Volume 5 ▪ Issue 3 ▪ August 2011

INTERNATIONAL JOURNAL OF

COMPUTER SCIENCE AND SECURITY (IJCSS)

ISSN : 1985-1553

Publication Frequency: 6 Issues / Year

CSC PUBLISHERS

http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND SECURITY (IJCSS)

VOLUME 5, ISSUE 3, 2011

EDITED BY
DR. NABEEL TAHIR

ISSN (Online): 1985-1553

International Journal of Computer Science and Security is published both in traditional paper form and in Internet. This journal is published at the website http://www.cscjournals.org, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJCSS Journal is a part of CSC Publishers
Computer Science Journals
http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND SECURITY (IJCSS)

Book: Volume 5, Issue 3, August 2011
Publishing Date: July / August 2011
ISSN (Online): 1985 -1553

IJCSS Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers, 2011

EDITORIAL PREFACE

This is third issue of volume five of the International Journal of Computer Science and Security (IJCSS). IJCSS is an International refereed journal for publication of current research in computer science and computer security technologies. IJCSS publishes research papers dealing primarily with the technological aspects of computer science in general and computer security in particular. Publications of IJCSS are beneficial for researchers, academics, scholars, advanced students, practitioners, and those seeking an update on current experience, state of the art research theories and future prospects in relation to computer science in general but specific to computer security studies. Some important topics cover by IJCSS are databases, electronic commerce, multimedia, bioinformatics, signal processing, image processing, access control, computer security, cryptography, communications and data security, etc.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 5, 2011, IJCSS appears in more focused issues. Besides normal publications, IJCSS intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

This journal publishes new dissertations and state of the art research to target its readership that not only includes researchers, industrialists and scientist but also advanced students and practitioners. The aim of IJCSS is to publish research which is not only technically proficient, but contains innovation or information for our international readers. In order to position IJCSS as one of the top International journal in computer science and security, a group of highly valuable and senior International scholars are serving its Editorial Board who ensures that each issue must publish qualitative research articles from International research communities relevant to Computer science and security fields.

IJCSS editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCSS. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJCSS provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Computer Science and Security (IJCSS)

EDITORIAL BOARD

TABLE OF CONTENTS

Volume 5, Issue 3, August 2011

Pages

298 - 309 Survey Paper: Cryptography Is The Science Of Information Security
Mohammed AbuTaha, Mousa Farajallah, Radwan Tahboub, Mohammad Odeh

310 - 315 Automatic Detection of Malaria Parasites for Estimating Parasitemia
S. S. Savkare, S. P. Narote

316 - 337 Information Security Maturity Model
Malik F. Saleh

338 - 345 Toward a New Algorithm for Hands Free Browsing
Murad Al-Rajab, Haifaa Kattan

346 - 355 A Distributed Approach to Defend Web Service from DDoS Attacks
Monika Sachdeva, Gurvinder Singh, Kuldip Singh

356 - 367 Development of an Efficient Computing Multilingualism Model for Diacritical Marks in Arabic and Hindi
Abu Sarwar Zamani, Nasser Al Arifi, Md. Mobin Akhtar

368 - 375 Black Box Backup System
iyad Ahmad Aldasouqi, Arafat Awajan

376 - 386 Face Recognition Using Neural Network Based Fourier Gabor Filters & Random Projection
Anissa Bouzalmat, Naouar Belghini, Arsalane Zarghili, Jamal Kharroubi, Aicha Majda

387 - 393 Finding Relationships between the Our-NIR Cluster Results
N.Sudhakar Reddy

Survey Paper: Cryptography Is The Science Of Information Security

Mohammed AbuTaha
College of Administrative Sciences and Informatics *m_abutaha@ppu.edu*
Palestine Polytechnic University
Hebron, Palestine

Mousa Farajallah
College of Engineering and Technology *mousa_math@ppu.edu*
Palestine Polytechnic University
Hebron, Palestine

Radwan Tahboub
College of Engineering and Technology *radwant@ppu.edu*
Palestine Polytechnic University
Hebron, Palestine

Mohammad Odeh
IT and Communications Dept *mhmdodeh@qou.edu*
Al-Quds Open University
Hebron, Palestine

Abstract

Cryptography in the past was used in keeping military information, diplomatic correspondence secure and in protecting the national security. However, the use was limited. Nowadays, the range of cryptography applications have been expanded a lot in the modern area after the development of communication means; cryptography is essentially required to ensure that data are protected against penetrations and to prevent espionage. Also, cryptography is a powerful mean in securing e-commerce. Cryptography is used to ensure that the contents of a message are confidentiality transmitted and would not be altered. Confidentiality means nobody can understand the received message except the one who has the decipher key, and data cannot be changed means the original information would not be changed or modified; this is done when the sender includes a cryptographic operation called a hash function in the original message. A hash function is a mathematical representation of the information, when any information arrives at its receiver; the receiver calculates the value of this hash function. If the receiver's hash function value is equivalent to the sender's, the integrity of the message is assured .

Keyword: Symmetric Encryption, A Symmetric Encryption ,Hash Algorithm, Caesar Table.

1. INTRODUCTION

Nowadays, cryptography plays a major role in protecting the information of technology applications. Information security is an important issue, for some applications. Have the top priority such as e-commerce, e-banking, e-mail, medical databases, and so many more, all of them require the exchange of private information. For example, let us consider a person named Alice a sender who wants to send a data message which has a length of m characters to a receiver called Bob. Alice uses an unsecure communication channel. Which could be a telephone line , computer network, or any other channel. If the message contains secret data, they could be intercepted and read by hackers. Also they may change or modify the message during its transmission in such a way that Bob would not be able to discover the change. In this survey a various ways of encryption is viewed and have been compared ,a lot of examples have been provided .

1.1 Cryptography Goals

By using cryptography many goals can be achieved, These goals can be either all achieved at the same time in one application, or only one of them, These goals are:

1. Confidentiality: it is the most important goal, that ensures that nobody can understand the received message except the one who has the decipher key.

2. Authentication: it is the process of proving the identity, that assures the communicating entity is the one that it claimed to be, This means that the user or the system can prove their own identities to other parties who don't have personal knowledge of their identities. (The primary form of host to host authentication on the Internet today is name-based or address-based; and both of them are notoriously weak).

3. Data Integrity: its ensures that the received message has not been altered in any way from its original form, This can be achieved by using hashing at both sides the sender and the recipient in order to create a unique message digest and compare it with the one that received.
4. Non-Repudiation: it is mechanism used to prove that the sender really sent this message, ,and the message was received by the specified party, so the recipient cannot claim that the message was not sent [2].

5. Access Control: it is the process of preventing an unauthorized use of resources. This goal controls who can have access to the resources, If one can access, under which restrictions and conditions the access can be occurred, and what is the permission level of a given access.

1.2 Basic Terminology of Cryptography

Computers are used by millions of people for many purposes. such as banking, shopping, military, student records, etc.... . Privacy is a critical issue in many of these applications, how are we need to make sure that an unauthorized parties cannot read or modify messages.

Cryptography is the transformation of readable and understandable data into a form which cannot be understood in order to secure data. cryptography refers exactly to the methodology of concealing the content of messages, the word cryptography comes from the Greek word "Kryptos", that means hidden, and "graphikos" which means writing [3].

The information that we need to hide, is called **plaintext** (P), It's the original text, It could be in a form of characters, numerical data, executable programs, pictures, or any other kind of information, The plaintext for example is the first draft of a message in the sender before encryption, or it is the text at the receiver after decryption.

The data that will be transmitted is called **cipher text** (C), it's a term refers to the string of "*meaningless"* data, or unclear text that nobody must understand, except the recipients. it is the data that will be transmitted Exactly through network, Many algorithms are used to transform plaintext into cipher text [4].

Cipher is the algorithm that is used to transform plaintext to cipher text, This method is called encryption or enciphers (encode), in other words, it's a mechanism of converting readable and understandable data into "*meaningless*" data, and it is represented as follows:

$$C = E_{(K)}(P) \quad (1)$$

Where $E_{(K)}$ is the encryption algorithm using key k.

The opposite of cipher mechanism is called **decipher (decode)** that is the algorithm which recovers the cipher text, this method is called decryption, in other words it's the mechanism of converting "*meaningless"* data into readable data.

$$P = D_{(K^{-1})}(C) \quad (2)$$

The Key is an input to the encryption algorithm, and this value must be independent of the plaintext, This input is used to transform the plaintext into cipher text, so different keys will yield different cipher text, In the decipher side, the inverse of the key will be used inside the algorithm instead of the key.

Computer security it's a generic term for a collection of tools designed to protect any data from hackers, theft, corruption, or natural disaster while allowing these data to be available to the users at the same time. One example of these tools is the A-vast antivirus program [1].

Network security refers to any activity designed to protect the usability, integrity, reliability, and safety of data during their transmission on a network, Network security deals with hardware and software, The activity can be one of the following anti-virus and anti-spyware, firewall, Intrusion prevention systems, and Virtual Private Networks [4].

Internet Security is measures and procedures used to protect data during their transmission over a collection of interconnected networks .while **information security** is about how to prevent attacks, and to detect attacks on information-based systems [2].

Cryptanalysis (code breaking) is the study of principles and methods of deciphering cipher text without knowing the key, typically this includes finding and guessing the secrete key, It's a complex process involving statistical analysis, analytical reasoning, math tools and pattern-finding, The field of both cryptography and cryptanalysis is called **cryptology** [4,15].
Symmetric encryption refers to the process of converting plaintext into cipher text at the sender with the same key that will be used to retrieve plaintext from cipher text at the recipient. while **asymmetric encryption** refers to the process of converting plaintext into cipher text at the sender with different key that will be used to retrieve plaintext from cipher text at the recipient [15].

Passive attacks mean that the attackers or the unauthorized parties just monitoring on the traffic or on the communication between the sender and the recipient, but not attempting to breach or shut down a service, This kind of attacks is very hard to discover, since the unauthorized party doesn't leave any traces. On the other hand **active attacks** mean that the attackers are actively attempting to cause harm to the network or the data. The attackers are not just monitoring on the traffic, but they also attempt to breach or shut down the service [4,15].

Authentication is the process of determining whether someone is the same person who really is, such as login and password in login pages while authorization is the process of ensuring that this person has the ability to do something [4, 9, 15].

Brute force is the attacker who is trying all of the possible keys that may be used in either decrypt or encrypt information [15].

1.3 A Brief History of Cryptography

The encryption process is as old as writing itself, Through this short historical combo, the most important stations in the progress of data encryption will be reviewed. It is believed that the first texts used or contained any encryption techniques were known 4000 years ago at the Veterans Egyptian where the hieroglyphic inscriptions on the tomb of the nobleman Khnumhotep II, They were written with a number of unusual symbols to confuse or obscure the meaning of the inscriptions [6].

2000 years ago, the Greek knew cylinder device called Scytale, which was the sender's part very similar to the recipient part, where a narrow strip of parchment or leather, was wounded around the Scytale and the message was written across it, so if anyone tries to read the text he will find meaningless letters, The only one that can read this text is the one who has the Scytale, This technique is similar to the transposition technique which will be later discussed in symmetric encryption section [5].
The Arab role in the data encryption, was since ancient times, Through the analysis of the text of the holy Qur'an text, Muslim scholars were able to invent frequency analysis technique for breaking monoalphabetic substitution ciphers about 1200 years ago, by Sheikh AL-Kindi in his famous book "Risalah fi Istikhraj al-Mu'amma (Manuscript for the Deciphering Cryptographic Messages)", which it was

the most advanced in cryptography since that time, until the World ware two, Figure 1 shows the first page of AL-Kindi's book, After AL-Kindi's invention, all cipher text became vulnerable to this cryptanalytic technique, until the development of the polyalphabetic cipher by Leone Battista Alberti, who is known as "The Father of Western Cryptology" in 1465 [6].

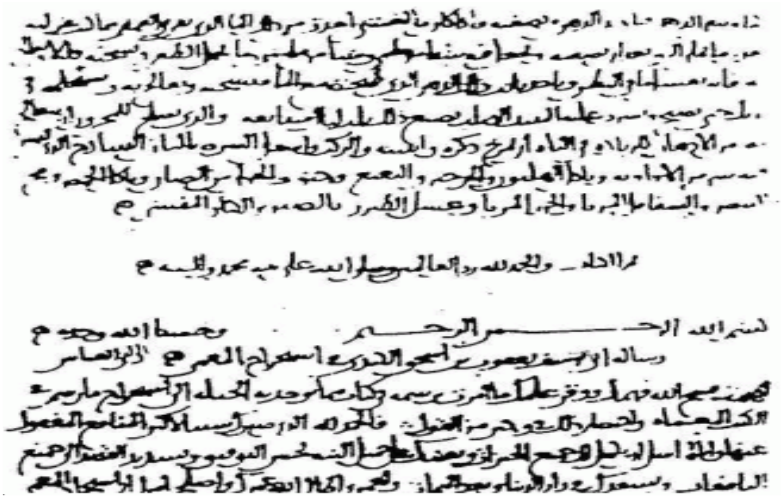


FIGURE.1: The first page of al-Kindi's manuscript On Deciphering Cryptographic Messages

The next step was in 1518 by Trithemius, a German monk, who wrote a table of Twenty-six column and Twenty-six row. Each row duplicate the above row but shifted by one letter.

In 1585, Blaise de Vigenere developed a Trithemius table by changing the way that the keywords system works. One of his used techniques is the plaintext as its own key.

Forty-three years later, a Frenchman named Antoine Rossignol helped his army to defeat the Huguenots, by deciphering a captured message. After that victory, Antoine was deciphering messages for the benefit of the French government many times. He used two lists to solve his ciphers: "one in which the plain elements were in alphabetical order and the code elements randomized, and one to facilitate decoding in which the code elements stood in alphabetical or numerical order while their plain equivalents were disarranged" [7].

The wheel cipher is a cylinder composed of Twenty six cylindrical piece of wood, The alphabetical letters inscribed randomly on each piece the [8].
The development in data encryption has begun to accelerate after the discovery of the telegraph, simply sending messages by the telegraph is not secure; therefore they had to provide means of data encryption before transmission.

In 1854, Charles Wheatstone and Lyon Playfair invented the Playfair system, which was consisted from 5X5 rectangle key, while the plaintext message divided into adjacent pairs, This system will be discussed later.

Before 1883, the encrypt ion process often depended on hiding of algorithm to protect data, Of course, This is not practical, but the first major advances in cryptography were made in the year 1883 by Kerkhoff by developing a set of principles which is now known as Kerkhoff principle, The major principle is, hidding the key of algorithm instead of hiding the algorithm itself [5].

Kerkhoff Principles [9]

1. Ciphertext should be unbreakable.
2. The cryptosystem should be convenient for the correspondent.
3. The key should be easily remembered and changeable.

4. The Ciphertext should be transmitted by the telegraph.
5. The cipher apparatus should be easily portable,
6. The cipher machine should be relatively easy to use.

In 1915, two Dutch navy officers invented the rotor machine; which is a combination of electrical and mechanical systems. The simple view of rotor machine is an electrical system with twenty-six switches pressed by the plaintext, These switches attached by a wire to a random contact letter on the output, for example if the plaintext letter is pressed, the wiring is placed inside a rotor, and then rotated with a gear every time a letter was pressed. So while pressing ***A*** the first time might generate character ***D***, the next time it might generate character ***S*** [10].

