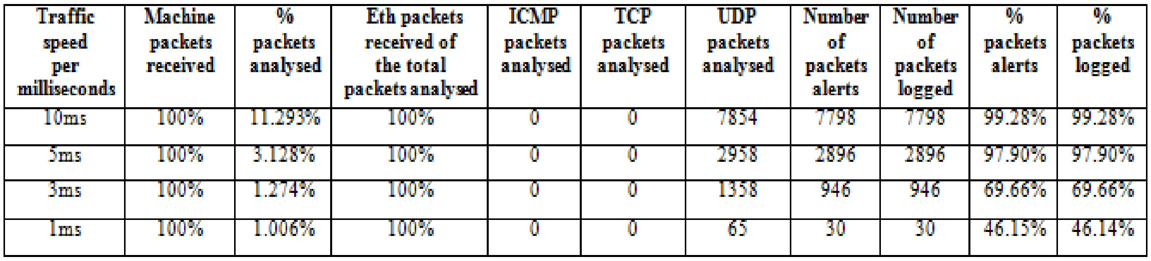
As the results show in Figure 2, Snort analysed every packet that reached the wire. When ICMP traffic was sent at 10ms, Snort alerted and logged nearly 100% of the total ICMP packets analysed (see Table 1). As the speed increased from 10ms to 1ms, Snort started missing alerts and logged packets. Also, Figure 2 shows that the number of missed alerts increased when the speed increased. The experiment shows that Snort detected 55.47 % of the total ICMP packets that it analysed (see Table 1).

5.2 Experiment 2: Detecting (alert and log) UDP Header

In this experiment, more than 1 million IP/UDP packets were sent at different speeds (10ms, 5ms, 3ms and 1ms), and the following rule was written to allow Snort to detect any UDP packets from any sources to any destination address and to any source and destination ports:

Alert udp any any ->any any (msg: ”Detect UDP Packets”; sid:100002;).

Table 2. Snort reaction to UDP header.



Traffic speed per milliseconds	Machine packets received	% packets analysed	Eth packets received of the total packets analysed	ICMP packets analysed	TCP packets analysed	UDP packets analysed	Number of packets alerts	Number of packets logged	% packets alerts	% packets logged
10ms	100%	11.293%	100%	0	0	7854	7798	7798	99.28%	99.28%
5ms	100%	3.128%	100%	0	0	2958	2896	2896	97.90%	97.90%
3ms	100%	1.274%	100%	0	0	1358	946	946	69.66%	69.66%
1ms	100%	1.006%	100%	0	0	65	30	30	46.15%	46.14%

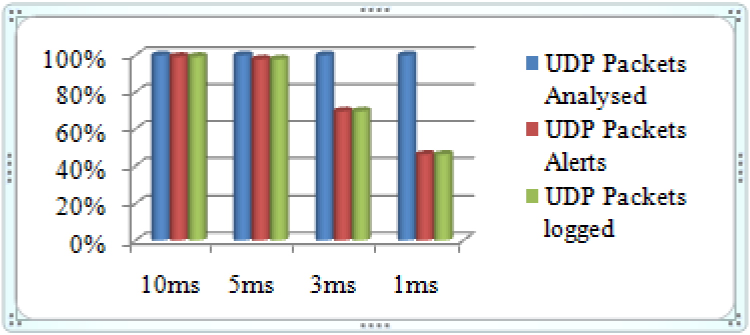


Figure 3. UDP packets detection.

As shown in Figure 3, when UDP traffic was sent at a speed of 10ms, Snort alerted and logged nearly 100% of the total UDP packets that it analysed (see Table 2). When the traffic’s speed increased to 5ms, Snort detected 97.90% of the total UDP packets analysed (see Table 2). Figure 3 shows that, as the speed increased, missed alerts and logs also increased. This experiment shows that Snort detected 46.14% of the total UDP packets that it analysed (see Table 2).

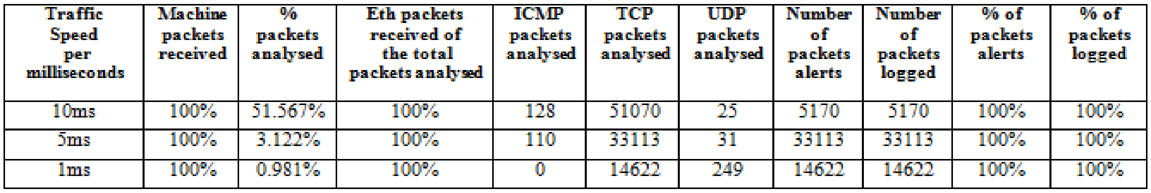
5.3 Experiment 3: Detecting (alert and log) UDP Header

Here, more than 1 million IP/TCP packets were sent at different speeds (10ms, 5ms and 1ms). The following rule was made to allow Snort to detect any TCP packets from any sources to any destinations, from and to any ports:

Alert tcp any any ->any any (msg: ”Detect tcp Packets”; sid:100003;).

As shown in Figure 4 Snort analysed every packet that reached the wire. The experiment shows that Snort detected all TCP packets that it analysed, even if the speed increased (see Table 3). This effectiveness occurred because TCP does not send the next packet until it receives an acknowledgement that the previous package has been received. These acknowledgements make the TCP packet slower than the UDP and ICMP packets.

Table 3. Snort reaction to TCP header



Traffic Speed per milliseconds	Machine packets received	% packets analysed	Eth packets received of the total packets analysed	ICMP packets analysed	TCP packets analysed	UDP packets analysed	Number of packets alerts	Number of packets logged	% of packets alerts	% of packets logged
10ms	100%	51.567%	100%	128	51070	25	5170	5170	100%	100%
5ms	100%	3.122%	100%	110	33113	31	33113	33113	100%	100%
1ms	100%	0.981%	100%	0	14622	249	14622	14622	100%	100%

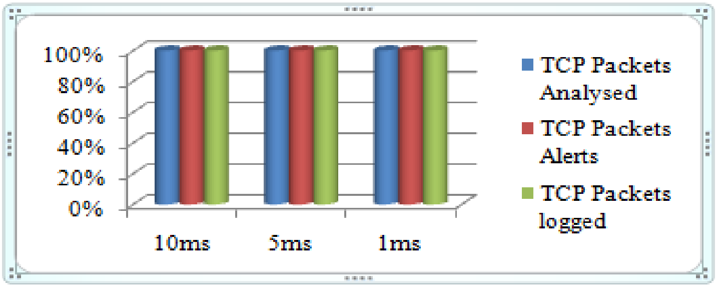


Figure 4. TCP packets detection.

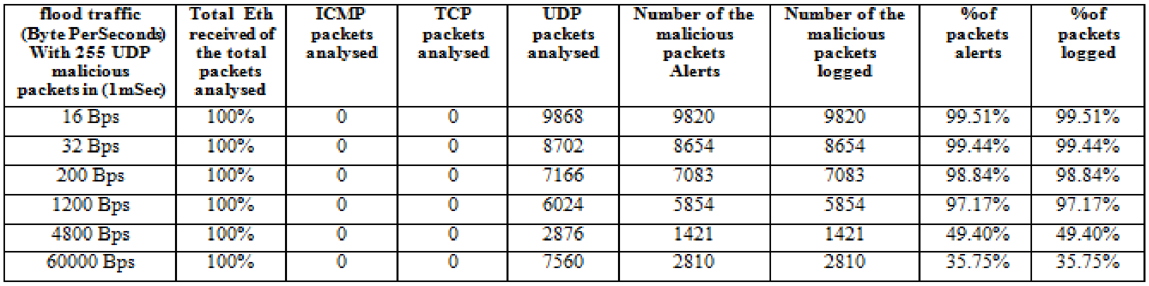
5.4 Experiment 4: Detecting Malicious Packet (UDP Threads)

In this experiment, WinPcap and Flooder packet tools were used to send flood traffic with malicious UDP packets (threads) to specific hosts or networks at different speeds. The UDP malicious packets contain variables and time to live 128. The following rule is written to permit Snort to alert and log any UDP threads or malicious packets that contain the variables ‘abcdef’ and time to live (TTL) 128 that comes from any source and port address and goes to any destination address and ports:

Alert udp any any ->any any (msg: “Detect Malicious UDP Packets”; ttl: 128; content:|’ 61 62 63 64 65 66 ’|; Sid: 100004 ;)

This experiment is different from the previous ones. The previous experiments tried to detect headers, such as TCP, UDP and ICMP. The system received the TCP, UDP and ICMP packets at different speeds, but in this experiment, we sent flood traffic in different bandwidths (speeds) with malicious UDP packets (threads) in interval packets with a delay of 1 microsecond (1 mSec), and then we tried to detect only the UDP threads by using two conditions of additional rules (TTL and content). These two key rules will detect any UDP malicious packet that is matched in order to determine that the TTL value is equal to 128 and to determine if a data pattern inside the malicious packet has variables (‘abcdef’). However, the hexadecimal number (‘61 62 63 64 65 66’), which the rule contained, is equal to the ASCII characters (‘a b c d e f’).

Table 4. Snort reaction to udp malicious packets.



flood traffic (Byte PerSeconds) With 255 UDP malicious packets in (1mSec)	Total Eth received of the total packets analysed	ICMP packets analysed	TCP packets analysed	UDP packets analysed	Number of the malicious packets Alerts	Number of the malicious packets logged	%of packets alerts	%of packets logged
16 Bps	100%	0	0	9868	9820	9820	99.51%	99.51%
32 Bps	100%	0	0	8702	8654	8654	99.44%	99.44%
200 Bps	100%	0	0	7166	7083	7083	98.84%	98.84%
1200 Bps	100%	0	0	6024	5854	5854	97.17%	97.17%
4800 Bps	100%	0	0	2876	1421	1421	49.40%	49.40%
60000 Bps	100%	0	0	7560	2810	2810	35.75%	35.75%

As shown in Figure 5, Snort analysed every packet that reached the wire. When malicious UDP packets were sent at a speed of 1 mSec and flood traffic at 16 bytes per second (Bps), Snort alerted and logged more than 99% of the total UDP packets that it analysed. As the flood traffic (speed) was increased to 200, 1200, 4800 and 60000 bytes per second (Bps), Snort alerted and logged packets to a decreasing degree, respectively, at 98.84, 97.17, 49.40 and 35.75% of the total malicious packets analysed (see Table 4). Figure 5 shows that the number of missed malicious packet alerts increased when the speed increased. The experiment shows that, when the speed was 60000 Bps, Snort only detected nearly 35 of 100% of the malicious packets analysed (see Table 4).

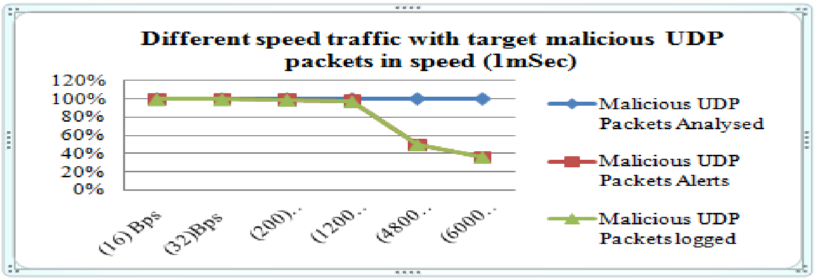


Figure 5. Malicious packets detection.

6. Proposed Solution and Evaluation

This section proposes a novel configuration as a solution to the problem of NIDPS dropped packets illustrated in section 5. It also presents an evaluation of the solution.

6.1 Proposed Solution

A critical analysis was conducted for experiments 1, 2 and 4 (see Figures 2, 3 and 5, respectively). It was found that Snort's performance detection throughout was affected by high-speed traffic, and there were more missed alerts and logs for packets as the speed of traffic increased. Also, when the malicious traffic was sent at high speed, it was found that Snort increased its missed malicious packet alerts and logs (see Figure 5), because Snort is capable of performing as a real-time traffic processor on the network. It is a multimode packet tool that can perform network traffic analysis, detection and content searching/matching in both real-time and for forensic post-processing [4, 5]. Although Snort has a limited time frame for processing packets and then alerting and logging any packets and malicious traffic successfully, because of the limitation of buffer size and processor speed (section 7 explains this more). If a network's traffic speed is higher than Snort's limit, Snort will miss alerts and logs.

To address this problem, A QoS configuration has been suggested in Layer-3 Cisco switches with parallel NIDPS nodes to increase packets throughput processing speed, even if traffic arrives at a high speed. Some mechanisms that QoS offers are queue, classification, policing and marking technologies, which can give a switch a new logical throughput-traffic-forwarding plan. A configuration of QoS offers two (2) input queues (ingress queues) and four (4) output queues (egress queues) at the physical switch interfaces (SVI or ports).As shown in Figure 6, the switch has been configured to two ingress queues and four egress queues to load a set of bytes (packets) into a number of processor input queues equally and to divide traffic (as a number of bytes) into parallel streams in order to speed up packets processing. It then uses parallel NIDPS to analyse each portion of traffic individually to determine whether it is free of malicious codes. A novel class map (marking) and a policy (policy map) were made for each input queue. The class map recognises and classifies a certain type of traffic for each input queue, and the policy map controls and recognises the speed limit for each input queue and applies it to interfaces [4]. The bandwidth, threshold, buffer, memory allocated and priority were configured for each ingress queue and each egress queue to treat and control traffic in order to help prevent congestion or disabled traffic in the input queues, even if traffic comes in at high load and speed.