In 1918, the german army during the world war one, used ADFGVX cipher system, which consisted from a table, the first row and first column was the key while the data entry was randomly replaced by the plaintext with pair of characters of text at the top of the corresponding row and corresponding column, The following figure shows the replaced character T with pair AD Figure 2 explain ADFGVC cipher system [11].

	A	D	F	G	X
A	B	T	A	L	P
D	D	H	O	Z	K
F	Q	F	V	S	N
G	G	J	C	U	X
X	M	R	E	W	Y

FIGURE 2 : Example of Using ADFGVX cipher system.

Lester Hill is one of the few scientists who had concluded that mathematics inevitably necessary for the success of encryption, and the encryption remained the same until 1941 when Adrian Albert Benefited from Hill theorem and built an encryption system based on mathematics [12].

In 1948, Shannon published "A Communications Theory of Secrecy Systems", In this paper Shannon's analysis demonstrates several important features of the statistical nature of language that make nearly the solution of all previous ciphers very straight forward, One of the most important result in this paper is that Shannon developed a measure for cryptographic strength called the "unicity distance" [14].

During a collaboration between Whitfield Diffie and Martin Hellman in 1976 , the Diffie-Hellman key agreement was invented, The method was based on the selected three variables at the sender ***(x, a, P)*** and generating of ***s***, then sending ***(s, a, P)*** to the recipient, the recipient chooses ***y*** and uses ***y*** with ***(a, P)*** to generate ***r*** and sends ***r*** to the sender, the sender use ***r*** with ***(x, P)*** to generate the public key, The recipient also uses ***s*** with ***(y, P)*** to generate the same public key, Figure 3 explains this idea [15].

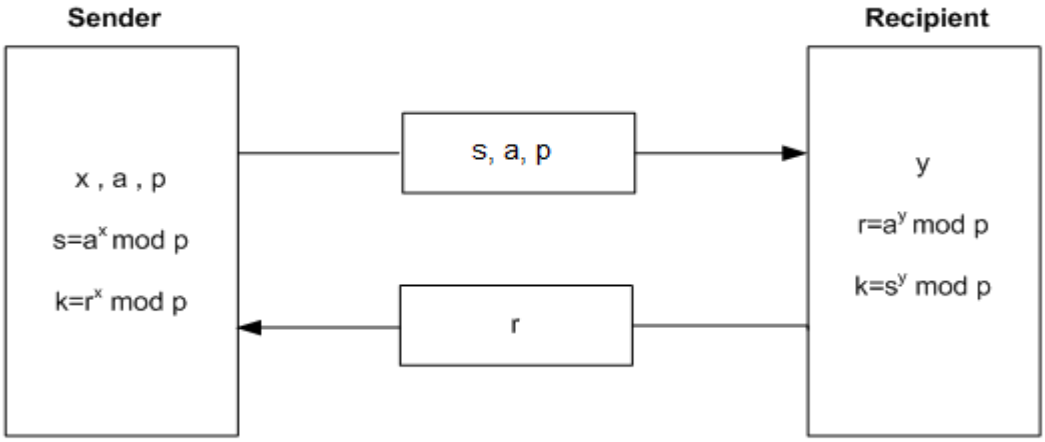


FIGURE .3 : Diffie-Hellman key generation

After Diffie-Hellman approach, the cryptography was divided into symmetric and asymmetric cryptography, and then many techniques and methods were developed. The next section is about the symmetric and asymmetric encryption [16].

2. SYMMETRIC AND ASYMMETRIC ENCRYPTION

Encryption is the strongest and the safest way in securing data. Certainly, it is the most common one. Encryption systems are divided into two major types or forms, symmetric and asymmetric.

Symmetric encryption is known as secret key or single key, The receiver uses the same key which the sender uses to encrypt the data to decrypt the message,. This system was the only system used before discovering and developing the public key., A safe way of data transfer must be used to moving the secret key between the sender and the receiver in symmetric encryption. Figure 4 shows how the system works. Symmetric encryption occurs either by substitution transposition technique, or by a mixture of both. Substitution maps each plaintext element into cipher text element, but transposition transposes the positions of plaintext elements.

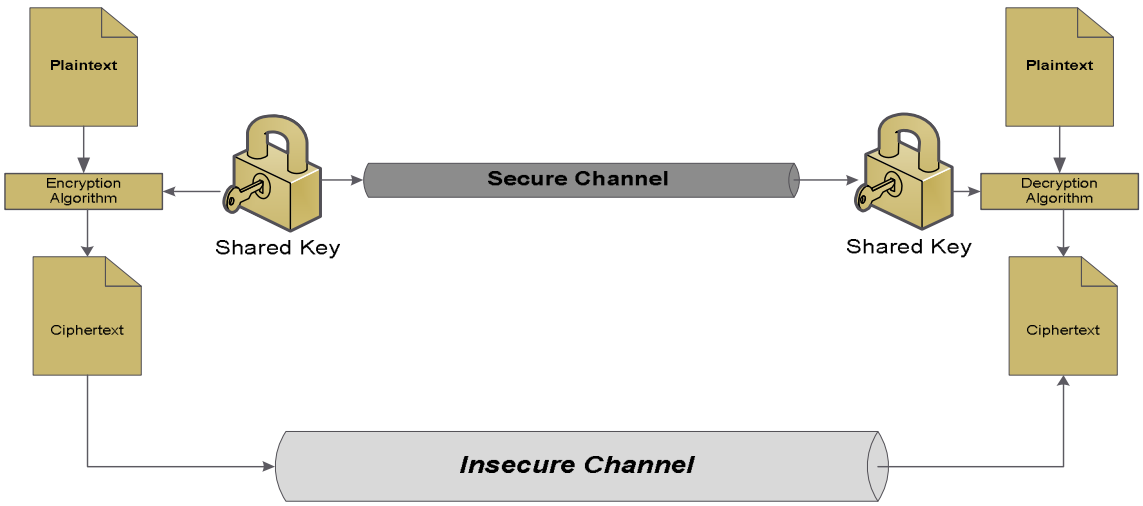


FIGURE .4 : Simplified model of conventional encryption

Plaintext	Encryption Process	Cipher text
p➔15	(15+12) mod 26	1➔ b
a➔ 0	(0+12) mod 26	12➔m
l ➔11	(11+12) mod 26	23➔x
e➔ 4	(4+12) mod 26	16➔q
s➔18	(18+12) mod 26	4➔ e
t➔19	(19+12) mod 26	5➔ f
i➔ 8	(8+12) mod 26	20➔u
n➔13	(13+12) mod 26	25➔z
e➔ 4	(4+12) mod 26	16➔q

The common simplified cipher algorithm which assigns each character of plaintext into numerical value is called Caesar cipher, , its sums the key value to the numerical value of plaintext character, and then assigns the rest of the division by modular value into cipher text character, where the modular value is the max numerical value plus one [17], The mathematical model of Caesar cipher is:

At encryption side: $$E_n(x) = (x + n) \bmod p \quad (3)$$

At decryption side: $$E_n(x) = (x - n) \bmod p \quad (4)$$

Where x is the plaintext character and x is shift value, the following example illustrates Caesar cipher model:

Example 1:
Let the plaintext message is "Palestine" and the key value=12 , and use the simplest symmetric encryption algorithm ,which called "Caesar cipher", the Caesar table will be:

Table .1: Caesar Table

a	b	C	d	e	f	g	h	i	j	k	L	m	n	o	p	q	r	s	T	u	v	W	x	y	z
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

The cipher text which arrive to the receiver is “bmxqefuzq“, and the cipher text is entered into decryption process in the receiver to decrypt the text as follow:

Cipher text	Decryption Process	Plaintext
b➔ 1	(1 - 12) mod 26	15➔ p
m➔12	(12 - 12) mod 26	0➔ a
x➔ 23	(23 - 12) mod 26	11➔ l
q➔ 16	(16 - 12) mod 26	4➔ e
e➔ 4	(4 - 12) mod 26	18➔ s
f➔ 5	(5 - 12) mod 26	19➔ t
u➔ 20	(20 - 12) mod 26	8➔ i
z➔ 25	(25 - 12) mod 26	13➔ n
q➔ 16	(16 - 12) mod 26	4➔ e

an advanced rail fence technique which is more sophisticated technique on symmetric encryption , uses the original plaintext to write it in row-by-row, and read the cipher text column-by-column, but at decryption side write the cipher text column-by-column and retrieve the plaintext by reading the message row-by-row, the mathematical model of advanced rail fence when $(key = d_1 d_2 d_3 \cdots d_n)$, where $(d_3 > d_1 > d_n > d_2)$:

$$\begin{array}{lccccc} key & d_1 & d_2 & d_3 & \cdots & d_n \\ & p_1 & p_2 & p_3 & \cdots & p_n \end{array} \quad (5)$$

$$\begin{array}{cccccc} key & d_1 & d_2 & d_3 & \cdots & d_n \\ & C_{2\times l/n+1} & C_1 & C_{3\times l/n+1} & \cdots & C_{l/n+1} \\ & C_{2\times l/n+2} & C_2 & C_{3\times l/n+2} & \cdots & C_{l/n+2} \\ & \vdots & \vdots & \vdots & \vdots & \vdots \\ & C_{3\times l/n} & C_{l/n} & C_{4\times l/n} & \cdots & C_{2\times l/n} \end{array} \quad (6)$$

Where d_1 is the smallest digit among digits of key that consist from n digits, l represent number of characters in plaintext message, p_i is the i^{th} character of plaintext message and C_i is the i^{th} character of cipher text output.

Example 2:

To understand and accommodate advance rail fence technique, let us consider $(key = 5236417)$, plaintext (p)"AES is a block cipher intended to replace DES for commercial application":

Using equation (5), the encryption message:

Key	5	2	3	6	4	1	7
Plaintext:	A	e	s	i	s	a	b
	L	o	c	k	c	i	p
	H	e	r	i	n	t	e
	N	d	e	d	t	o	r
	E	p	l	a	c	e	d
	E	s	f	o	r	c	O
	M	m	e	r	c	i	A
	L	a	p	p	l	i	C
	A	t	i	o	n	x	X

Output: Aitoeciixeoedpsmatscrelfepiscntcrclnalhneemlaikidaorpobperdoacx

Using equation (6), the decryption message (plaintext):

Key	5	2	3	6	4	1	7
Plaintext:	A	e	s	i	s	a	B
	i	o	c	k	c	i	P
	h	e	r	i	n	t	E
	n	d	e	d	t	o	R
	e	p	l	a	c	e	D
	e	s	f	o	r	c	O
	m	m	e	r	c	i	A
	l	a	p	p	l	i	C
	a	t	i	o	n	x	X

Output: Aesisablockcipherintendedtoreplacedesforcommercialapplication

From previous examples, the plaintext is translated into different cipher text and then transferred throw unsecured channel to the receiver, while the secrete key which is been used in encryption process will be transferred throw secured channel, At the receiver side the inverse of the secret key or/and the inverse of encryption process are used to decrypt the cipher text and to retrieve the original plaintext, Caesar mechanism is the core for all encryption model, from easy to very complicated one, in other word, the encryption process needs key to convert the plaintext into cipher text, but at the receiver the inverse of processes will retrieve the original plaintext.

Symmetric encryption has many advantages over asymmetric. Firstly, it is faster since it doesn't consume much time in data encryption and decryption. Secondly, it is easier than asymmetric encryption in secret key generation. However, it has some disadvantages, for example key distribution and sharing of the secret key between the sender and the receiver, also symmetric key encryption incompleteness, since some application like authentication can't be fully implemented by only using symmetric encryption [18].

In 1976 Diffie and Helman invented new encryption technique called public key encryption or asymmetric encryption; Asymmetric encryption is the opposite of symmetric encryption in safety, since it doesn't require sharing the secret key between the sender and the receiver. And this is the main difference between symmetric and asymmetric encryption, the sender has the public key of the receiver. Because the receiver has his own secret key which is extremely difficult or impossible to know through the public key, no shared key is needed; the receiver is responsible for establishing his private and public key, and the receiver sends the public key to all senders by any channel he needs, even unsecured channels to send his public key, asymmetric key can use either the public or secret key to encrypt the data. Also it can use any of the keys in decryption, asymmetric encryption can be used to implement the authentication and non-repudiation security services, and also it can be used for digital signature and other application that never be implemented using symmetric encryption. Figure.5 shows how the system works.

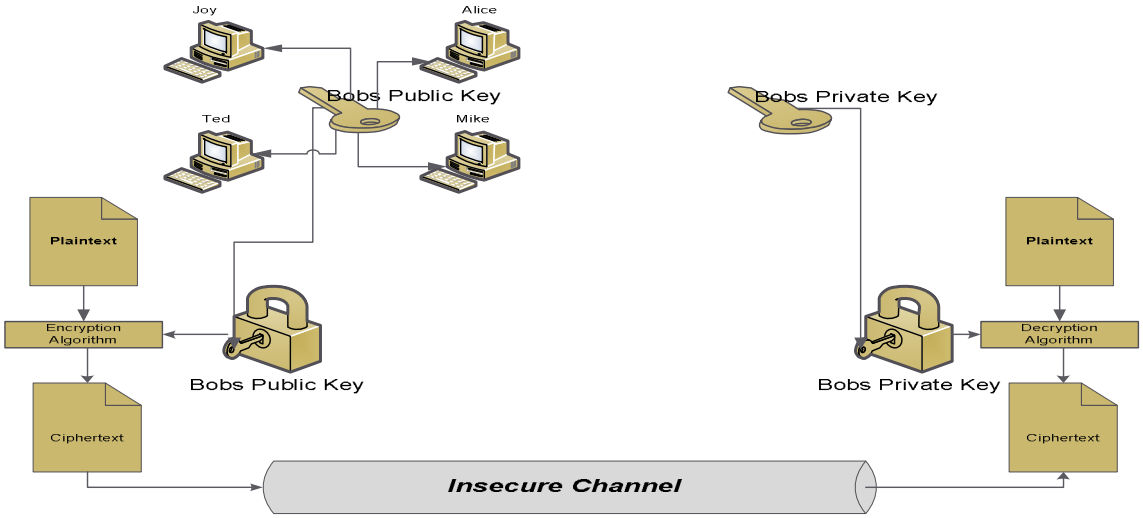


FIGURE 5 : Simplified model of asymmetric encryption

Asymmetric encryption is slower and very complicated in calculations than symmetric encryption . Therefore, asymmetric encryption deals with plaintext as a group of numbers which are manipulated in mathematics, while the plaintext in symmetric encryption deal as group of symbols and characters, the encryption process may permute these symbols, or may substitute one symbol by another.

So the nature of the data determines the system of encryption type. And every system has its own uses. For example, asymmetric encryption may be used in authentication or in sending secret key for decryption.

To understand asymmetric encryption, lets us take RSA model which is an example on asymmetric encryption, RSA model main steps:

RSA Model Steps:

- Each user generates a public/private key pair by selecting two large primes at random p, q.
- Computing modular value $n = p \times q$
- Calculating the Euler's function $\phi(n) = (p-1) \times (q-1)$
- Selecting at randomly the public encryption key e, where $1 < e < \phi(n)$, and e is prime relative to the $\phi(n)$.

- Solving the following equation to find private decryption key d, $e \times d = 1 \; mod \; \phi(n)$, and $0 \leq d \leq n$.
- Publishing their public encryption key: $P_K = (e, n)$.
- Keeping secret private decryption key: $P_R = (d, n)$.
- At the encryption side the sender uses encryption mathematical equation $C = P^e \; mod \; n$.
- At the decryption side the receiver uses decryption mathematical equation $P = C^d \; mod \; n$.