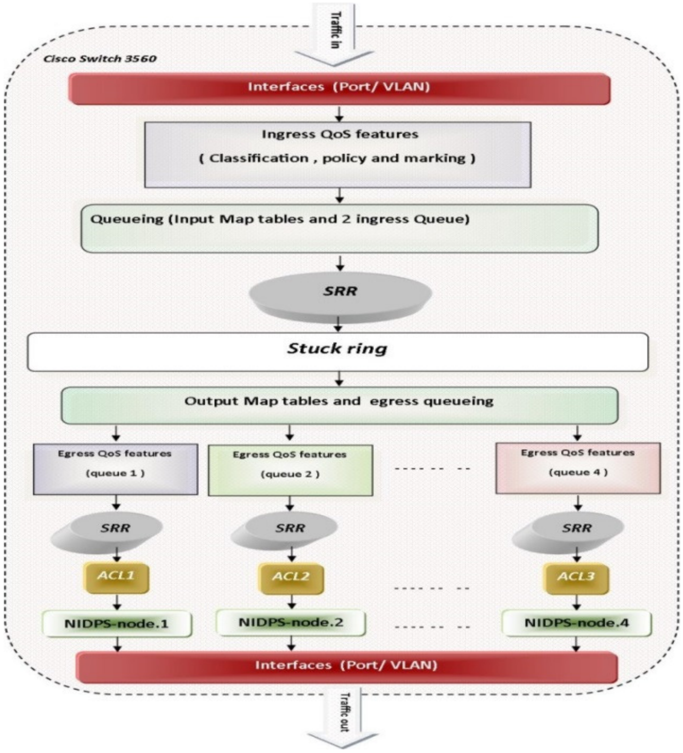

Traffic in
Cisco Switch 3560
Interfaces (Port/ VLAN)
Ingress QoS features
(Classification , policy and marking)
Queueing (Input Map tables and 2 ingress Queue)
SRR
Stuck ring
Output Map tables and egress queueing
Egress QoS features (queue 1)
Egress QoS features (queue 2)
Egress QoS features (queue 4)
SRR
SRR
SRR
ACL1
ACL2
ACL3
NIDPS-node.1
NIDPS-node.2
NIDPS-node.4
Interfaces (Port/ VLAN)
Traffic out

Figure 6. Parallel NIDPS node with QoS architecture.

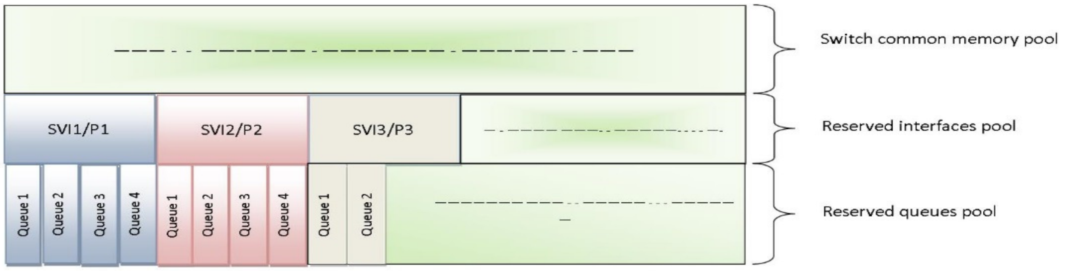


Figure 7: novel egress Queue buffer space reservation

QoS Shaped/Share Round Robin (SRR) technology was used to guarantee the bandwidth for an interface, and also for each ingress queue and egress queue. A network device must be able to identify packets in all the IP traffic flowing through it. The Shaped task only exists on output queues, and a queue books a percentage of a total port's bandwidth. One queue may be set as priority and queue do not share bandwidth. The Share function (SRR) is offered on both input and output queues. It provides a queue a portion of a total port's bandwidth, but unused bandwidth can be used by any queue. In our novel architecture (see Figure 6), the Share queue is used in input queues to help prevent congestion, and each output queue has an individual processing by using the Shaped queue to control speed traffic (bandwidth) A single NIDPS node is implemented for each output queue. One of the queues can be the expedited queue, which is serviced until empty before the other queues are serviced.

Also a memory buffer queue reservation was configured for each queue which provides more buffer space over its limit (over 100% of buffer space) when needed by reserving more space from available queue buffer, ports or SVI memory buffer or switch common memory pool buffer (see Figure 7). However, headers, such as ICMP, TCP and UDP, and even hackers have different characteristics, features and techniques. Using SRR, Threshold and Priority methods for each output queue can offer a wider range to deal with the behaviour of different IP headers and hackers.

The main aim of our novel architecture design is to manage and allocate a traffic load (number of bytes) into each input queue and process each output queue individually in order to permit a limited group of bytes that are divided into output queues to be processed at same time, thereby increasing NIDPS throughput processing time and reducing traffic congestion, even if the traffic is high load and speed.

6.2 Evaluation of Solution

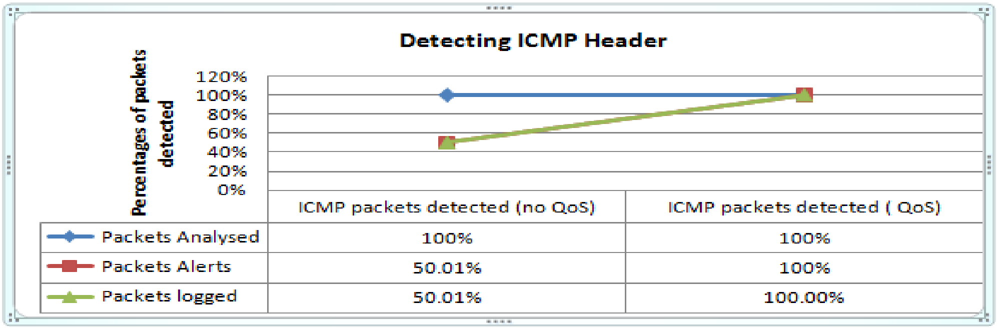
We ran (4) experiments to test performance of our novel NIDPS architecture design to detect: (1) ICMP header; (2) UDP header; (3) TCP header; and (4) Malicious packets. Each experiment tests Snort's detection rate without and with QoS and parallel technologies under high-speed traffic.

6.2.1 Experiment 5: Parallel Snort with QoS Reaction to Detect ICMP Header

In this experiment, more than 38,000 ICMP/IP packets were sent in high-speed traffic (1ms). Each packet carried 1KByte.

Table 5: Snort with QoS reaction to ICMP header in high speed traffic

Test type	The number of Packets received	Total packets analysed of the total the packets received	Total Eth packets received of the total packets analysed	ICMP packets analysed	TCP packets analysed	UDP packets analysed	ICMP packets alerts	ICMP packets logged	% packets alerts	% Packets logged
Without QoS	39121	37.259%	100%	14438	0	97	7220	7220	50.007%	50.007%
	Snort Processor Times = 155s -> (Pkts/min:7228 – Pkts/sec:94)									
With QoS	38668	99.943%	100%	37393	0	757	37393	37393	100.00%	100.00%
	Snort Processor Times = 293s -> (Pkts/min:9661 – Pkts/sec:131)									



	ICMP packets detected (no QoS)	ICMP packets detected (QoS)
Packets Analysed	100%	100%
Packets Alerts	50.01%	100%
Packets logged	50.01%	100.00%

Figure 8: Snort with QoS reaction to detect ICMP packets in 1ms.

As the results show in Figure 8 and Table 5, when more than 38,000 ICMP/IP packets were sent in an interval time of 1ms, packet processing speed was 94 packets per second (94 Pkts/sec) and number of alerts and logs was 7220 of the 14,438 ICMP packets that were analysed (see Table 5). When the same number of packets was sent at the same speed, using QoS and parallel technologies, packet processing speed was increased from 94 to 131 Pkts/sec and Snort detected all of the ICMP packets that it analysed (see Figure8). This experiment shows that when Snort NIDPS was used without QoS, it only detected 50% of the total packets analysed, but when Snort was used with QoS, Snort detected 100% of the total packets that it analysed (see Figure 8 and Table 5).