Example 3:

Let a part of the plaintext message be "Palestine", then the RSA key generation process is:

- Select two prime numbers: p=23 & q=17
- Computing $n = p \times q = 23 \; x \; 17 = 391$
- Computing $\phi(n) = (p-1) \times (q-1) = 22 \; x \; 16 = 352$
- Selecting $e: gcd(e, 352) = 1; \; choose \; e = 7$
- Determining $d: d \times e = 1 \; mod \; 352 \; and \; d < 352$ Value is $d = 151$ since $151 x 7 = 1057 = 352 \times 3 + 1$
- Publishing public key $P_K = (7, 391)$.
- Keeping private key secrete $P_R = (151, 391)$.

The encryption process and decryption process then is applied to previously calculated parameters as follow:

Plaintext			Encryption Process
p	➔	15	$15^7 \; mod \; 391 = 195$
a	➔	00	$00^7 \; mod \; 391 = 000$
l	➔	11	$11^7 \; mod \; 391 = 122$
e	➔	04	$04^7 \; mod \; 391 = 353$
s	➔	18	$18^7 \; mod \; 391 = 052$
t	➔	19	$19^7 \; mod \; 391 = 383$
i	➔	08	$08^7 \; mod \; 391 = 219$
n	➔	03	$13^7 \; mod \; 391 = 055$
e	➔	04	$04^7 \; mod \; 391 = 353$

The cipher text will arrive the receiver, and at the receiver the cipher text will be entered into decryption process to decrypt the text as follow:

Decryption Process	Plaintext		
$195^{151} \; mod \; 391 = 015$	015	➔	p
$000^{151} \; mod \; 391 = 000$	000	➔	a
$122^{151} \; mod \; 391 = 011$	011	➔	l
$353^{151} \; mod \; 391 = 004$	004	➔	e
$052^{151} \; mod \; 391 = 018$	018	➔	s
$383^{151} \; mod \; 391 = 019$	019	➔	t
$219^{151} \; mod \; 391 = 008$	008	➔	i
$055^{151} \; mod \; 391 = 003$	003	➔	n
$353^{151} \; mod \; 391 = 004$	004	➔	e

The mathematical model for symmetric and asymmetric encryption consists of key, encryption and decryption algorithm and powerful secured channel for transmitting the secrete key or any channel for transmitting the public key from the sender to the receiver, the mathematical model similar to equations (1 - 2):

At encryption side: $\boldsymbol{C = E_K(P)}$

At decryption side: $\boldsymbol{P = D_K(C)}$

Where C is the cipher text to be sent, E is the encryption algorithm, P is the plaintext, D is the decryption algorithm, and K is the key used inside the encryption and/or decryption process.

3. RESULTS AND COMPARISON

When it comes to encryption, the latest isn't necessarily the best. You should always use the encryption algorithm that is right for the job and has been extensively publicly analyzed and tested, something the cryptographic community won't have had the chance to do with a brand new algorithm. Let's have a look at some of the most widely-used algorithms. For most people, encryption means taking plaintext and converting it to cipher text using the same key, or secret, to encrypt and decrypt the text. This is symmetric encryption and it is comparatively fast compared to other types of encryption such as asymmetric encryption. The most widely-used algorithm used in symmetric key cryptography is AES (Advanced Encryption Standard). It comprises three block ciphers, AES-128, AES-192 and AES-256, each of which is deemed sufficient to protect government classified information up to the SECRET level with TOP SECRET information requiring either 192 or 256 key lengths.

The main disadvantage of symmetric key cryptography is that all parties involved have to exchange the key used to encrypt the data before they can decrypt it. This requirement to securely distribute and manage large numbers of keys means most cryptographic services also make use of other types of encryption algorithms. Secure MIME (S/MIME) for example uses an asymmetric algorithm - public/private key algorithm - for non-repudiation and a symmetric algorithm for efficient privacy and data protection.

Asymmetric algorithms use two interdependent keys, one to encrypt the data, and the other to decrypt it. This interdependency provides a number of different features, the most important probably being digital signatures which are used amongst other things to guarantee that a message was created by a particular entity or authenticate remote systems or users. The RSA (Rivest, Shamir and Adleman) asymmetric algorithm is widely used in electronic commerce protocols such as SSL, and is believed to be secure given sufficiently long keys and the use of up-to-date implementations. As RSA is much slower than symmetric encryption, what typically happens is that data is encrypted with a symmetric algorithm and then the comparatively short symmetric key is encrypted using RSA. This allows the key necessary to decrypt the data to be securely sent to other parties along with the symmetrically-encrypted data.

4. SUMMARY

Cryptography is used to ensure that the contents of a message are confidentiality transmitted and would not be altered. Confidentiality means nobody can understand the received message except the one that has the decipher key, and "data cannot be changed" means the original information would not be changed or modified; this is done when the sender includes a cryptographic operation called a hash function in the original message. A hash function is a mathematical representation of the information, when information arrives at its receiver; the receiver calculates the value of this hash function. If the receiver's hash function value is equivalent to the sender's, the integrity of the message is assured [15].in this survey paper we describe and compare between symmetric and asymmetric encryption technique ,provide many example to show the differences .

5. REFERENCES

[1] J. Badeau.,: " The Genius of Arab Civilization ", Second Edition. MIT Press,(1983), USA.

[2] M .Chapple., M Solomon,: " Information Security Illuminated " First Edition. Jones and Bartlett Publishers, (2005), USA.

[3] J.R Childs: " General Solution of the ADFGVX Cipher System ". Aegean Park Press, ,(2000), USA.

[4] D.Delfs., and K. Helmut.,: " Introduction To Cryptography: Principles and applications ", Second Edition. Springer Science & Business Media, (2007), Germany.

[5] G .Dieter: " Computer Security ", Second Edition. John Wiley & Sons, , (2005), UK.

[6] A. Forouzan.,: " Cryptography and Network Security ", First Edition. McGraw-Hill, (2007), USA.

[7] R Hamamreh., M Farajallah., " Design of a Robust Cryptosystem Algorithm for Non-Invertible Matrices Based on Hill Cipher ". International Journal of Computer Science and Network Security, (2009): Vol (9), pp: 12-21.

[8] J Hoffstein., et al, " An Introduction to Mathematical Cryptography ", First Edition. Springer Science & Business Media, (2008):, Germany.

[9] H.Kenneth, " Elementary Number Theory and Its Applications " Third Edition. Addison-Wesley, (1992): Germany.

[10] M.Lucas, " Thomas Jefferson wheel cipher ", Monticello Research Department, Thomas Jefferson Foundation, Charlottesville, (1995):, VA.

[11] S .Maret," Cryptography Basics PKI ", First Edition. Dimension Data SA, ., (1999):, Switzerland.

[12] E.Ralph., F Weierud" Naval Enigma: M4 and Its Rotors ". Crypologia, ,. (1987):, Vol(11),pp:235-244.

[13] W .Reinhard., " Cryptology Unlocked ", Translation Edition By Angelika Shafir. John Wiley & Sons, (2007):, UK.

[14] H. Rodríguez, et al,: " Cryptographic Algorithms on Reconfigurable Hardware ", First Edition. Springer, (2006), USA.

[15] D.Salomon" Data Privacy and Security " First Edition. Springer-Verlag New York, ., (2003):, Inc. USA.

[16] C .Shannon,. " Communication Theory of Secrecy Systems ". Bell Syst, (1949):, Tech. J., Vol (28), pp: 656-715.

[17] W .Stallings, " Cryptography and network security, Principles and practices ", Fourth Edition. Pearson Prentice Hall, (2006):, USA.

[16] K .Thomas, : " The Myth Of The Skytale ". Taylor & Francis, (1998), Vol (33), pp: 244-260.

Automatic Detection of Malaria Parasites for Estimating Parasitemia

S. S. Savkare *swati_savkare@yahoo.com*
Moze College of Engineering,
University of Pune,
Pune, India

S. P. Narote *snarote@rediffmail.com*
Sinhgad College of Engineering,
University of Pune,
Pune, India

Abstract

Malaria parasitemia is a measurement of the amount of Malaria parasites in the patient's blood and an indicator for the degree of infection. In this paper an automatic technique is proposed for Malaria parasites detection from blood images by extracting red blood cells (RBCs) from blood image and classifying as normal or parasite infected. Manual counting of parasitemia is tedious and time consuming and need experts. Proposed automatic approach is used Otsu thresholding on gray image and green channel of the blood image for cell segmentation, watershed transform is used for separation of touching cells, color and statistical features are extracted from segmented cells and SVM binary classifier is used for classification of normal and parasite infected cells.

Keywords: OTSU Thrsholding, Watershed Transform, Feature Extraction, SVM Classifier.

1. INTRODUCTION

Malaria is a serious disease caused by a blood parasite named Plasmodium spp. It affects at least 200 to 300 million people every year and causes an estimated 3 million deaths per annum. Diagnosis and medication of it is necessary [1], [2]. In blood sample visual detection and recognition of Plasmodium spp is possible and efficient via a chemical process called (Giemsa) staining [4]. The staining process slightly colorizes the RBCs but highlights Plasmodium spp parasites, white blood cells (WBC), and artifacts. Giemsa stains nuclei, chromatin in blue tone and RBCs in pink color. It has been shown in several field studies that manual microscopy is not a reliable screening method when performed by non-experts. Malaria parasites host in RBCs when it enter in blood stream. In Malaria parasitemia count it is important step to segment RBCs from blood image and classify it as parasite infected or normal. In thin blood images morphology of cells can be observed clearly. The present paper describes the techniques used in segmenting normal and infected RBCs for purpose of Malaria parasitemia (number of infected blood cells over total red blood cell) count.

This paper is organized as follows: Section 2 summarizes literature related to segmentation of cells and count Malaria parasitemia. Section 3 illustrates the system architecture which includes pre-processing, cell segmentation, RBCs segmentation, feature extraction and classification. Section 4 and 5 include results and conclusion of this paper.

2. RELATED WORK

Minh-Tam Le et. al. [3], proposed a comparison-based analysis, which differentiates solid components in blood smears. The semiautomatic method uses statistical measures and cross-referencing validations yields a reliable detection scheme. The nucleated components are identified using adaptable spectral information. Cells and parasites are isolated from the background, by comparing the input image with an image of an empty field of view. The range of erythrocyte sizes is determined by input of isolated RBC.

Jesus Angulo et. al. [4], presents a technique to automatically detect the working area of peripheral blood smears stained with Giemsa. The approach consists of two stages. First, an image analysis procedure using mathematical morphology is applied for extracting the erythrocytes, the centers of erythrocytes and the erythrocytes with center. Second, the number of connected components from the three kinds of particles is counted.

D. Ruberto et. al. [5] follow morphological method for detection of parasites in Giemsa stained blood slides. Different objects in blood are identified using their dimensions and color. The parasites are detected by means of an automatic thresholding based on morphological approach, using Granulometrices to evaluate size of RBCs and nuclei of parasite. A segmentation method using morphological operators combined with the watershed algorithm.

Silvia et. al. [6], proposed a technique for estimating parasitemia. Template matching is used for detection of RBCs. Parasites are detected using variance-based technique from grayscale images and second approach is based on color co-occurrence matrix. Support Vector Machine (SVM) as the classifier which exploits the texture, geometry and statistical features of the image.

Stanislaw Osowski et. al. [7], presents the application of a genetic algorithm (GA) and a support vector machine (SVM) to the recognition of blood cells on the image of the bone marrow aspirate. GA is used for the selection of the features for the recognition of the neighboring blood cells belonging to the same development line. The SVM is used for final recognition and classification of cells.

3. SYSTEM ARCHITECTUR

System architecture used for Malaria parasite detection involves following steps: Image Acquisition, Pre-processing, cell segmentation, Feature Extraction, and Classification. Block diagram of system architecture is shown in Figure1.

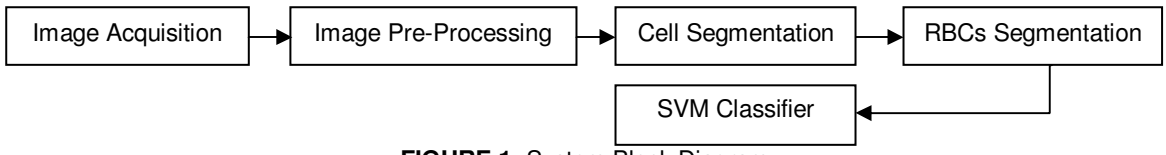


FIGURE 1: System Block Diagram.

3.1 Image Acquisition

For slide preparation working solutions of Giemsa were made by adding 100 µl stock solution to each milliliter of distilled water. Dried thin blood films were fixed with methanol for 30 s, poured off and stained with Giemsa for 20 min [4], [8]. The stain was rinsed off with tap water for 10 s. Upon drying, slides were used immediately or stored for future use. Image was captured by connecting high resolution Digital camera to microscope. By adjusting microscope magnification image is captured.

3.2 Image Pre-Processing

Pre-processing step includes noise reduction, smoothening of image. In this paper we used median filter for smoothening of color image and Lapalcian filter is used for edge sharpening. This result is subtracted from original to enhance the image. The median filter [8] is a non-linear digital filtering technique, used to remove noise from images. In median filtering pixel replaces with the median of its neighboring pixel values. Lapalcian filter takes second order derivative of pixel. After pre-processing image is send to cell segmentation block to segment cells.

3.3 Cell Segmentation

To segment foreground from background Global threshold and Otsu threshold [10] is used on grayscale enhanced image. For low contrast image segmentation applied on enhanced green channel of the image. Result of thresholding on both images is added to get binary image of cells. A 3 x 3 median filter was applied on this binary cell mask to fill the holes in blood cells and to remove the unwanted points from binary image of cells and background [11]. Using morphological operation cells having larger area is identified which is overlapping of the cells.

Distance transform is applied on it followed by watershed transform [5]. This gives separation of overlapping cells. This final binary image of cells is given to next block.

3.4 RBCs Segmentation

First rule check for White blood cells which are bigger than the RBCs, and second check for platelets which are smaller than RBCs. Using morphological operation platelets are removed from binary image. By labeling this binary image total number of cells is calculated.

3.5 Feature Extraction

Since the chosen features affect the classifier performance, selection of feature which is to be used in a specific data classification problem is as important as the classifier itself [12]. The features which give predominant difference between normal and infected cells are identified and used for training purpose. The selected features are geometrical, color and statistical based. The mathematical morphology provides an approach to the processing of image based on shape. The set of parameters corresponds to the geometrical features are as follows:

Radius -measured by averaging the length of the radial line. *Perimeter* - the total distance between consecutive points of the border, *Area* - the number of pixels on the interior of the cell. *Compactness* - is the ratio of perimeter^2 by area, *Metric* – (Perimeter)^2/4π·Area which is 1 for circle.

The values of saturation histogram is used for classification it is spread for infected cell and lye towards left if normal cell. Histogram of green plane of normal cell is spread and for infected cell it lies towards right [7].

$$Skewness = \frac{1}{\sigma^2}\sum_{b=0}^{L-1}(b-\bar{b})^3$$

$$Kurtosis = \frac{1}{\sigma\ b^4}\sum_{b=0}^{L-1}(b-\bar{b})^4 P(b) - 3$$

$$Energy = \sum_{b=0}^{L-1}[p(b)]^2$$

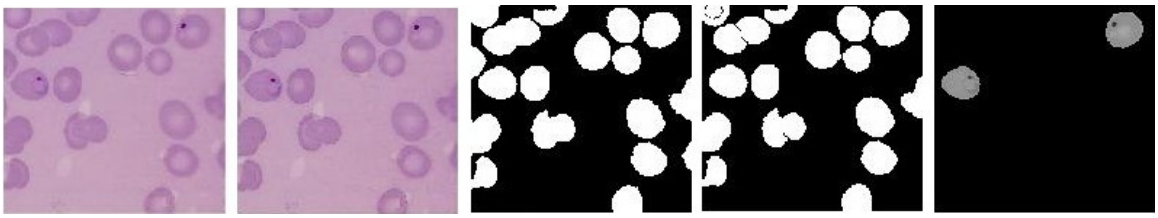
$$StndardDev\quad iation = \left[\sum_{b=0}^{l-1}\left(b-\bar{b}\right)\right]^{1/2}$$

P(b) is the first-order histogram estimate, Parameter *b* is the pixel amplitude value. *L* is the upper limit of the quantized amplitude level. The above parameters are used for feature extraction. The statistical features use gray level histogram and saturation histogram of the pixels in the image and based on such analysis, the mean value; angular second momentum, Skewness, Standard deviation, Kurtosis are treated as the features [14] and calculated using above equations.

3.6 SVM Classifier

The SVM is a powerful solution to the classification problems. In this paper, it has been used for the recognition and classification of cells. The main advantage of the SVM network used as a classifier is its very good generalization ability and extremely powerful learning procedure, leading to the global minimum of the defined error function. Linear SVM is a linear discriminant classifier working on the principle of maximum margin between two classes. The decision function of the N-dimensional input vector x for K-dimensional feature space (K>N) is defined as $D(x) = w^T \varphi(x) + b$ through the use of function $\varphi(x)$. Where $\varphi(x) = [\varphi 1(x), \varphi 2(x), \ldots, \varphi K(x)]$, w as the weight vector of network $w=[w_1, w_2, \ldots, w_k]^T$, and b as the bias weight [12]. All values of weights have been arranged in decreasing order and only the most important have been selected for each pair of classes and then used in the final classification system.

The learning of the SVM network working in the classification mode is aimed at the maximization of the separation margin between two classes. Simple classification algorithm is proposed that classifies points by assigning them to the closer of two parallel planes (in input or feature space). Standard support vector machines (SVMs), which assign points to one of two half spaces. SVM classifier is used for classification of normal and infected cells. Results pre-processing, Otsu's threshold to get binary image of cells, separation of overlapping cells and finally detection of infected cells is shown in Figure 2.



(a) (b) (c) (d) (e)

FIGURE 2: a) Original Image, b) Pre-processed Image, c) Binary Image of Cells, d) Separation of Overlapping cells, e) Detected parasite infected cells

4. RESULT

The described methods of feature extraction produce a very rich group of parameters. Skewness of healthy cells is up to 2 and for infected cell it is above 2. Kurtosis of normal cell is below 3 and for infected it is up to 9. Standard deviation of infected cell is very high as compare to normal cell. Thus all extracted features are sends to next block for classification. The binary classifier using RBF kernel is used for classification.

Image	Manual Parasitemia	Automatic Parasitemia
1	25.00	25.00
2	13.33	6.67
3	11.11	11.11
4	12.50	12.50
5	6.67	7.14
6	16.67	25.00
7	3.03	3.03
8	4.76	4.76
9	18.18	18.18
10	2.78	2.78
11	0.00	0.00
12	4.00	4.00
13	20.00	20.00
14	10.00	18.18
15	2.94	2.94

TABLE 1: Summary of Manual and Automatic Parasitemia.

The cost parameter C and Lagrange multiplier λ are taken 1000 and 10^{-7} respectively. Image processed through automatic system segments RBCs from input image, separate overlapping cells, counts total number of erythrocytes and SVM binary classifier detect infected cells. Finally system gives number of normal cells and infected cell, and percentage parasitemia in command window. 15 images processed through automatic system. Table 1 summarizes result of manual and automatic parasitemia for 15 images. Figure 3 shows graphical comparison of manual and automatic parasitemia count.

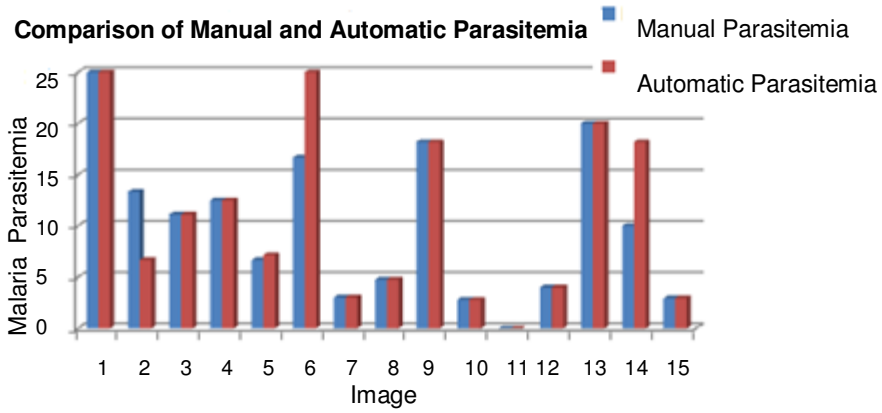


FIGURE 3: Graphical comparision of Manual and Automatic Parasitemia.

5. CONCLUSION

The proposed automated method of segmentation and classification of cell is simple. An approach is proposed to detect red blood cells with consecutive classification into parasite infected and normal cells for estimation of parasitemia. The extraction of red blood cells achieves a reliable performance and the actual classification of infected cells. Sensitivity of system is 93.12%, and Specificity is 93.17%.

Shape based and statistical features are generated for classification. The features are selected for recognition of two classes only. This approach leads to the high specialization of each classifier and results in an overall increase in accuracy. The above algorithms are implemented using MATLAB.

6. REFERENCES

[1] Shiff, C., 2002. Integrated approach for malaria control. Clin. Microbiol. Rev.15, 278–293.

[2] World Health Organization What is malaria? Factssheetno94. http://www.who.int/mediacentrefactsheetsfs094/en/./factshee. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp 68–73.

[3] M. Tam Le, T. Bretschneider,"*A novel semi-automatic image processing approach to determine Plasmodium falciparum parasitemia in Giemsa-stained thin blood smears*", Research article, BMC Cell Biology, 28 March 2008.

[4] J. Angulo, G. Flandrin, "*Automated detection of working area of peripheral blood smears using mathematical morphology*", U. S. National Library of Medicine, Analytical Cellular Pathology 25(1), pp 39-47, 2003.

[5] C.D. Ruberto, A.G. Dempster, S. Khan, and B. Jarra, "*Automatic Thresholding of Infected Blood Images Using Granulometry and Regional Extrema*", in Proceedings of International Conference on Pattern Recognition, pp 3445-3448, 2000.

[6] S. Halim et al., "*Estimating Malaria Parasitaemia from Blood Smear Images*", in Proceedings of IEEE international conference on control, automation, robotics and vision, pp 1-6, 2006.

[7] S. Osowski et al., "*Application of Support Vector Machine and Genetic Algorithm for Improved Blood Cell Recognition*", in proceedings of IEEE transaction on Instrumentation and Measurement, Vol. 58, No. 7, pp 2159-2168, July 2009.

[8] S.W.S. Sio et al., "*Malaria Count: An image analysis-based program for the accurate determination of parasitemia*", Journal of Microbiological Methods 68, Science Direct, pp 11-18, 2007.

[9] D. Anoraganingrum et al.' "*Cell Segmentation with Median Filter and Mathematical Morphology Operation*", in proceedings of on Image Analysis and Processing, Italy, pp 1043-1046, 1999.

[10] N. Otsu, "A threshold selection method from gray-level histograms", in proceedings of IEEE Transactions on Systems, Man and Cybernetics, 9(1), pp 62-66, 1979.

[11] K. Kim et al., "Automatic Cell Classification in Human's Peripheral Blood Images Based on Morphological Image Processing", Lecture Notes in Computer Science, vol. 2256. pp 225-236, 2001.

[12] T. Markiewicz, S. Osowski, "*Data mining techniques for feature selection in blood cell recognition*", European Symposium on Artificial Neural Networks, Bruges (Belgium), 26-28 April, pp 407-412, 2006

[13] N. Ritter, J. Cooper, "*Segmentation and Border Identification of Cells in Images of Peripheral Blood Smear Slides*", in Proceedings of Thirtieth Australasian Computer Science Conference (ACSC2007), CRPIT, 62, 161-169, 2007.

[14] G. Diaz et al., "A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images", Journal of Biomedical Informatics 42, Science Direct, pp 296–307, 2009.

Information Security Maturity Model

Dr. Malik F. Saleh *msaleh@pmu.edu.sa*
Management Information Systems, Chair
Prince Mohammad Bin Fahd University
Al Khobar, 31952, Saudi Arabia

Abstract

To ensure security, it is important to build-in security in both the planning and the design phases and adapt a security architecture which makes sure that regular and security related tasks, are deployed correctly. Security requirements must be linked to the business goals. We identified four domains that affect security at an organization namely, organization governance, organizational culture, the architecture of the systems, and service management. In order to identify and explore the strength and weaknesses of particular organization's security, a wide range model has been developed. This model is proposed as an information security maturity model (ISMM) and it is intended as a tool to evaluate the ability of organizations to meet the objectives of security.

Keywords: Maturity Model, Security Maturity Model, Security Measure, Security self study.

1. INTRODUCTION

The traditional information security objectives are confidentiality, integrity, and availability. Achieving these three objectives does not mean achieving security. Security is achieved by the prevention of attacks against information systems and from achieving the organization's mission despite attacks and accidents. One problem with organizations' security is that it is often viewed in isolation and organizations do not link the security requirements to the business goals. The rationale for these organizational problems is linked to the financial obligations that organizations face for unnecessary expenditure on security and control. Some of the information security efforts may not achieve the intended business benefit, resulting in lack of security and financial investments in systems that do not represent the core systems of an organization. For example managers can justify the need for a system that manages the resources at an organization. It is a relatively simple task to identify a system that adds value to an organization but to justify a second system to protect the first one might result in cancelling the investment of both systems. Any additional security investments are thought of as future projects that can wait until the business prospective is improved. Then, organizations are faced with the challenging task of recovering from an attack that disrupts the business process.

To ensure security, it is important to build-in security in both the planning and the design phases and adapt a security architecture which makes sure that regular and security related tasks, are deployed correctly [1]. Security requirements must be linked to the business goals through a process-oriented approach. The process must take into consideration many of the factors that affect the goals of an organization. We identified four domains that affect security at an organization. First, organization governance is one factor that affects the security of an organization. Second, the organizational culture affects the implementation of security changes in the organization. Third, the architecture of the systems may represent challenges to the implementation of security requirements. Finally, service management is viewed as a challenging process in the implementation.

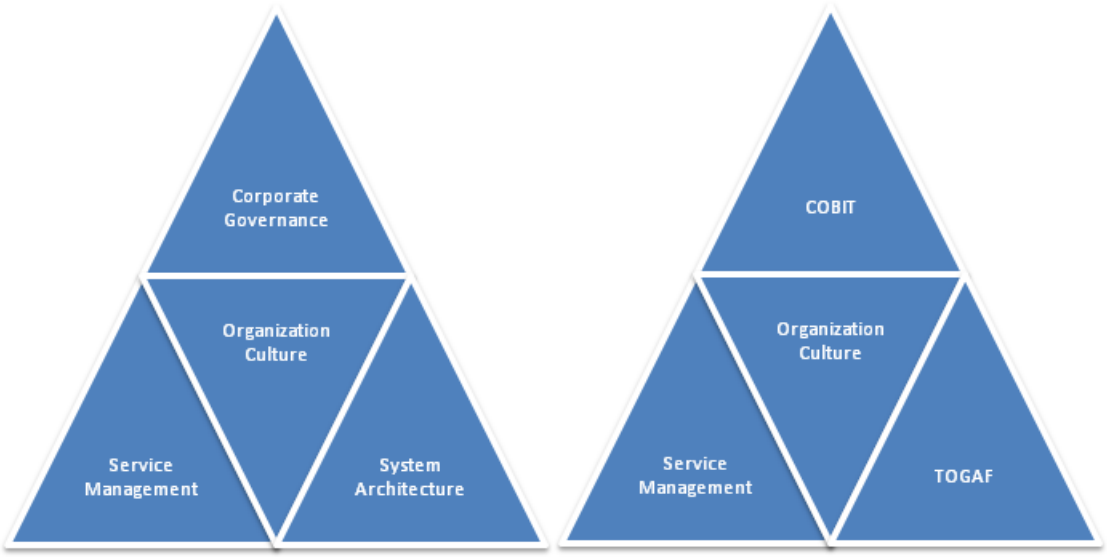


FIGURE 1: Domains mapped to implementation standards

This research narrows the gap between theory and practice for information security management by following the process of a security maturity model and by identifying the benefits of implementing a standard for organization security needs. We stress the fact of using a domain based approach to develop a model that can be widely used by organizations. This approach, if developed without an understanding of the organizational culture, will impact the effectiveness of the implementation and the human reaction to the use of new technologies. The organization culture often hinders the success of this approach and the delivery of the intended benefits of the implemented security model or standard.

2. BACKGROUND AND RELATED WORK

The motivation for this paper was due to challenges of assessing the implementation of security at organizations. In addition to implementation challenges, accomplishing best practices in the implementation of security is needed and it was undertaking in this research in the form of a self study that organizations would use to measure their information security practices.

Some attempt were undertaken to establish information security management maturity model. The ISM3 system was introduced to prevent and mitigate attacks, errors, and accidents that jeopardize security [2]. While this attempt recognized three levels of management responsibility, it did not provide best practices for the implementation of security.

An information assurance model was introduced based on none risk assessment model. It is based on diligence model where assurance is achieved by using threat and vulnerability reviews and countermeasures based on tangible best practices. The model did benchmarking, risk assessment and followed a diligence model [3]. This model introduces benchmarking but it did not provide best practices for security.

A certification and accreditation model through the identification of operational risks and the determination of conformance with established security standards and best practices was introduced by [4]. Its idea was to effectively establish trustworthiness for security services. While an organization policy and defined processes will introduced by [5] with appropriate accountability standards to facilitate compliance, monitoring and enforcement of security guidelines.

2.1 Domain-Oriented Approach

Senior management at organizations must become more IT literate to effectively synergize business strategy. In security, people, information, systems, and networks affect each others. These four domains

provide a vital link to all of the dynamic interconnections at an organization. Inside each domain, there are processes that identify, measure, manage and control risk.

Connecting different domains together requires securing each domain and securing the interconnection between the different parts. For the purposes of creating a widely used model that has good practices, security is looked at as domains, were each particular category of security represents knowledge in the organization. According to [6] there is no one-size-fits-all approach for maximizing the alignment of IT with the business and all of its components. Much depends upon the nature of the business, its size, markets, culture, and leadership style. Additional factors that help dictate the organization's alignment components and structure include the in-house IT capabilities and the dependence upon outsourcing.

2.1 Maturity Model

The concept of maturity models is increasingly being applied within the field of Information Systems as an approach for organizational development or as means of organizational assessment [7-9]. Any systematic framework for carrying out benchmarking and performance improvement can be considered as a model and if it has continuous improvement processes it can be considered a maturity model. Maturity implies a complete system. Generally, in the constituent literature maturity implies perfect or explicitly defined, managed, measured, and controlled system [10]. It is also a progress in the demonstration of a specific ability or in the accomplishment of a target from an initial to a desired end stage.