6.2.2 Experiment 6: Parallel Snort with QoS Reaction to Detect UDP Header

In this experiment, more than 38,000 UDP/IP packets were sent in high-speed traffic (0.5ms). Each packet carried 1KByte.

Table 6: Snort with QoS reaction to UDP header in high speed traffic.

Test type	The number of Packets received	Total packets analysed of the total the packets received	Total Eth packets received of total packets analysed	ICMP packets analysed	TCP packets analysed	UDP packets analysed	UDP packets Alerts	UDP packets logged	% packets alerts	% packets logged
Without QoS	36213	12.391%	100.00%	763	0	59	4	4	6.780%	6.780%
	Snort Processor Times = 59s -> (Pkts/sec:76)									
With QoS	37783	99.942%	100.00%	0	0	37564	37257	37257	99.182%	99.182%
	Snort Processor Times = 292 -> (Pkts/min:9440 – Pkts/sec:129)									

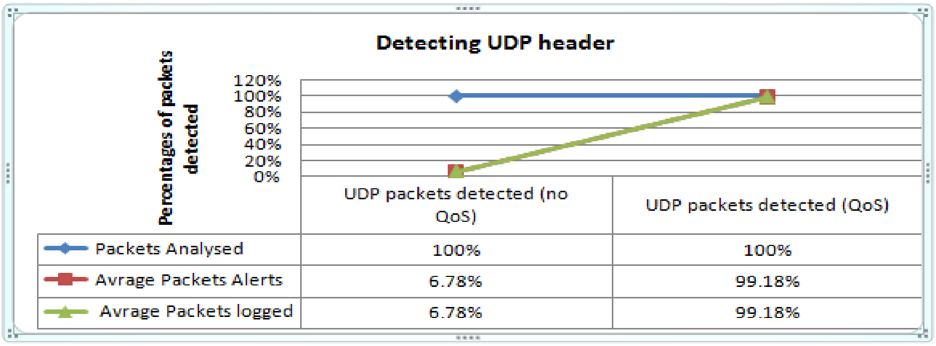


Figure 9: Snort with QoS reaction to detect UDP packets in 0.5ms

As the results show in Figure 6 and Table 6, when the traffic (packets) was sent at an interval time of 0.5ms, the number of packet alerts and logs was nearly 4 of the 59 UDP packets that were analysed with packet processing speed of 76 Pkts/sec (see Table 6). It detected fewer than 7% of all UDP packets that it analysed (see Figure 9). When QoS architecture was implemented and packets were sent again in the same traffic and interval speed of 1ms, the packet possessing speed was increased to 129 Pkts/sec and Snort detected more than 99% of the total UDP packets analysed (see Figure 9). This experiment shows that the Snort NIDPS performance detection improved from 7 to 99% when the novel QoS configuration was used.

6.2.3 Experiment 7: Parallel Snort with QoS Reaction to Detect TCP Header

In this experiment, more than 38,000 IP/TCP packets were sent in high-speed traffic (0.5ms); each packet carried 1KByte.

Table 7: Snort with QoS reaction to TCP Header in high speed traffic

Test type	The number of Packets received	Total packets analysed of the total the packets received	Total Eth packets received of total packets analysed	ICMP packets analysed	TCP packets analysed	UDP packets analysed	TCP packets Alerts	TCP packets logged	% Packets alerts	% Packets logged
Without QoS	32108	1.884%	100.00%	0	497	85	497	497	100.00%	100.00%
	Snort Processor Times = 34s -> (Pkts/sec:17)									
With QoS	31058	99.997%	100.00%	0	30057	779	30057	30057	100.00%	100.00%
	Snort Processor Times = 322s -> (Pkts/min:6211 – Pkts/sec:96)									

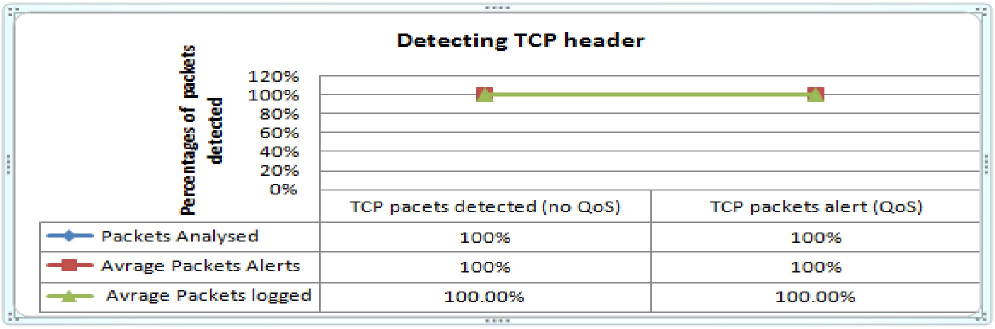


Figure 10: Snort with QoS reaction to detect TCP packets in 0.5ms.

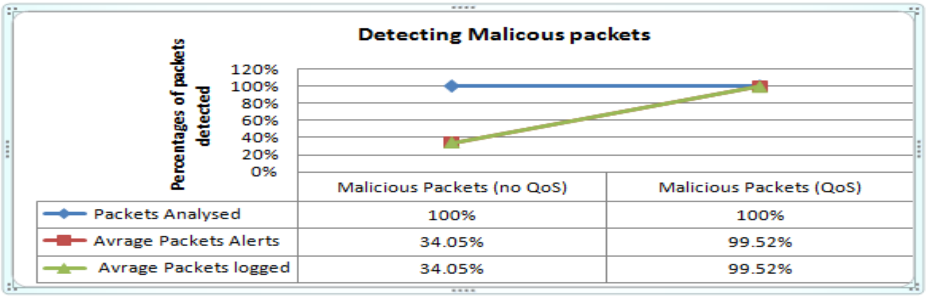
Figures 10 and Table 7 show that Snort detected 100% of the total TCP packets analysed, even without QoS. The packet processing speed is improved from 17 to 96 Pkts/sec with QoS.

6.2.4 Experiment 8: Parallel Snort with QoS Reaction to Detect Malicious Packets

In these experiments, a flood traffic was generated with UDP malicious packets (threads) in high-speed traffic (60000Bps, flooded traffic with 225 threads, sent at an interval time of 1mSec) by using NetScanPro, WinPcap and Flooder packets tools. Two tests were conducted: one test of Snort without QoS and one of Snort with QoS.

Table 8: Snort with QoS reaction to malicious packets in high speed traffic.

Test type	The number of Packets received	Total packets analysed of the total the packets received	Total Eth packets received of total packets analysed	ICMP packets analysed	TCP packets analysed	UDP packets analysed	UDP malicious packets Alerts	UDP malicious packets logs	% packets alerts	% packets logged
Without QoS	985686	1.004%	100.00%	0	0	793	270	270	34.048%	34.048%
	Snort Processor Times = 125s -> (Pkts/min:5090 – Pkts/sec:82)									
With QoS	996005	99.999%	100.00%	0	0	995155	990344	990344	99.517%	99.517%
	Snort Processor Times = 19.22m -> (Pkts/min:52421 – Pkts/sec:857)									



	Malicious Packets (no QoS)	Malicious Packets (QoS)
Packets Analysed	100%	100%
Avrage Packets Alerts	34.05%	99.52%
Avrage Packets logged	34.05%	99.52%

Figure 11: Snort with QoS reaction to detect malicious packets in high speed traffic.