The Total Quality Management (TQM) maturity models is a structured system for meeting and exceeding customer needs and expectations by creating organization-wide participation in the planning and implementation of breakthrough and continuous improvement processes. It integrates with the business plan of the organization and can positively influence customer satisfaction and market share growth [11]. This structured system encompasses the entire organization and the goal is communicated on a regular bases while practicing what is being breached [12]. Quality can take many forms but its perception is dependant on the beholder. However, the emphasis is on things being done right the first time.

In order to identify and explore the strength and weaknesses of particular organization's security, a wide range model has been developed. The purpose is to identify a gap between the practice and theory which then can be closed by following a process-oriented approach. We introduce a maturity model that provides a starting point for security implementation, a common and shared vision of security, and a framework for prioritizing actions. Moreover, this information security model has five compliance levels and four core indicators to benchmark the implementation of security in organizations.

3. INFORMATION SECURITY MATURITY MODEL (ISMM)

This proposed information security maturity model (ISMM) is intended as a tool to evaluate the ability of organizations to meet the objectives of security, namely, confidentiality, integrity, and availability while preventing attacks and achieving the organization's mission despite attacks and accidents. The proposed model defines a process that manages, measures, and controls all aspect of security. It relies on four core indicators for benchmarking and as an aid to understanding the security needs in the organization. These indicators are goal-driven to achieve the security needs.

3.1 Levels of Compliance

It is hard for security practitioners and decision makers to know what level of protection they are getting from their investments in security. It is even harder to estimate how well these investments can be expected to protect their organizations in the future as security policies, regulations and the threat environment are constantly changing [13]. An information system would transition between several distinct vulnerability states. The first state is hardened and it occurs when all security-related corrections, usually patches, have been installed. The second is vulnerable and it occurs when at least one security-related correction has not been installed. The final state is compromised and it occurs when it has been successfully exploited [14]. Within these states, metrics need to indicate how secure the organization is so that the window of exposure can be minimized by the security operations teams in an organization by following a standard patching process to eliminate vulnerability and any associated risks. The security team either deploys patches after vulnerability was first disclosed or updates signatures that are associated with attacks.

The longer the window of exposure, the more the organization is exposed to attacks and exploits. The magnitude of risks is minimized if organizations are conscious about their security needs. Therefore the proposed ISMM considers five levels of compliance. Security is believed to improve as the organization moves up these five levels:

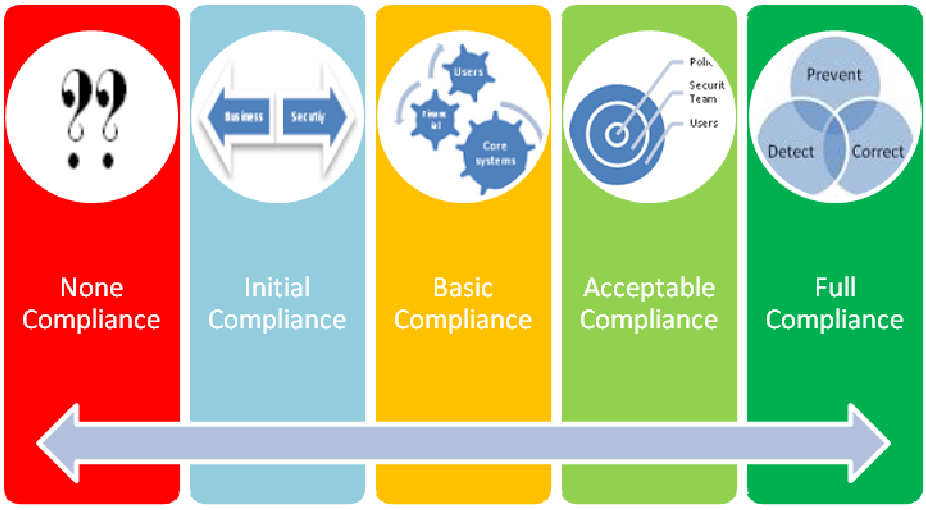


FIGURE 2: Levels of Compliance

3.2 None Compliance

This state is characterized by none existence of policies and procedures to secure the business. Management does not consider investing in security related systems necessary for the overall business strategies. In addition, the organization does not assess the business impact of its vulnerabilities and it does not understand the risks involved due to these vulnerabilities.

3.3 Initial Compliance

This state is the starting point for any organization. As long as an organization is conscious about the threats that their information systems face then that organization is considered in the initial state of compliance. This state is characterized by being chaotic, inconsistent, ad hoc, and in response to attacks and possibly because of losing resources due to an attack. Organizations recognize the business risks due to vulnerabilities but have no defined policies or procedures to protect the organization. In addition, the organization would have little practical implementation in security systems. Most implemented control will be reactive and not planned. The goals at the initial state are usually centered on the business activities of the organization and little attention is focused on securing the organization. The goals will change in response to attacks by implementing some kind of protection but it will not be continuous.

3.4 Basic Compliance

This state is the starting point for any organization that wants to protect its investment and ensure continuity. Application and network security is implemented but changes are not centrally managed and ad hoc security requests are common. In this state, organizations trust the interaction between the user and the systems. Security awareness programs are being considered for key resources only. IT security procedures are informally defined and some risk assessments taking place. In addition, responsibilities for IT security have been assigned but enforcement is inconsistent. Some intrusion and detection testing can also be performed.

A fundamental process to most systems is the interaction between the system and the user. According to [15], this interaction is the greatest risk. Organizations don't classify their users as threats to their systems. The user does not always cause a threat in isolation; rather, the actions of users are the starting point for some attacks, and in some cases, the users themselves may launch the attacks. Weak passwords, susceptibility to social engineering attacks, and failure to install security updates are some examples of why the user is classified as the weak human factor and the user's interaction with the systems create threats [16].

The goals at this level are usually centered on the business activities of the organization and the protection of core systems. Usually, an organization will consider the security of a system after the system's implementation. Two restrictions are faced at this stage: First, financial restriction and spending on systems that don't add value to the income of the business. Second, organizations classify their initial investments in security as completed. Organization will have a perception that their systems are protected and they become unaware of the threats and vulnerabilities.

3.5 Acceptable Compliance

This state is characterized by central management of all security related issues and policies. Users are trusted but their interactions with the systems are viewed as vulnerability. No ad hoc changes and central configuration models, from which all configurations are derived, are implemented. Security policies and procedures are now in place together with adequate delivery mechanisms to aid awareness and compliance. Access controls are mandatory and are closely monitored. Security measures are introduced on a cost/benefit basis and ownership concept is in place.

There is a school of thought that maintains that it is not the users' fault that they perform the easiest action; rather, it is the designers fault to have made the most insecure operation the easiest operation [16]. Since the actions of users are the starting point for some attacks, there is a need to inculcate a "culture of security" in users. Many users have to remember multiple passwords. They use different passwords for different applications and have frequent password changes, which reduces the users' ability to remember passwords and increases insecure work practices, such as writing passwords down [17]. For organizations to secure the interactions with their systems, communication between the security team and the users must take place to keep the users informed of possible threats. In addition, the users do not understand security issues, while the security team lacks an understanding of users' perceptions, tasks, and needs. The result according to [16] is that the security team typecast the users as threats that need to be controlled and managed, at worst, they are the enemy within. Users, on the other hand, perceive many security mechanisms as an overhead that gets in the way of their real work.

The goals at this state are usually centered on the business activities, the users, and monitoring security threats and all related patches are tested and implemented. Usually, organizations at this state are conscious about their security needs and they invest in systems that protect the organization.

3.6 Full Compliance

This state is characterized by having control over the security needs of the organization, monitoring the systems, being aware of threats and benchmarking by comparing the organization itself to other similar organizations and to international standards. In addition, a comprehensive security function has been established that is both cost effective and efficient which delivers high quality implementation. This comprehensive plan has formal policies and procedures in place to prevent, detect, and correct any security related issues. Also, corporate governance is aligned with the security needs of an organization. Corporate governance has policies for internal auditing which is independent and objective activity designed to add value and improve the security of the organization. The result of any audit activity is published and actions are implemented.

For organization to have full compliance security is managed by identifying the security concerns and security incidents are tracked in a systematic way. The organization must have proper policies for security in a formal sense and business plans would have items for security. The use of specific technologies throughout the organization is in a uniform manner and the implementation came to existence out of a business plan.

Full compliance also considers the security architecture in an organization. While the business architecture considers all external factors in an organization, the security architecture considers all users in the implementation. Policies are created to meet the needs of the users but information in or out of the organization is captured. A system for providing traceability through the organization is in place. Users are also involved in architectural analysis and the organization offers training for the users in security related issues.

As for management of security, policies in the full compliance state have preventive, detective and corrective control. The organization must have a system for reporting security incidents and for tracking the status of each incident. Installing anti-virus software and firewall is not enough to control the threats the organizations face. Email filters and intrusion detection systems must also be used to prevent many types of incidents.

4. MEASUREMENTS

Metrics are often used to predict future behaviors, based on historical data and trends.[13] argue that Security metrics are created and monitored as a way to get insights about the performance of these controls and to identify failure points or anomalies. However, the metrics are collected across organizations and they are operational metrics without the context of the overall security processes. On the other hand, measurement of any complex, operational system is challenging and security risks introduce another dimension of complexity. Risk management and the availability of different measurements and their properties will vary during the overall system lifecycle. Any measurement framework needs to be able to adapt to both the changes in the target of measurement and in the available measurement infrastructure. Security assurance measurements often require aggregation of several metrics, because direct measurement of the relevant properties is not often possible in practical complex systems and aggregation strategies can change from time to time, depending on the environment and the many risk factors [18].

4.1 ISMM Metric and Core Indicator

The principle that is followed here is what you can't measure, you can't manage. Therefore four core indicators are developed to manage and measure the compliance with this maturity model. Each indicator has its own key performance indicators that show the overall compliance with the model. These four indicators are domain specific rather than being process specific but they measure the aspect of structure, the management, the practices and the overall performance of the of the organization in term of its security.

The specific practices are intended as a guide for those responsible for the activities to draw their attention to good practices and to assist them to evaluate the practices at their organization. For each individual item, two responses are called for, but some items may not be applicable to the organization, therefore it should be marked with NA and ignored. The second response if applicable should be measured in term of assigning a five points rating scale to evaluate how well the practices are carried out. Certain activities require combining ratings to develop a broader rating. An overall rating of all domains would reflect the compliance with this maturity model according to table 1.

5. LIMITAION, IMPLICATION, AND RECOMMENDATION

The results of this paper clearly showed that there are metrics that can assess the implementation of security at organization. However, the use of a qualitative method incorporates various disadvantages and it is often criticized for being subjective and it lacks criteria to judge the trustworthiness and relevance of the results.

Much more research needs to be undertaken to accomplish best practices in the implementation of security by using a combination qualitative and quantitative research. Quantitative work will be undertaken to demonstrate the effectiveness of the proposed model. A survey of will be distributed to different organization and the result will be published in the near future.

6. CONCLUSION AND CONTRIBUTION

A systematic framework for carrying out benchmarking and performance improvement was developed. This model of best practices can be considered a maturity model which implies a complete system with continuous improvement. The objective of the proposed solution is to provide an organization with a way to conduct a self study of its implementation of security. The result will be measured in terms of compliance to the model. There are five compliance levels and each level consists of goals. An organization that continuously measure and audit its security implementation will achieve the highest level and it will achieve the objectives of security.

Full compliance to the model is characterized by having control over the security needs of the organization, monitoring the systems, being aware of threats and benchmarking by comparing the organization itself to other similar organizations and to international standards. Acceptable compliance is characterized by central management of all security related issues and policies. Other levels exist to raise a red flag for organizations that their security is weak and improvements are required.

The measurement indicators were domain specific rather than being process specific but they measure the aspect of the structure, the management, the practices and the overall performance of the of the organization in term of its security.

Management of Security

Security must be clearly and appropriately defined in the organization.

The scales ask you to indicate whether these practices are followed in your organization and to show how well this is done.
Evaluations should be based on valid evidence.

Good practices	**Is this true? Yes/No**	**How well is this done? (0-5 stars)**
1.1 Appropriateness of Management Practices		
1.1.1 Proper policies for security exist in a formal sense	☐	☐
1.1.2 Management considers it necessary to have policies for security at the organization	☐	☐
1.1.3 Management considers the organization security when making business plans. .	☐	☐
1.1.4 Management support and approval is vital to success of security implementation	☐	☐
1.1.5 The use of specific technologies throughout the organization is in a uniform manner	☐	☐
Overall Assessment (Average)		☐

Comment_____

_____

Priorities for improvement
_____

_____

Each sub-question is assigned a zero or one point. The sum of all section is assigned to the group

1.2 Types of Computer Systems Security used by the organization

1.2.1 Anti-virus software. . . . . ☐ . . . . . (0 or 1)

1.2.2 Firewall. . . . . ☐ . . . . . (0 or 1)

1.2.3 Passwords changed every 30, 60 days, etc. . . . . ☐ . . . . . (0 or 1)

1.2.4 E-mail filters. . . . . ☐ . . . . . (0 or 1)

1.2.5 Intrusion detection system. . . . . ☐ . . . . . (0 or 1)

Overall Assessment (Sum) ☐

Comment_____

_____

Priorities for improvement
_____

_____

1.3 Computer Security Concerns

1.3.1 What are the computer security concerns for your organization:
1.3.1.1 Computer viruses. . . . . [ ] . . . . . (0 or 1)

1.3.1.2 Denial of service . . . . . [ ] . . . . . .(0 or 1)

1.3.1.3 Theft of information. . . . . [ ] . . . . . .(0 or 1)

1.3.1.4 Breach of computer systems . . . . . [ ] . . . . . (0 or 1)

1.3.1.5 Misuse of computers by users. . . . . [ ] . . . . . .(0 or 1)

Overall Assessment (Sum) [ ]

Comment_____

_____

Priorities for improvement_____

_____

1.4 Computer Security Incidents

1.4.1 Security incidents at your organization:
1.4.1.1 Computers in your organization were used to commit fraud or embezzlement. . . . . (0 or 1) [ ]

1.4.1.2 Your organization detected viruses which infected your computer systems. . . . . (0 or 1) [ ]

1.4.1.3 A number of employees lost or forgot their passwords. . . . . (0 or 1) [ ]

1.4.1.4 Your organization detected a noticeable interruption of its Internet connection or e-mail service. . . . .(0 or 1) [ ]

Overall Assessment (**5** - Sum) [ ]

Comment_____

_____

Priorities for improvement_____

_____

Overall Assessment

1.1 **Appropriateness of Management Practices** . . . . . .

1.2 **Types of Computer Systems Security**. . . . . .

1.3 **Computer Security Concerns**. . . . . .

1.4 **Computer Security Incidents**. . . . . .

Combined Assessment (Average) . . . . . .

Comment _____

_____

_____

Service Management

Security must be clearly and appropriately defined in the organization.

The scales ask you to indicate whether these practices are followed in your organization and to show how well this is done.
Each sub-question is assigned a zero or one point. The sum of all section is assigned to the group
Evaluations should be based on valid evidence.