As the results show in Figures 11 and Table 8, when malicious traffic was sent at high speed and volume, the number of malicious packets detected was 270 of the 793 malicious packets analysed and the packet processing speed was 82 Pkts/sec (see Table 8). Snort detected fewer than 35% of the total malicious packets analysed (see Figure 11). When the same traffic was generated with the same speed and value, but Snort was supported by the novel QoS architecture, the packet processing speed wan improved from 82 to 857 Pkts/sec and Snort detected more than 99% of the total malicious packets that it analysed (see Table 8). This experiment showed that the Snort NIDPS performance detection improved while novel QoS was used.

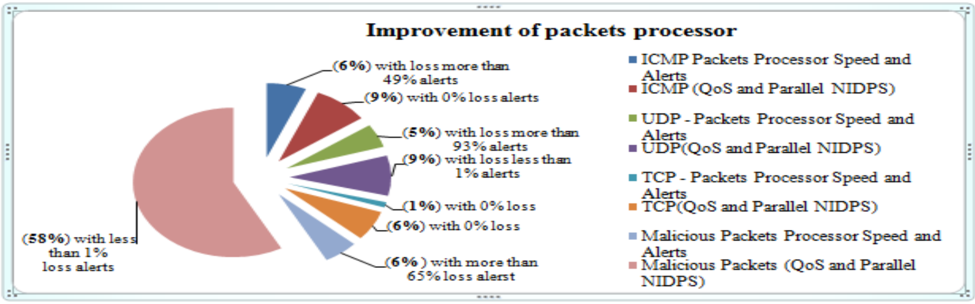


Figure 12: Improve NIDPS detection performance and packets processor through QoS and Parallel technologies.

Figure 12 provides a synopsis of our results. The experiments show that packets processor speed and Snort NIDPS detection performance have been improved when we used QoS, queue and parallel NIDPS technologies (see Figure 12).

7. TECHNICAL DISCUSSION

The performance of an NIDPS could be described as ineffective if the NIDPS is unable to detect or stop unwanted packets that could reach the system. There are two main causes of ineffective NIDPS: buffer size and processing speed.

When traffic moves through the network interface card (NIC) to the NIDPS node, the packets are stored on the buffer until the other relevant packets have completed transmission to processing nodes. In the event of high speed and heavy traffic in multiple directions, the buffer will fill up. Then packets may be dropped or left outstanding [20, 21, 22]. In this case, there is no security concern about the packets dropped; the packets are dropped outside the system. The outstanding packets that are waiting or have not been processed by the NIDPS node may affect the system.

However, packets can also be lost at the host. Most software tools use a computer program such as the kernel, which manages input/output (I/O) requests from software and decodes the requests into instructions to direct the CPU's data processing. When traffic moves from the interface (NIC) through the kernel's buffer to the processor space, where most of processing nodes are executed, the packets will be held in the kernel buffer before being processed by the CPU. When some nodes experience a high volume of data, the buffer will fill up and packets may be dropped. Configuring the kernel parameter in the New Application Programming Interface (NAPI) can enhance kernel performance by increasing the level of optimization and selecting multivariate features such as kernel complex quantitative near-infrared (K-NIR), kernel support vector regression (k-SVR), or kernel partial least squares (K-PLS) to improve the accuracy of packet processing [23, 24, 25, 26]. In order to hold and process packets quickly, these kernel performance enhancements pull a high value of packets from interfaces and bind them with obtainable CPU cycles, which limit packet speed and time and have no buffer memory. Furthermore, it requires a great deal of CPU to process a vast amount of data buffered in the kernel; the CPU cycles may run out of time. In this case, the packets that were dropped in the kernel and NIC might drop very early in the CPU cycles, which cannot buffer packets [27, 28]. In these two cases of host and processor packets loss, the NIDPS node is affected because packets are dropped before it is analysed. In this work, we are not focused on the network-based packets drop (NIC interfaces) as much as focusing on the host and processor based packets loss, because in the network-based, the packets are dropped from NIC and do not hit the system, but in the host based, the packets go through the system and then are dropped, which affects NIDPS detection performance. Our solution uses multi-layer switch technology, which supports QoS. The parallel NIDPS nodes are associated with queues each with a specific buffer and bandwidth thus increasing the queue buffer size automatically over its limit when needed. The NIDPS is faced with high speed and volume traffic. The appropriate number of NIDPS nodes is dependent in network speed limit. We therefore needed to operate with the class and quality of service technologies within the network switch.

By default, most of the Cisco switches work in Layer 2, the Data Link layer, and use the Class of Services (CoS) value [4, 29, 30] (see Figure 23). In this layer there is insufficient commands to support switch features such as QoS features, dynamic access control lists (ACLs), VLAN features, static IP routing, Routing Information Protocol (RIP), Policy-based routing (PBR) Cisco-default Smartports, etc. Other mechanisms operate at Layer 3 (see Figure 13). However, in the network switch operation, all the traffic should has equal priority and an equal chance of being delivered in a manner that is timely for the network. Some packets have an equal chance of being dropped when congestion network traffic occurs. DiffServ (Differentiated Service) allows different types of service to be offered depending on a code. For instance there can be a policy to give a certain type of package priority. QoS implementation is based on DiffServ architecture, which specifies that each packet be classified upon entry and has an equal priority and an equal chance of being delivered into network which is adjusted for different traffic speed in a good and timely manner. Furthermore, QoS makes network performance more predictable and bandwidth utilisation more effective [29, 30, 31, 32].

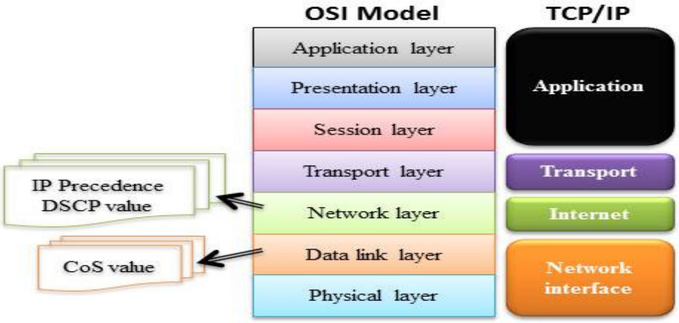


Figure 13. Place for CoS and DSCP values [4].

In our novel QoS architecture, the CoS values in Layer 2 were mapped to DSCP value (Differentiated Services Code Point) in Layer 3 so that appropriate categories could be matched at the network level. The static or dynamic classification methods involved Layer 3 header information matching, and a mechanism such as IP precedence or DSCP values was used to carry the IP packet header. For example, a dynamic classification access list can be used to identify IP traffic and place the traffic into a reserved queue [30, 31, 32]. Classification can also take place in the Layer 2 frame. Packet classification can be processor-intensive, so it should occur as far out toward the edge of the network as possible because every hop needs to make a determination on the treatment a packet should receive. A simpler classification is achieved through marking or setting the type of service (ToS) field in the IP header [29, 30, 31, 32]. In Layer 2, 802.1Q and 802.1p frames used 3 bits for IP type of service (ToS) field, and Layer 3 IPV4 packets used 6 bits for DSCP in (ToS) field to carry the classification (class) information (see Figure 14). The DSCP values allow for a higher degree of differentiation. CoS values range from 0 to 7 (8 values) and DSCP values range from 0 to 63 (64 values).