Good practices of Management	**Is this true? Yes/No**	**How well is this done? (0-5 stars)**
2.1 Appropriateness of the Service Management		
2.1.1 Does your organization classify Incidents.	☐	☐
2.1.2 Did you organization establish a Major Incident Response Team	☐	☐
2.1.3 Does you organization implement a problem management system.	☐	☐
2.1.4 Incidents Submitted via Automated Monitoring.	☐	☐
2.1.5 Does your organization maintain inventory of software and hardware equipment	☐	☐
2.1.6 Implemented changes must be approved by management.	☐	☐
2.1.7Change management is coordinated between the different teams	☐	☐
2.1.8 Annual budget proposals are submitted with detailed security requirements.	☐	☐
2.1.9 Financial resources are available and sufficient for security related systems. . . .	☐	☐
2.1.10 If performance is considered less than satisfactory clear requirements are established. . .	☐	☐
2.1.11 Employees are given appropriate and fair opportunities for development.	☐	☐
2.1.12 Recruitment processes ensure qualifications and verifications of candidates. . .	☐	☐
2.1.13 Effective systems are in place to ensure security of the property.	☐	☐
2.1.14 Effective systems are in place to ensure the personal security of employees . . .	☐	☐
Overall Assessment (Average)		☐

Comment_____

_____

Priorities for improvement

_____

_____

3.1 Management of Major Incidents

2.2.1 No. of problems not controlled through formal problem management. . . . . ☐

2.2.2 No. of problems with delays and deviations. . . . . ☐

2.2.3 Frequency of similar service level failures. . . . . ☐

2.2.4 No. of problems reported with known fixes. . . . . ☐

2.2.5 Managed escalations of problems. . . . . ☐

2.2.6 No. of problems escalated to higher levels. . . . . ☐

2.2.7 The level of service offered is more than what is expected . . . . . ☐

2.2.8 Period between request and implementation. . . . . ☐

2.2.9 No. of reruns and restarts. . . . . ☐

2.2.10 Frequency of support meetings. . . . . ☐

2.2.11 No. of operator interventions. . . . . ☐

2.2.12 Average age of equipment. . . . . ☐

2.2.13 No. of unplanned maintenance. . . . . ☐

Overall Assessment (5 - Average) ☐

Comment_____

_____

_____

_____

Priorities for improvement

_____

_____

_____

_____

_____

Overall Assessment

3.1 **Appropriateness of the Service Management** . . . . . .

3.2 **Management of Major Incidents**. . . . . .

Combined Assessment (Average) . . . . . .

Comment _____

_____

_____

_____

_____

_____

Enterprise Architecture

Security must be clearly and appropriately defined in the organization.

The scales ask you to indicate whether these practices are followed in your organization and to show how well this is done.
Evaluations should be based on valid evidence.

Good Practices of Management	Is this true? Yes/No	How well is this done? (0-5 stars)
4.1 Appropriateness of the Enterprise Architecture		
3.1.1 Users are involved in architectural analysis.		
3.1.2 Applications are upgraded to meet new architectural requirements		
3.1.3 The business capability (What the organization does) is known to all stakeholders		
3.1.4 The business architecture considers all external factors to an enterprise (including its customers, suppliers, and regulators).		
3.1.5 Information in or out of the organization is captured.		
3.1.6 Organization strategic goals that drive an organization forward are captured. . . .		
3.1.7 Strategic goals are mapped to metrics that provide ongoing evaluation of how successfully the organization in achieving its goals.		
3.1.8 A system for providing traceability through the organization is in place.		
3.1.9 A set of strategic, core and support processes that transcend functional and organizational boundaries as in place.		
3.1.10 The organization identifies and describes external entities such as customers, suppliers, and external systems that interact with the business.		
3.1.11 The organization describes which people, resources and controls are involved in its processes.		
3.1.12 The organization identifies gaps between the current and target business capabilities. . . .		
3.1.13 Business Architecture is directly based on business strategy.		
3.1.14 The business architecture derives the organizational structure.		
Overall Assessment (Average)		

Comment_____

_____

Priorities for improvement
_____

_____

3.2 Security Architecture

3.2.1 Centralized User Provisioning and Single Sign-On is implemented.		
3.2.2 New security architecture emerges as a result of security assumptions and designs being refreshed and updated to manage emerging threats		
3.2.3 Security is viewed as a service.		
3.2.4 In a service oriented architecture, your organization implement none centralized security		
3.2.5 In a none service oriented architecture, your organization implement centralized security		
3.2.6 Security is built-in in both planning and design phases		
3.2.7 The security architecture is capable of adapting new changes in technology, policy, and strategies.		
3.2.8 Security architecture implements policies, standards, and risk management decisions . . .		
3.2.9 Specialized security architecture is implemented for different security assumptions		
3.2.10 Central authentication service is implemented		
3.2.11 Different layers of security are implemented.		
3.2.12 Physical security is implemented.		
3.2.13 Personal security is implemented (host based).		
3.2.14 Network security is implemented.		
3.2.15 Information security is implemented.		
3.2.16 Application features are identified for security implementations.		

3.2.17 Software protection, that includes memory protection and proof-carrying code, is implemented.		
3.2.18 Database security ensures integrity, confidentiality, and availability.		
3.2.19 System audits are done regularly.		
3.2.20 Job descriptions include level of security risk.		
Overall Assessment (Average)		

Comment_____

_____

_____

_____

Priorities for improvement

_____

_____

_____

_____

_____

3.3 Continuous Improvement

3.3.1 Your organization continuously identifies gaps and addresses security issues. .		
3.3.2 The organization implements an incident reporting systems and the security team learns from incidents.		
3.3.3 Employees are trained on Security and threat awareness.		
3.3.4 All levels of the organization understand the importance of security and security is made into a priority.		
3.3.5 The security processes are documented and feedback is collected		
3.3.6 The organization measures the effectiveness of the security processes by tracking the number of attacks and the number of threats		

3.3.7 The organization plans for security changes		
3.3.8 Small scale changes are implemented		
Overall Assessment (Average)		

Comment_____

_____

_____

_____

Priorities for improvement

_____

_____

_____

_____

Overall Assessment

3.1 **Enterprise Architecture**.	
3.2 **Security Architecture**.	
3.3 **Continuous Improvement**.	
Combined Assessment.	

Comment _____

_____

_____

_____

Corporate Governance

Corporate governance must support security in the organization.

The scales ask you to indicate whether these practices are followed in your organization and to show how well this is done.
Evaluations should be based on valid evidence.

Good practices of Corporate Governance	Is this true? Yes/No	How well is this done? (0-5 stars)
4.1 Appropriateness of the Corporate Governance		
4.1 The organization complies with security policies.		
4.2 The organization explains why it is not complying with some security policies . . .		
4.3 The organization discloses the scope and responsibilities of the internal auditors		
4.4 The security team has Independent decision making		
4.5 Regulatory Compliance On Time		
4.6 Frequency of compliance reviews		
4.7 Frequency of internal compliance reports		
4.8 Level of satisfaction of the internal audit process.		
4.9 Delay between internal control deficiency and reporting.		
4.10 Number of auditors who are qualified.		
4.11 Number of incidents of non-compliance with internal controls		
Overall Assessment (Average)		

Comment_____

_____

Priorities for improvement
_____

_____

Overall Assessment

1.1 Corporate Governance. . . . . .

Comment _____

_____

_____

_____

_____

_____

Overall Assessment

1. **Management of Security**. . . . . .
2. **Service Management**. . . . . .
3. **Enterprise Architecture**. . . . . .
4. **Corporate Governance** . . . . . .

Combined Assessment (Average) . . . . . .

Combined Assessment	Stars	Compliance Level
0 – 1.5	One star	None Compliance
1.6 – 2.5	Two Stars	Initial Compliance
2.6 – 3.5	Three Stars	Basic Compliance
3.6 – 4.5	Four Stars	Acceptable Compliance
Above 4.6	Five Stars	Full Compliance

TABLE 1: Overall rating and Compliance Levels

7. REFERENCES

1. Amer, S.H. and J. John A. Hamilton, Understanding security architecture, in Proceedings of the 2008 Spring simulation multiconference. 2008, Society for Computer Simulation International: Ottawa, Canada. p. 335-342.

2. Aceituno, V. Information Security Management Maturity Model 2007 [cited 2011 July 11]; Available from: www.ism3.com/page1.php.

3. Al-Hamdani, W.A., Non risk assessment information security assurance model, in 2009 Information Security Curriculum Development Conference. 2009, ACM: Kennesaw, Georgia. p. 84-90.

4. Lee, S.W., R.A. Gandhi, and G.-J. Ahn, Establishing trustworthiness in services of the critical infrastructure through certification and accreditation. SIGSOFT Softw. Eng. Notes, 2005. **30**(4): p. 1-7.

5. Walton, J.P., Developing an enterprise information security policy, in Proceedings of the 30th annual ACM SIGUCCS conference on User services. 2002, ACM: Providence, Rhode Island, USA. p. 153-156.

6. Williams, P.A. IT Alignment: Who Is in Charge. [cited 2011 May 21]; Available from: http://www.isaca.org/Knowledge-Center/Research/Documents/IT-Alignment-Who-Is-in-Charge.pdf.

7. Ahern, D., A. Clouse, and R. Turner, CMMI distilled: A practical introduction to integrated process improvement. 2004, Boston, London: Addison-Wesley.

8. Chrissis, M.B., M. Konrad, and S. Shrum, CMMI: Guidelines for Process Integration and Product Improvement. 2008, Upper Saddle River, NJ: Addison-Wesley.

9. Mettler, T. and P. Rohner. Situational Maturity Models as Instrumental Artifacts for Organizational Design. in Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology. 2009. Philadelphia, Pennsylvania: ACM.

10. Fraser, M.D. and V.K. Vaishnavi, A formal specifications maturity model. Commun. ACM, 1997. **40**(12): p. 95-103.

11. V., P.P. Total Quality Management - A Strategic Initiative Gaining Global Compitative Advantage. 2010 May 21 [cited 2011; Available from: http://www.indianmba.com/Faculty_Column/FC1174/fc1174.html.

12. TQM - Total Quality Management. 2003 [cited 2011 May 21]; Available from: http://www.six-sigma-material.com/TQM.html.

13. Beres, Y., et al., Using security metrics coupled with predictive modeling and simulation to assess security processes, in Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement. 2009, IEEE Computer Society [download]. p. 564-573.

14. Arbaugh, W.A., W.L. Fithen, and J. McHugh, Windows of Vulnerability: A Case Study Analysis. IEEE Computer, 2000. **33**(12): p. 52 - 59

15. Schneier, B., Secrets and Lies: Digital Security in a Networked World. 2000, New York: John Wiley & Sons, Inc.

16. Vidyaraman, S., M. Chandrasekaran, and S. Upadhyaya, Position: the user is the enemy, in Proceedings of the 2007 Workshop on New Security Paradigms. 2008, ACM: New Hampshire. p. 75-80.

17. Brostoff, S. and M.A. Sasse, Safe and sound: a safety-critical approach to security, in Proceedings of the 2001 workshop on New security paradigms. 2001, ACM: Cloudcroft, New Mexico. p. 41-50.

18. Kanstrén, T., et al., Towards an abstraction layer for security assurance measurements: (invited paper), in Proceedings of the Fourth European Conference on Software Architecture: Companion Volume. 2010, ACM: Copenhagen, Denmark. p. 189-196.

Toward a New Algorithm for Hands Free Browsing

Murad Al-Rajab *m.alrajab@alhosnu.ae*
Faculty of Engineering/Software Engineering Department
ALHOSN University
Abu Dhabi, 38772, UAE

Haifaa Kattan *h.kattan@alhosnu.ae*
Faculty of Engineering/Software Engineering Department
ALHOSN University
Abu Dhabi, 38772, UAE

Abstract

The role of the Internet in the life of everyone is becoming more and more important. People who usually use the internet are normal people with use of all physical parts of their body. Unfortunately, members of society who are physically handicapped do not get access to this medium as they don't have the ability to access the internet or use it.
The Integrated Browser is a computer application that browses the internet and displays information on screen. It targets handicapped users who have difficulties using a standard keyboard and/or a standard mouse. In addition, this browser, unlike the conventional method, works by responding to voice commands.
It is characterized as an economical and humanitarian approach because it is handicap-friendly and is especially helpful to those who cannot use their hands. The Integrated Browser is a state-of-art technology that provides better and faster Internet experience. It is a program that will facilitate and enhance learning of slow or/and handicapped learners

Keywords: Internet Browsing, Handicapped, Computer Application, Voice.

1. INTRODUCTION

Internet is a global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP) to serve billions of users worldwide [1].
A web browser is a software application for retrieving, presenting, and traversing information resources on the World Wide Web [2].
The handicapped person is a person who has a special condition that markedly restricts his ability to function physically, mentally or socially [3]. The IB (Integrated Browser) application tries to combine these three terms (Internet, web browsing and handicapped) to work all together and become a reality.
There are several different web browsers available in the internet world. The several types of browsers could be used by normal people who can use their hands. Internet usage has become a part of the daily life in a society that contains different types of people such as children, adults, and handicapped, each with their different abilities. Unfortunately, the handicapped faces difficulties in using the different general browsers. There is a need to create a special browser to target these people. This paper aims to develop an internet browser with a new approach in the technology world. The change in the routine style of browsers is a state-of-art design which provides a multifunction process through combining voice.
This paper has worked out to combine the voice with the internet browsing, which would help the handicapped to use the internet in an easy and direct way. Assuming that the computer's operating system is running on and that the program is in the startup programs, the focus would be on browsing the WWW through voice commands.
A new focus on accessibility should be kept in mind as speech technology has "always" (at least since the mid-1990s) implicitly addressed users with visual or mobile disabilities, sometimes

"disguised" as a more general goal of enabling eyes-free/eyes-busy or hands free/hands-busy access to web browsing and other applications [4].
What makes the IB a different browser is its backbone algorithm that supports the voice commands. The paper will discuss an algorithm which will develop a voice command browser that targets handicapped people. In addition, the IB can be used by any individual, not only the handicapped.
The IB has its own new style of browsing that is shown in the new design of the browser. It has many built in functions, such as News Line, Add Rotator, Games, Quick Search Button, Quick Email, and other useful functionality.
Bringing up the idea of having a voice command browser is one way of assisting and supporting the handicapped. There are other approaches of developing a voice command browser, but the idea discussed in this paper is to develop the IB by applying a new algorithm technique.

1.2. Technical Objectives

1. To allow the handicapped to interact and navigate the web using their voice.
2. To investigate the possibility of using the voice Recognition Technology, which is a special technique, needed to achieve success and use special components such as the programming language and specific grammar to match the key words.
3. To help workers multitask. If a worker needs to type, explore or navigate the internet all at one time, the voice would help achieve this simultaneously.
4. To investigate the difference between other browsers we used different functions and applications imbedded to show the IB different style such as:
   - The voice recognition for facilitating the work on browsing the internet.
   - The ability to change the background color through a choice of different colors to allow for personal style.
   - The particular style of other browsers was avoided through using different menus on different locations and adding the advertisement rotator.
   - A rotating line for news was added.
   - The user has many other facilities in this browser such as making pop ups optional, selecting size and other additional features.

2. LITERATURE REVIEW

2.1 Definition of the Internet Browser

A browser is a software program that allows you to view and interact with various kinds of internet resources available on the World Wide Web (WWW). A browser is commonly called a web browser [5]. Internet browsers are of different types and designs; some are closed source (i.e. Internet Explorer) and some are open source (i.e. Mozilla Firefox). Also, browsers are available with different add on tools that facilitate their functions and usability.

2.2. Significance of the Voice Internet Browser

Access to information has become a major economic and social factor [6]. Voice browsing technology is a rapidly-growing field. Whether or not it proves to be the next internet, it deserves a careful examination in its present form [7], as the need for an easy and direct way to access the internet has become a demand for many types of people, especially the handicapped. From this, the idea of the paper was raised in order to help implement a voice command IB browser.