Regardless of the method by which the network is able to classify and identify IP traffic (either through port address information or through the ToS filed), those hops can then provide each IP packet with the required QoS. At that point, special techniques can be configured to provide priority queueing in order to ensure that large data packets do not interfere with packet data transmission. “If a node can set the IP Precedence or DSCP bits in the ToS field of the IP header as soon as it identifies traffic as being IP traffic, then all of the other nodes in the network can be classified based on these bits. However, in most IP networks, marking IP Precedence or DSCP should be sufficient to identify traffic as IP traffic” [4, 30, 31, 32].Differentiated services technology can be used such that each packet can be classified upon entry into the network and adjustments can be made for different traffic speeds and loads. In this work, the switch frame has been changed from Layer 2 to Layer 3. [4, 30, 31, 32].

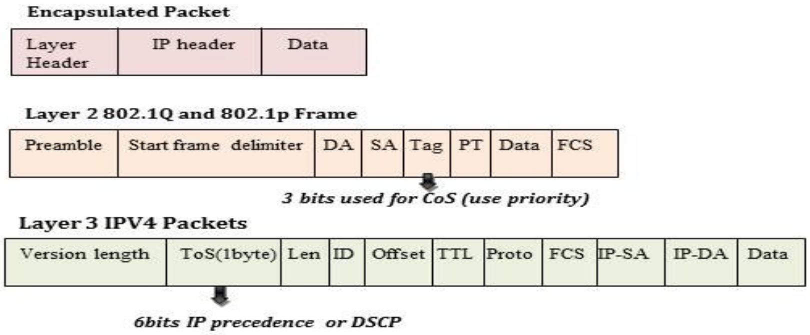


Figure 14. QoS classification Bits in Frames and Packets [30].

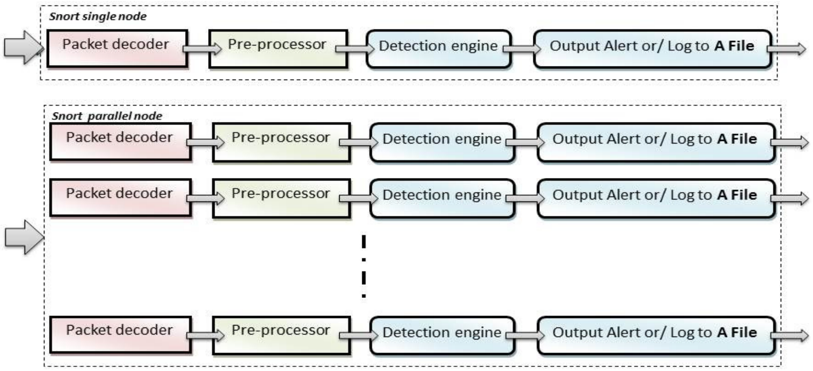


Figure 15: Snort NIDPS Parallel node.

Features of QoS, such as a policy map and class map, can be used to classify the traffic inside the switch with the same policy and class plan; different management can be given to packets with a different class and policy plan. Classification is process of identifying the data packets to a class or group in order to manage the packet appropriately [4, 30, 31, 32]. Network devices use several match criteria to place traffic into a certain number of classes. This can be processor-intensive if nodes must repeat classifications based on access list (ACL) matches. Therefore, nodes should mark packets as soon as they have identified and classified the IP traffic. Policing involves creating a policy that specifies the bandwidth limits for the traffic and applies it to the interface. Policing can be applied to a packet per direction and can occur on the ingress and egress interfaces [30]. Different types of traffic can be recognised in terms of, for instance, type and ports, and differentiated policies can be set.

In our work, Snort NIDPS is configured from single NIDPS detection node to multi NIDPS node (see Figure 15) and also configured CISCO QoS switch technology. Network QoS technology enables a new logical and throughput-traffic-forwarding plan to be implemented in the switch. A physical interface (port) was configured to two input queues (ingress queues) and four output queues (egress queues). A buffer was set for each queue in order to organise and hold more traffic by using a dynamic memory technology (buffer memory located). The buffer's memory space was divided between the switch common memory pool, the SVI, and the queue reserved pool (see Figure 7). We implemented a buffer allocation scheme to reserve a minimum amount for each egress buffer. Thus, all buffers cannot be consumed by one egress queue, and the system can control whether to grant buffer space to a requesting queue. Furthermore, a specific buffer memory space was defined for each queue, including ingress and egress thresholds. Packets were divided between two ingress and four egress queues via configured queue-sets. The remaining free common pool interfaces were set to reserve up to 50% of the available switch memory pool.

After all traffic has been placed into input queues and classes and policy based on their QoS requirements has been defined, appropriate services can be provided, for instance bandwidth guarantees, thresholds, memory buffering and priority servicing through an intelligent output queueing mechanism. The output queues were processed separately by implemented parallel NIDPS nodes in order to increase packet possessing speed. Other mechanisms that QoS offers is Shaped or Share Round Robin (SRR) technologies which can vary the bandwidth provided for the queues in the interface [17, 18, 19]. Shaped function (SRR) can guaranty each queue a bandwidth limit but queues cannot share bandwidth if queues reach their bandwidth limit. Share function (SRR) can guaranty a bandwidth limit for each queue and the other queues can share with each other if one of queues reached bandwidth limit. We utilized Share in the ingress queues but Shaped in the egress queues to ensure appropriate bandwidth for each egress queue.

Queue technology is placed at specific points in Cisco switches to help prevent congestion. The total inbound bandwidth of all interfaces may exceed a ring space of internal bandwidth [29, 30, 31]. After packets are processed through classification, policing, and marking, and before packets pass into the switch fabric, the system allocates them to input queues. Because multiple input queue interfaces can simultaneously send packets to output queue interfaces, outbound queues are allocated after the internal ring in order to avoid congestion. The SRR ingress queue sends packets to the internal ring, while the SRR egress queue sends the packets to the output queue. The novel configurable research architecture has a large limit of buffer space and a generous bandwidth allocation for each queue. Queue 1 was set as a priority queue for each ingress and egress queue, which allowed the system to prioritise packets with particular DSCP values and thereby allocate a large buffer. It also allows buffer space to be used more frequently, and then adjusts the thresholds for each queue and packets so that packets with lower priorities are dropped when queues are full. This allows the system to ensure that high priority traffic is not dropped.

Parallel NIDPS is a form of computation in which many NIDPS nodes work simultaneously, operating on the principle that the large incoming data can be divided into smaller sets, which are processed at the same time [4, 5]. Parallelism of NIDPS can occur at three general levels: the high-level processing node (entire system), the component level (part of the system with specific task), and the sub-component level parallelism (function within a specific task). In our novel QoS architecture, the entire Snort NIDPS was replicated on a number of machines, each of which processed a specific portion of the incoming traffic. Parallel queues (2 input queues and 4 output queues) were designed through QoS configuration on a switch virtual interface (SVI) where component level parallelism of NIDPS nodes were implemented with the aim of improving NIDPS throughput performance and reducing NIDPS processor time (see Figure 16).

The NIDPS node was configured from a single node NIDPS to a multi-node NIDPS. One group of rules were implemented for all nodes to perform a quick check for each packet with one group of rules to determine if the node contains the rule group associated with the given packet. Each node was configured to do a different type of task and therefore can be considered to be component level parallelism. Component level parallelism is defined as function parallelism of the NIDPS processing node. In component parallelism, individual components of NIDPS were isolated, and each output queue is given its own processing element. Furthermore, the component parallelism level could be created as a thread of a lightweight processing node and existing threads to schedule threads of NIDPS processing nodes.

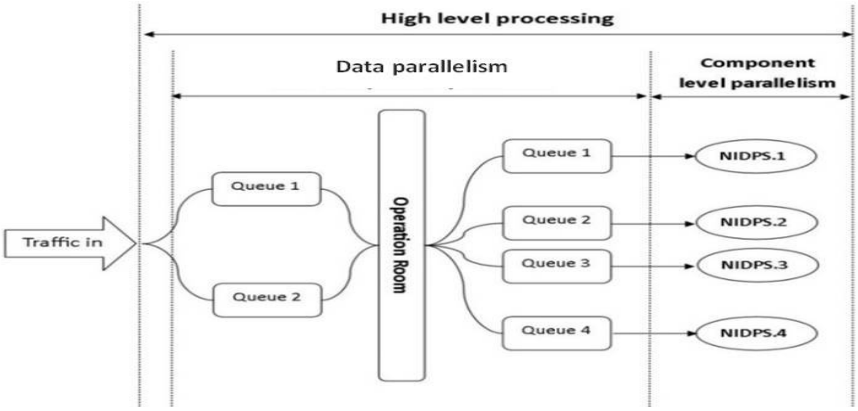


Figure 16: High level parallel processing.

8. CONCLUSIONS, RECOMMENDATIONS AND FURTHER WORK

8.1 Conclusion

NIDPS have become important components in securing today's computer networks. To be highly effective, an NIDPS must perform packet examination of incoming traffic at or near network speed. Failing to do so will permit malicious packets to infiltrate through the network undetected, and thus threaten network security. Our paper proposes and evaluates a new design of NIDPS architecture which uses a novel and unique infrastructure of QoS configuration and parallel technology in Layer 3 Cisco Catalyst switches to improve the NIDPS performance. The novel architecture reduces difficulties in maintaining security due to multiple characteristics of advanced computer networks, such as processing in real time, high speeds and high loads, which increase difficulties for defenders and reduce difficulties for attackers. Our experimental results show vast improvement in packet detection in such environments and therefore give better protection against attacks.

8.2 Recommendations and Further Work

Intruders usually use signatures that behave similarly to viruses used in computers. They also analyses data packets related to IP, which contains known anomalies, as either a single signature or a set of signatures. The detection system is capable of detecting suspicious activity in logs and generating alerts, based on these signatures and rules. NIDPSs are used to capture data and detect malicious packets that travel on the network media (cables, wireless) and match them to a database of signatures. Signature-based NIDPS are able to detect known attacks, but the major problem of the signature-based approach is that every signature should have an entry in a database in order to compare with the incoming packets. New signatures arise constantly and an issue is how to keep track up with new signatures. Another problem is processing time required to check all signatures. Knowledge sharing may provide a solution. Cloud computing which provides for massive processing distribution and sharing is a possible future direction [34, 35] but this also raises issues of trust. Our future work will investigate the use of specialized and trustworthy security clouds.

REFERENCES

[1] Nazer, G. M. & Selvakumar, A. A. L, (2011) “Current intrusion detection techniques in information technology—A detailed analysis”, European Journal of Scientific Research, 65, 4, pp 611-24.

[2] Radhakishan, V. & Selvakumar, S, (2011) “Prevention of man-in-the-middle attacks using ID based signatures”, In Proc. 2nd Int, Conf. Networking and Distributed Computing, (Sept. 2011). IEEE Press, pp165-169, DOI= http://dx.doi.org/10.1109/ICNDC.2011.40.

[3] Beg, S., Naru, U., Ashraf, M., & Moshin, S, (2010) “Feasibility of intrusion detection system with high performance computing: a survey”, Int. J. Advances in Computer Science, 1 (Dec. 2010), pp26-35.

[4] Bul'ajoul, W., James, A., & Pannu, M. 2015 “Improving network intrusion detection system performance through quality of service configuration and parallel technology”, Journal of Computer and System Sciences, 81, 6 (Sep. 2015), pp981–999, DOI= http://dx.doi.org/ 10.1016/j.jcss.2014.12.012.

[5] Bul’ajoul, W., James, A., & Pannu, M. (2013), Network intrusion detection systems in high-speed traffic in computer networks”, In e-Business Engineering (ICEBE), 2013 IEEE 10th International Conference (Sept. 2013), IEEE, pp168-175, DOI= http://dx.doi.org/10.1109/ICEBE.2013.26.

[6] Sangeetha, S., Ramya, R., Dharani, M. K., & Sathya, P (2015) “Signature based semantic intrusion detection system on Cloud”, In Information Systems Design and Intelligent Applications, Springer India, 339 (Jan. 2015), pp657-666, DOI= http://dx.doi.org/ 10.1007/978-81-322-2250-7_66.

[7] Rehman, R U (2003). Intrusion detection systems with Snort: advanced IDS techniques using Snort, Apache, MySQL, PHP, and ACID, Prentice Hall Professional.

[8] The Snort Documents (2014) [Online]. Available: https://www.snort.org/documents. [Accessed: 11-March-2015].

[9] Friedberg, I., Skopik, F., & Fiedler, R (2015) “Cyber situational awareness through network anomaly detection: state of the art and new approaches”, E & i Elektrotechnik und Informationstechnik, 132,2 (March. 2015), pp101-105, DOI= http://dx.doi.org/ 10.1007/s00502-015-0287-4.

[10] Tesfahun, A., & Bhaskari, D. L (2015) “Effective hybrid intrusion detection system: a layered approach”, International Journal of Computer Network and Information Security (IJCNIS), 7, 3 (Feb 2015), pp35, DOI= http://dx.doi.org/ 10.5815/ijcnis.2015.03.05.

[11] Jiang, W., Song, H., & Dai, Y (2005) “Real-time intrusion detection for high-speed networks”, Computers and Security, 24, 4 (Jun. 2005), pp287–294, DOI= http://dx.doi.org/ 10.1016/j.cose.2004.07.005.

[12] Kenkre, P. S., Pai, A., & Colaco, L (2015) “Real time intrusion detection and prevention system”, In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014 (Jan. 2015), Springer International Publishing, 405-411, DOI= http://dx.doi.org/ 10.1007/978-3-319-11933-5_44.