2.3. Previous Work and Algorithms

In many studies, algorithms and applications have been implemented to facilitate the internet browsing or to revolutionize the traditional way of surfing the www in different angles. One application was implemented to display enumerated links in the browser window and to have also a compass mouse with a curser positioned over a mouse-over pull-down menu by speech recognition [8].
On the other hand there was an implementation for a multi-model browsing system that allows adding automatic speech recognition functions to standard Internet browsers. It was designed for a real Web Application designed for a medical domain [9].

Other implementation for a speech recognition system based on an Internet client-server model as a Java Applet that records voice at the client computer sends the recorded speech file over the Internet, and the server on the end point recognizes the speech and displays the recognized text [10].

In addition there was a contribution for Navigation by Speech system in which allows the user to control by speech a subset of navigation facilities like the basic tool bar commands, URL spelling, following hyperlinks and page printing [11].

This paper will track the voice browsing in a new algorithm approach through the "vMatrix" algorithm.

3. Design and Implementation Criteria

3.1. Methodology Adopted in This Research

A combination of the case study and the personal interview was deemed appropriate for this research. The case study is an in-depth examination of a behavior, concept, or phenomenon. Complementary aspects of the case study are experiments and surveys. This research approach can be helpful in analyzing a real situation, and can serve as a strong basis for debate.

Interviews with system analysts were conducted to clarify their ideas about browsers, understand the main performance and applications of browsers and to gain an insight into the main problems they may have faced through working with such applications. In addition, interviews with handicapped people were also conducted to know their demands for such a browser and the degree they like to navigate the internet. A group discussion with some computer academics took place in order to discover the algorithm efficiency to be implemented.
From all the previous interviews conducted, a clear idea was forming about browsers' applications, their performance, and the user requirements needed for that application.
IB user interface was designed using a visual programming language (Visual Basic 6.0). A flash programming application (Swish max program) was used in addition to the pictures in order to create an attractive style. The sound was enhanced inside the browser which would facilitate many functions and improve the efficiency of the browser.
The IB explores and navigates the internet in a user friendly way that will synchronize the feature requests.
Questioners were distributed to a selection of individuals in the research area for the purpose of gathering the names of the most popular and enjoyable websites navigated. Then we added what was suggested to the implemented browser as a step in developing the algorithm.

There are two options to start the IB software. The first option is to add it to the active task bar, while the second option is to install a small tool that converts the operating system functions to voice.

In this paper, an algorithm solution on how to open and navigate websites only was provided. This means that the system will operate on a one voice command and browse the internet. This will be done without the interaction within the website which also can be implemented using the same algorithm technique.

4. Requirements Analysis

4.1. System Requirements

The system requirements set out the system's functions, services, and operational constraints in detail, and they should be precisely expressed, and define exactly what is to be implemented. They are often classified as functional requirements and non-functional requirements.

4.2. Non-Functional Requirements

4.2.1. Operational Requirements
The software will operate in a Visual Basic environment.
The software will be able to import .Jpeg and .Gif graphics files.
The software will be able to import .WAV sounds files.
The software will be a portable one (instable on a CD)
The software will be able to import .swf files (flash).
The software will be able to interact with the internet.

4.2.2. Security Requirements
Users are not allowed to make any changes to the system. Only programmers have the key to modify and change the system.

4.2.3. Cultural and Political Requirement
No special cultural and political requirements were included, but the system can be enhanced to cater to the Arabic culture and/or political environment.

4.2.4. Usability of the System
Since the system targets handicapped people, implementing a user friendly interface was taken into consideration during the design and the implementation phase. In this application we used a comic character that assists the users in using the voice system, redistributed the functional buttons, and color controlled the background. These are key in a user friendly interface which would make it easier and preferable.

4.3. Functional Requirements
1. The user can interact with the interface of the system to find what he/she wants and the system will response directly.
2. The user can select by voice the sites he/she would like to navigate.
3. The user can have the choice to enable/ disable pop ups on the sites.
4. The user has the option to turn the voice On/Off in order to minimize noise effect.

5. Design
The design phase is a part of the System Development Life Cycle (SDLC), which is a blue print for the new system. Also, it guides the project team through planning.
An important initial part of the design phase is the examination of several design strategies to decide which would be used to build the system. The next step would be designing the user interface, system inputs, and system outputs, which involve user interaction with the system.

5.1. Design Strategy (Methodology)
The study of this software has revealed that the appropriate approach to create the new system is to use the custom development. Through this approach, there was room for flexibility and creativity how business problems were resolved. Also, it helps build technical and functional knowledge within the team.
The business need of the software is to build an effective, useful, and compatible package for the internet browsers.

5.2. Architecture Design
Software Specification:

	Standard client	Standard App. window
operating System	Windows	Windows
special Software	VB Components Swish Flash	Visual Basic 6.0

TABLE 1: Software Specifications

6. IMPLEMENTATION

This section discusses the activities needed to successfully build an information system that consists of programming and coding. Programming is often seen as the focal point of system development as system development is basically writing programs. It is the reason why all the analysis and designs are done.

6.1 Coding

During designing the user interface, different objects and items were inserted. Behind each of these objects and items, codes were added to activate and make them work properly.

6.2 Interface Design

Interface Design is the process of defining how the system will interact with the external entities (system users or other systems). It describes the layout of the pages and the flow of events. It is also concerned with where and how data are represented on the pages [12].
However, in a User Interface Design, needs, experience, and capabilities of the system users must be taken into account. In addition, the designers should be aware of the users' physical and mental limitations (e.g. limited short-term memory and people's tendency to make mistakes) [13].

The following illustrates some of the system's graphical interfaces.

- **News Line:** This Line displays the latest news from all around the world. It is just an RSS linker that is linked to a news agency.
- **Address Bar:** This allows the user to type the site he/she would like to navigate.
- **Add Rotator:** This is called the Advertisement Rotator, which is used to display the different online advertisements.
- **Search Button:** This button will take the user to a search engine page which will help him/her to search for any information.
- **E–Mail:** This button will take the user to an E–Mail Editor Form that enables the user to write an E–Mail and send it.
- **Add to Favorite Button:** This button helps the user to store any sites to quickly reference them without the need to retype their URL in the Address Bar again.
- **Games:** This is a button which allows the user to choose from different types of games to play.
- **Micro – Browser:** This is a button when clicked, will allow a Micro (smaller) size of the IB to appear in front of the old one but in a smaller size.
- **Calculator Button:** This button will display a Mathematical Calculator.
- **Speak Help:** This is a hint that facilitates the usage of the voice browser.
- **Site Display Area:** This is the area on which the navigated site will be displayed for user.

IB begins with a main page which has Buttons and Components as shown in Figure 1 below:

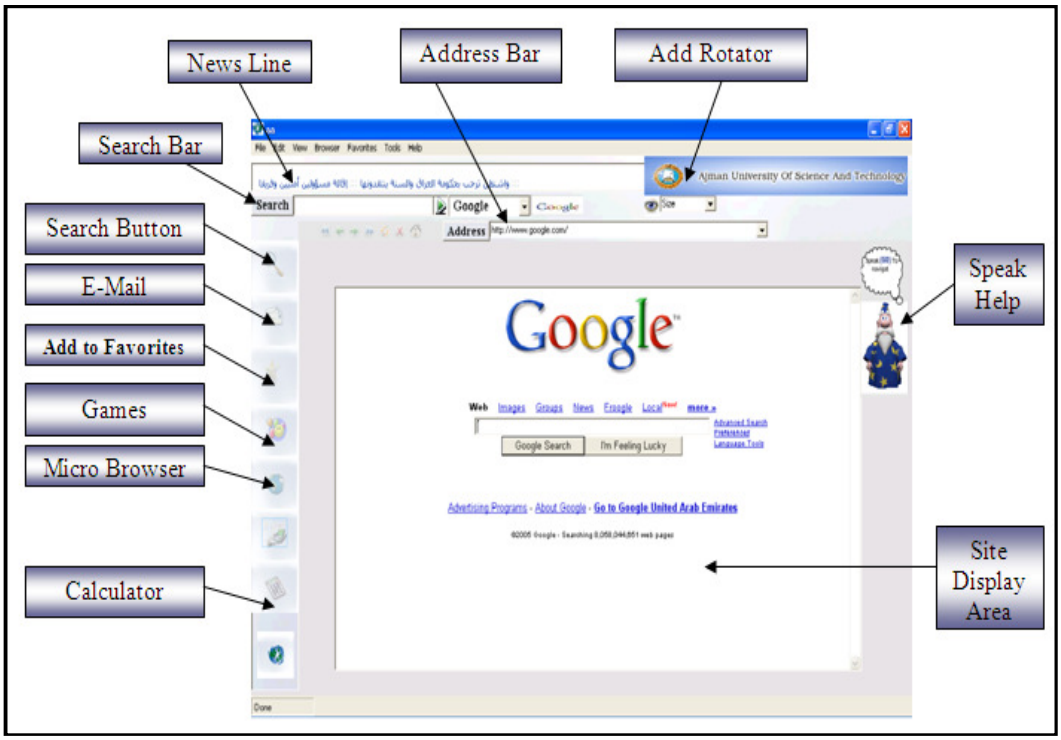


FIGURE 1: User Interface Design Shows Interaction between the System and the User.

7. Algorithm Behind

An Algorithm named "vMatrix Algorithm", which stands for the voice matrix idea, was implemented. A grammar was implemented as our database engine. A survey was conducted for the most favorite websites, using a random sample of students, faculty, and staff at a university. Then, a form of one matrix (which can be expanded) was implemented. The row and column headings were alphabeticized sequentially from a to z , because we implemented only a sample of database websites (36 websites), that were fit in a 2-D matrix inside 1 form (window) as shown in Figure 2, the headings were selected of the rows and columns to start from a up to f.

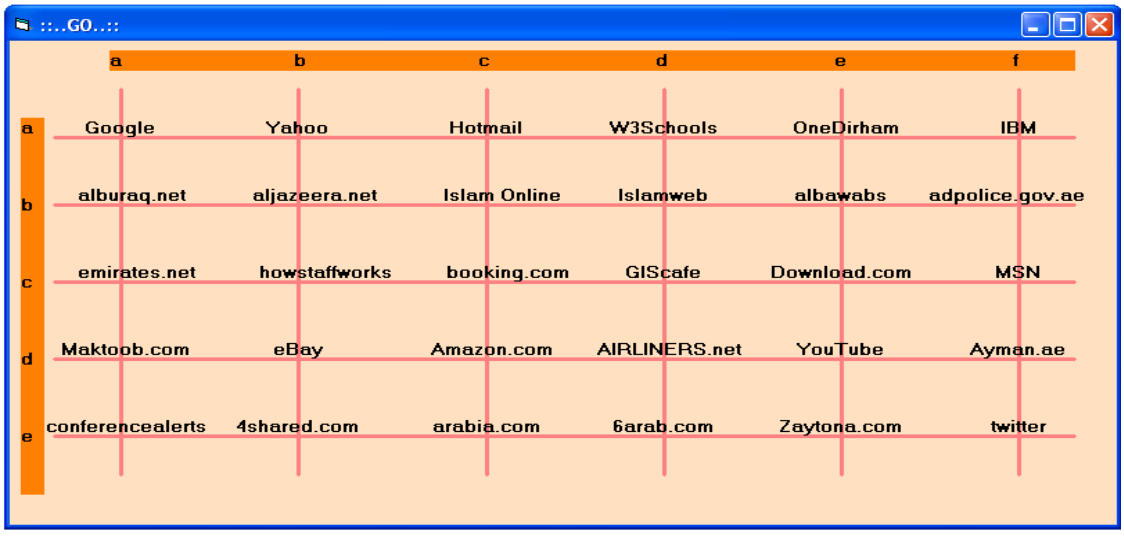


FIGURE 2: Implementing Form of One Matrix.

Once the IB user sees the Comic Character (Speak Help Picture), the above form will be displayed just once for the user to speak or utter the word “Go”. The matrix would help the user choose the site he would like to navigate by matching the intercept of the column and row. For example, if the speaker says “ca”, then the system will respond and open the “Hotmail” site and so on each time the user navigate a different website.

In addition, the user has the ability to say “ON”, which would switch the microphone on or say “OFF” which would turn the voice recognition of the microphone to the off mode.

Also, the functionality of the IB was enhanced by enabling the user to deal with the browser functions such as (Home, Refresh, Backward, and Forward) through voice command control. For example, the user can say “Home” and the browser will respond and take the user to his/her home page.

8. RESULTS AND DISCUSSIONS

The “vMatrix” algorithm can efficiently facilitate the voice recognition of the website spoken but it depends on the characters announced rather than words. For example if the user wants to go to “Hotmail”, he/ she is not going to say “Hotmail” instead he/ she will say “c” “a” as two consequence characters. In this way the proficient of recognizing the voice will be easier and less confusion. While if the user will pronounce “Hotmail” the system may recognize it as two words like “Hot” “mail”, and may think that there is another site that starts with the word “Hot”. The 2-D “vMatrix” algorithm character recognition is more efficient in this case and less noise and confusion.

Also if compared to other implementations, most had just attempted to apply voice for a static single web-page but not for surfing different websites, like adding numeric points beside the links inside the webpage or just say the links as words. Others tried to have a compass mouse algorithm which put reasonable numbers on the website in compared to our user interface which simply has no numbers to search for in order to pronounce them. In our user interface and our algorithm efficiency the user will reach the site he /she desires through an easy and forward steps and spoken tokens.

9. CONCLUSION AND FUTURE WORK

These days, internet is playing a major role all over the world. The main gate to access the internet is browsing. As browsing was designed to cater to only normal people with no physical disabilities, the physically handicapped were isolated from accessing the internet. IB was developed to help enable the handicapped to browse the internet easily. This objective was achieved through a combination of voice commands with internet browsing; IB is a voice command driven browser for the disabled.

It should be recognized that working using voice commands would give us opportunities to find ideas that could be developed in the near future. Some of these ideas can be the diversity and difference of the pronounced accent while searching, interaction between the methods of the net browser with the matrix system, and the work of the PC by recognizing the eyes of the user (the vision method) as long as he/she is browsing the net.

10. ACKNOWLEDGEMENT

The authors acknowledge the appreciated support of the following people: Mr. Hasan Abdul Rahman, Mr. Maher Wasef and Ms. Hiba Moamen.

11. REFERENCES

[1] Wikimedia Foundation, I. (2001). “Web browser”. Available at: http://en.wikipedia.org/wiki/Web_browser [Accessed 7th November 2010]

[2] Wikimedia Foundation, I. (2001). “Internet”. Available at Wikipedia: http://en.wikipedia.org/wiki/Internet [Accessed 10th October 2010]

[3] Farlex, I. (2004, December). “Handicapped person”. Available at: http://www.thefreedictionary.com/handicapped+person [Accessed 21st December 2010]

[4] T. Brøndsted, E. Aaskoven. “Voice-controlled Internet Browsing for Motor-handicapped Users. Design, and Implementation”. Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal. 2005. pp. 185-188.

[5] Consulting, N. C. (2000). “Internet Browser”. Available at: http://www.netdigix.com/gloss/browser.php. [Accessed 22nd September 2010]

[6] K.Peter, C. Wootton, I. Wechsung, S. Moller. “Evaluation of a Voice-Based Internet Browser with Untrained and Trained Users”. In proceedings of the 5th International Conference on Universal Access in Human-Computer Interaction. 2009. pp 482-491.

[7] Community, A. M. (2000). "Voice Browsing". Available at: http://www.wirelessdevnet.com/channels/voice/features/voicebrowser.phtml. [Accesssed 12th December 2010]

[8] T Brondsted, E. Aaskoven. “Voice-controlled Internet Browsing for Motor-handicapped Users”. Annual Conference of the International Speech Communication Association – INTERSPEECH. 2005. pp. 185-188.