[13] Khamphakdee, N., Benjamas, N., & Saiyod, S (2015) “Improving Intrusion Detection System Based on Snort Rules for Network Probe Attacks Detection with Association Rules Technique of Data

Mining", Journal of ICT Research and Applications, 8.3 (Mar. 2015), pp234-250, DOI= 10.5614/itbj.ict.res.appl.2015.8.3.4.

[14] Khamphakdee, N., Benjamas, N., & Saiyod, S (2014) "Improving Intrusion Detection System based on Snort rules for network probe attack detection", In Information and Communication Technology (ICoICT), 2014 2nd International Conference on, IEEE, (May. 2014), pp69-74, DOI= http://doi.org/10.1109/ICoICT.2014.6914042.

[15] Saboor, A., Akhlaq, M., & Aslam, B. (2013) "Experimental evaluation of Snort against DDoS attacks under different hardware configurations", In Information Assurance (NCIA), 2013 2nd National Conference on, IEEE, (Dec. 2013), pp31-37, DOI= http://doi.org/10.1109/NCIA.2013.6725321.

[16] Buchanan, W. J., Flandrin, F., Macfarlane, R., & Graves, J (2011) "A methodology to evaluate rate-based intrusion prevention system against distributed denial-of-service (DDoS)", In: Cyberforensics, (Jun. 2011).

[17] Chen, M. J., Hsiao, Y. M., Su, H. K., & Chu, Y. S (2015), "High-throughput ASIC design for e-mail and web intrusion detection", IEICE Electronics Express, 12 (Jan. 2015), DOI=http://doi.org/10.1587/elex.12.20140854.

[18] Jiang, H., Zhang, G., Xie, G., Salamatian, K., & Mathy, L (2013) "Scalable high-performance parallel design for network intrusion detection systems on many-core processors", In Proceedings of the Ninth ACM/IEEE Symposium on Architectures for Networking and Communications Systems (Oct. 2013), IEEE Press, pp137-146, DOI= http://dx.doi.org/10.1109/ANCS.2013.6665196.

[19] Tilera, TILE-Gx8036 tm Processor specification brief [Online]. Available: http://www.meganovo.cn/uploadfile/web/product/tilera/processor/TILE-Gx8036_PB033-02_web.pdf. [Accessed: 20-May-2015].

[20] Naouri, Y., & Perlman, R (2015) "Network congestion management by packet circulation", Washington, DC: U.S, Patent and Trademark Office , 8,989,017(Mar. 2015).

[21] Kishore, K. R., Hendel, A., & Kalkunte, M. V (2015) "System, Method and Apparatus for Network Congestion Management and Network Resource Isolation", Washington, DC: U.S. Patent and Trademark Office, 20,150,146,527 (May. 2015).

[22] Zhu, Y., et al (2015) "Packet-Level Telemetry in Large Datacenter Networks", In Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication, ACM, 15(Aug. 2015), pp479-491 ,DOI= http://doi.org/10.1145/2785956.2787483.

[23] Wu, H., Cadambi, S., & Chakradhar, S. T (2015) "Optimizing data warehousing applications for GPUs using dynamic stream scheduling and dispatch of fused and split kernels", Washington, DC: U.S. Patent and Trademark Office, 8,990,827 (Mar. 2015).

[24] Lutz, T., Fensch, C., & Cole, M (2015) "Helium: a transparent inter-kernel optimizer for OpenCL", In Proceedings of the 8th Workshop on General Purpose Processing using GPUs , ACM, (Feb. 2015), pp70-80, DOI= http://doi.org/10.1145/2716282.2716284.

[25] Fraser, N. J., Moss, D. J., Lee, J., Tridgell, S., Jin, C. T., & Leong, P. H (2015) "A Fully Pipelined Kernel Normalised Least Mean Squares Processor For Accelerated Parameter Optimisation", In Proc. International Conference on Field Programmable Logic and Applications (FPL), page to appear.

[26] Lee, J., Chang, K., Jun, C. H., Cho, R. K., Chung, H., & Lee, H. (2015) "Kernel-based calibration methods combined with multivariate feature selection to improve accuracy of near-infrared spectroscopic analysis", Chemometrics and Intelligent Laboratory Systems.

[27] Emmerich, P., et al (2015) "Optimizing Latency and CPU Load in Packet Processing Systems", In International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS).

[28] Smith, S. C., Hammell, R. J., Parker, T. W., & Marvel, L. M (2014) "A theoretical exploration of the impact of packet loss on network intrusion detection, In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2014 15th IEEE/ACIS International Conference on. IEEE, (Jun/Jul. 2014), pp1-6, DOI=http://doi.org/10.1109/SNPD.2014.6888699.

[29] Cisco (2015) Catalyst 3560 Switch Software Configuration Guide, Cisco IOS Release 15.0(2)SE and Later,[Online], Available: http://www.cisco.com/c/en/us/td/docs/switches/lan/catalyst3560/software/release/15-0_2_se/configuration/guide/scg3560.html. [Accessed: 11-April-2015].

[30] Szigeti, T., Hattingh, C., Barton, R., & "Briley, K (2013) "End-To-End QoS Network Design: Quality of Service for Rich-Media and Cloud Networks", Pearson Education.

[31] Wallace, K (2011) Implementing Cisco Unified Communications Voice Over IP and QoS (CVOICE) Foundation Learning Guide: (CCNP Voice CVoice 642-437), Cisco Press.

[32] Americas Headquarters, C (2010) Catalyst 3560 Switch Software Configuration Guide, Cisco IOS Release 12.2(55)E.

[33] Parker, D. B (1981) Computer Security Management, Reston Publishing Company, Reston, VA.

[34] Agrawal, D., Das, S., & El Abbadi, A (2011) "Big data and cloud computing: current state and future opportunities", In Proceedings of the 14th International Conference on Extending Database Technology (March. 2011), ACM, pp530-533, DOI= http://dx.doi.org/ 10.1145/1951365.1951432.

[35] Patel, A., Taghavi, M., Bakhtiyari, K., & Júnior, J. C (2013) "An intrusion detection and prevention system in cloud computing: a systematic review", Journal of Network and Computer Applications, 36, 1 (Jan. 2013), pp25-41,DOI= http://dx.doi.org/ 10.1016/j.jnca.2012.08.007.

AUTHORS

Waleed Bul'ajoul, a doctoral research and lecturer assistant,at Coventry University, UK, with background in networking and security especially in NIDPS, Network configuration, Quality of services and parallel technologies.

Anne James is Professor of Data Systems Architecture at Coventry University, UK. She obtained her BSc degree at Aston University, UK in 1980 and her PhD at the University of Wolverhampton UK in 1986. The research interests of Professor James are in the general area of creating distributed systems to meet new and unusual data and information challenges.

Dr Siraj Shaikh, is a Reader in Cyber Security at the Centre for Mobility and Transport at Coventry University, UK .He has been involved in research, development and evaluation of large-scale distributed secure systems for over fifteen years. His main research interest lies in systems security, essentially at the intersection of cyber security, systems engineering and traditional computer science. He has addressed a range of problem domains including monitoring of insider and stealthy attacks, automotive cybersecurity, rail safety, and software assurance.

Dr Mandeep Pannu is an academic at Kwantlen Polytechnic University Canada, She obtained her MSc and PhD degrees at Coventry University, UK. Her research interests are in information retrieval and computer security

AUTHOR INDEX