[9] C. Eccher, L. Eccher, D. Falavigna, L. Nardelli, M. Orlandi, A. Sboner. “ON THE USAGE OF AUTOMATIC VOICE RECOGNITION IN AN INTERACTIVE WEB BASED MEDICAL APPLICATION”. IEEE International Conference on Acoustics, Speech, and Signal Processing. 2003. pp 289 - 92 vol.2.

[10] Z. Tu, P. Loizou. “SPEECH RECOGNITION OVER THE INTERNET USING JAVA”. IEEE International Conference on Acoustics, Speech, and Signal Processing. 1999. pp 2367 - 2370 vol.4.

[11] L.J. Adams, R.I. Damper, S. Harnad and W. Hall. “A System Design for Human Factors Studies of Speech-Enabled Web Browsing”. In: ESCA Workshop on Interactive Dialogue in Multi-Modal Systems, Kloster Irsee, Germany. 1999. pp. 137 – 140.

[12] Wixom., A. D. (2009). "Systems Analysis and Design", John Wiley & Sons, Inc.

[13] T. Lemlouma, N. Layaida. "Context-Aware Adaptation for Mobile Devices".In Proceedings of the 5th IEEE International Conference on Mobile Data Management. Barkeley, California, USA, 2004.

A Distributed Approach to Defend Web Service from DDoS Attacks

Monika Sachdeva monika.sal@rediffmail.com
Assistant Proff./Department of Computer Science & Engineering
SBS College of Engineering & Technology,
Ferozepur, Punjab, India

GurvinderSingh gsbawa71@yahoo.com
Associate Proff./Department of Computer Science & Engineering,
Guru Nanak Dev University,
Amritsar, Punjab, India

Kuldip Singh kds56fec@riitr.ernet.in
Retd. Proff./Department of Electronics and Computer Engineering,
Indian Institute of Technology,
Roorkee, Uttrakhand, India

Abstract

Most of the business applications on the Internet are dependent on web services for their transactions. Distributed denial of service (DDoS) attacks either degrade or completely disrupt web services by sending flood of packets and requests towards the victim web servers. An array of defense schemes are proposed but still defending web service from DDoS attacks is largely an unsolvable problem so far. In this paper, DDoS defense schemes are classified into centralized and distributed and their relative advantages and disadvantages are explored. An ISP based distributed approach is a pragmatic solution to defend from DDoS attacks due to its autonomous control, more resources, and incremental scope. Traffic cluster entropy is conceptualized from source address entropy and the combination is used to detect various types of DDoS attacks against the web service. A framework is proposed which can detect the attack, characterize attack sources, and filter the attack packets as early as possible so as to minimize the collateral damage

Keywords: DDoS, Centralized Defense, Distributed Defense, Deployment, Detection, Response, Source Address Entropy, Traffic Cluster Entropy.

1. INTRODUCTION

Internet has changed the way traditional business models are operated. Web service is one of the most important facilities used by commercial and government organizations to perform their activities. However DDoS attacks against high profile sites in the recent past have manifested their devastating power and have raised unresolved issues related to Web security [1]. A lot of research [2][3][4][5] has been carried out to defend web service from DDoS attacks, but none of these schemes are able to handle DDoS attacks in a comprehensive manner. The stumbling barrier has been the vulnerabilities in the Internet infrastructure and the volume of legitimate looking attack traffic generated towards the web server which makes defense system itself susceptible against these attacks [6]. Due to sheer volume, most of these schemes crumble as their bandwidths, data structures and CPU cycles are exhausted in handling the spurious attack traffic only [7]. So the biggest need is to design a DDoS resistant scheme.

In this paper an emphasis is laid on use of distributed approach to defend web service from sheer volume of DDoS attack traffic by dividing computational overheads at multiple points so that the approach itself should be DDoS resistant. Since detection of DDoS attacks requires monitoring

and analysis of complete traffic, so a technique, which can monitor and analyse traffic at distributed points, but actually behave as if the total traffic is monitored and analysed at single point, is good for DDoS attack detection [8]. Moreover characterization of attack traffic and then filtering also consumes computational resources, so they should also be distributed as far as possible. A framework is proposed in this paper to defend web service from DDoS attacks. It has following characteristics:-

1. Monitoring and analysis of traffic is distributed.
2. Complete traffic analysis for DDoS attack detection is carried out.
3. Defense is DDoS resistant so that automatic response may be generated.
4. Characterization of attack traffic is separately carried out from attack detection.
5. Filtering is done at distributed points.

An amalgamate approach of source address and traffic cluster entropy is used for attack detection. Kumar et al [8] formula has been used to compute entropies at one point collected from distributed points. The computed entropies are compared with base line entropies for signaling attacks. Characterization of attack traffic is based on finding new source addresses and cluster based on profiled traffic matrices. The attack signatures are communicated to the entry points so that they may be filtered without wasting core bandwidth.

The rest of the paper is organized as follows. Section 2 focuses on justifying distributed approach rather than centralized in a pragmatic manner. Section 3 discusses our detection approach. Section 4 explains proposed framework. Finally section 5 concludes our paper.

2. RATIONALE BEHIND DISTRIBUTED APPROACH

A comprehensive DDoS solution requires three effective modules namely traffic monitoring, traffic analysis, and attack traffic filtering [6] [8]. In a centralized solution all the modules are deployed at same place whereas voluminous and distributed nature of DDoS traffic demands a distributed DDoS solution because centralized solutions cannot handle high overheads of monitoring, analyzing and filtering. Components of distributed defense system are deployed at different locations and cooperate with each other to defend from the attacks. Compared with the centralized defense systems, distributed defense systems can discover and fight the attacks with more resources and at more than one point of the Internet. It is very difficult for the centralized defense system to detect the attack at the beginning. When the attacks are full-fledged, it becomes more difficult for defense system to resist the flooding. Moreover centralized defense systems are themselves more vulnerable to be attacked by hackers. The centralized defense systems are mostly deployed on the victim network because of economic reasons. Thus such defense systems are irresponsible systems which could only detect the attacks but cannot generate automatic alert and are also not able to filter the attack traffic themselves.

Distributed defense systems overcome the shortcomings of centralized and isolated defense systems. Deployed on all around the Internet, distributed defense systems can detect the attacks before they are launched by inspecting the traffic on many edge networks in which the computers are compromised by hackers. The most important and attractive feature of the distributed defense system is that the components in the distributed defense system can cooperate with each other to fight against DDoS attacks.

Centralized	Distributed
All the component modules are deployed at same place.	Whereas in distributed they are deployed at multiple places.
Highly Vulnerable and hence not robust against DDoS attacks.	Less Vulnerable and hence robust against DDoS attacks.
No cooperation and communication framework required.	Cooperation among various modules and proper communication framework required
Lesser resources are available for fighting against the attacks	More resources are available for fighting against the attacks
Mostly deployed at Victim site	Deployed at Victim-Core, Throughout the Internet and Victim-Source

TABLE 1: Centralized Vs Distributed defense

The advantage of distributed over centralized defense has been recognized in [9-11] [12]. A comparison of centralized Vs distributed is given in table 1.

Clearly distributed defense is the only workable solution to combat DDoS attacks. Some recently proposed defense systems use collaborating source-end and victim-end nodes [10], while others deploy collaborating nodes at the victim and core networks [13]. While they perform well against a variety of attacks, they do not completely handle the flooding DDoS threat. Specifically, source-victim defense systems fail to handle large attacks launched from legacy networks, while victim-core defense inflict high collateral damage to legitimate traffic. A few defense schemes combine defense nodes at all three locations [9] [11]. These defense mechanisms achieve higher effectiveness, but focus on a single approach to defense (e.g., a capability mechanism in [11], victim-hiding in [9]), which ultimately discourages integration with other defense approaches and wide deployment and hence are not practical. So a practical distributed defense mechanism which can have wide deployment is the need of the hour. Many distributed defense techniques are proposed in the literature. Distributed DDoS defense can be deployed at source, victim and intermediate, source-victim, and victim-intermediate networks.

Distributed defense techniques are likely to be the proper solution for handling the DDoS threat [14]. However, they are infrastructural solutions i.e. they span multiple networks and administrative domains and represent major undertakings of many Internet participants. Such systems are difficult to deploy and maintain. Further, the required cooperation of defense systems is hard to achieve due to distributed Internet management and strictly autonomous operation of administrative domains. Securing and authenticating the communication channels also incurs a high cost if the number of participants is large. In light of above said issues and Internet design vulnerabilities [3], a practical DDoS defense system deployment should have following important characteristics:

- Autonomous system i.e. whole defense location under one administrative control so that different defense nodes can collaborate in a secure manner.
- Large and infrastructure wise rich enough to handle high voluminous traffic from evenly distributed flood sources.
- Capability to evolve DDoS defense in incremental fashion.
- Sufficient financial motivation for value-added DDoS security service.

The Internet consists of thousands of Autonomous Systems (ASes) i.e., networks that are each owned and operated by a single institution. Usually each ISP operates one AS, though some ISPs may operate multiple ASes for business reasons (e.g. to provide more autonomy to administrators of an ISP's backbones in the United States and Europe) or historical reasons (e.g. a recent merger of two ISPs) [15]. An ISP has total autonomy to collaborate defense nodes in a secure manner. Enough infrastructures can be provided for DDoS defense to handle high volume at ingress points. Moreover, once agreement is reached between various ISPs then inter co-operation among ISPs is also possible [16, 17]. Accordingly, there is scope of incremental DDoS defense. If a provider's infrastructure is attacked (routers, DNS, etc.), all services to its customers fail, resulting in service level agreement (SLA) violations. Moreover, ISPs normally host most of the services available on the Internet. The cost of DDoS protection is insurance against catastrophic failures that would cost the business orders of magnitude more in terms of both revenue and negative customer relations. However, Cost-avoidance is not the only motivation to implement a complete DDoS solution in ISP domain. For the users, DDoS protection can also be offered as a value-added service that creates new revenue streams and provides competitive differentiation for ISPs. In nutshell, ISP level DDoS defense is most practical and viable at this stage. Though, longer term objective "how to achieve inter ISPs cooperation" still remains as the biggest challenge.

3. SOURCE ADDRESS AND TRAFFIC CLUSTER ENTROPY AS A DETECTION METRIC

Most of detection schemes in the literature fail to address a very important scenario comprising of legitimate increase in traffic called Flash events (FE) [18]. We have proposed an anomaly based approach to detect DDoS attack as well as to discriminate it from FE. Clearly first a base line behaviour of the system is required and then the same is compared with actual behaviour. If actual behaviour significantly deviates from normal behaviour then we raise an alarm for attack. Shannon entropy [19] has been used to conceptualize source address entropy [8][20] and traffic cluster entropy. The source address entropy and traffic cluster entropy are compared in different scenarios: normal and DDoS attacks, normal and flash, and flash with DDoS attack. Basic terminology and symbols used are explained below:-

Source IP address (src_IP):- A 4-byte logical address used in the packets to represent its source IP.

Traffic cluster (tc):- The traffic generated from same networks or administrative domains is defined as traffic cluster.

16-bit traffic cluster identifier ($tc16\_id$):- All the packets which share the same initial 16 bits of their src_IP are in same group called 16-bit traffic cluster. It is obtained by bit-wise AND operation of src_IP and 16-bit mask i.e. 255.255.0.0. A unique identifier assigned to such a traffic group or cluster is defined as 16-bit traffic cluster identifier.

24-bit traffic cluster identifier ($tc24\_id$):- All the packets which share the same initial 24 bits of their src_IP are in same group called 24-bit traffic cluster. It is obtained by bit-wise AND operation of src_IP and 24-bit mask i.e. 255.255.255.0. A unique identifier assigned to such a traffic group or cluster is defined as 24-bit traffic cluster identifier.

Source address entropy $H(src\_IP)$:- A metric that captures the degree of dispersal or concentration of distribution of a random variable is called sample entropy [8][20]. Let the random variable src_IP can take values $\{src\_IP_1, src\_IP_2, src\_IP_3, \ldots\ldots src\_IP_n\}$ in different packets. Let number of packets received per src_IP are $\{X_1, X_2, X_3, \ldots\ldots X_n\}$ respectively. Then as per Shannon criteria sample entropy is

$$H(src\_IP) = -\sum_{i=1}^{n} p(src\_IPi) \times log_2 p(src\_IPi) \quad (1)$$

Here the probability of occurrence of src_IP i.e. $P(src\_IP) = \{p(src\_IP_1), p(src\_IP_2), \ldots\ldots p(src\_IP_n)\}$ is computed as $p(src\_IPi) = \frac{Xi}{S}$ where $S = \sum_{i=1}^{n} Xi$

Traffic cluster entropy $H(tc\_id)$:- Let the random variable tc_id can take values $\{tc\_ID_1, tc\_ID_2, tc\_ID_3, \ldots\ldots tc\_ID_m\}$ in different packets. Let number of packets received per tc_id are $\{Y_1, Y_2, Y_3, \ldots\ldots Y_m\}$ respectively. Then as per Shannon criteria traffic cluster entropy is

$$H(tc\_ID) = -\sum_{i=1}^{m} p(tc\_IDi) \times log_2 p(tc\_IDi) \quad (2)$$

Here the probability of occurrence of tc_ID i.e. $P(tc\_ID) = \{p(tc\_ID_1), p(tc\_ID_2), \ldots\ldots p(tc\_ID_m)\}$ is computed as $p(tc\_IDi) = \frac{Yi}{S}$ where $S = \sum_{i=1}^{m} Yi$

Eq. (2) is used to compute 16-bit traffic cluster entropy $H(tc16\_ID)$ and 24-bit traffic cluster entropy $H(tc24\_ID)$ by finding 16-bit and 24-bit traffic clusters respectively.

In our approach, the packets destined to web server W_s are monitored at the point of presence PoPs of the protected ISP. PoPs of the ISP provide access of the Internet to its customers as well as are used for peering between ISPs. Packets are monitored in a short sized time window $\{t - \Delta, t\}$ to minimize memory overheads. Here Δ seconds is the size of time window. At time t,

the monitoring process yields packets arrival distribution of src_IP and tc_id. Then the probability of occurrence of each src_IP and tc_ID i.e. $P(src\_IP)$ and $P(tc\_ID)$ are respectively computed. In the next step source address entropy $H(src\_IP)$ and traffic cluster entropies $H(tc16_ID)$and $H(tc24\_ID)$ are computed for the time window $\{t-\Delta, t\}$ as per flowchart in figure 3. The computed entropies with total number of packets are sent by every PoP to the PoP which connects web server to the protected ISP. Here cumulative source address and traffic cluster entropies are computed as per equation 3 and 4 given below.

$$Hs(src\_IP) = (1/Sf)\sum_{i=1}^{N} Si\,(Hi(src\_IP) - \log(Si)) + \log(Sf) \qquad (3)$$

$$Hs(tc\_ID) = (1/Sf)\sum_{i=1}^{N} Si\,(Hi(tc\_ID) - \log(Si)) + \log(Sf) \qquad (4)$$

If there is no significant increase in $H_S(src\_IP)$ as well as $H_S(tc16\_ID)$ and $H_S(tc24\_ID)$, it signifies legitimate traffic as during normal event number of traffic sources and network domains do not vary much. But during FE number of traffic sources increases however there less variation in network domains. So a significant increase in $H_S(src\_IP)$ but minor variations in $H_S(tc16\_ID)$ and $H_S(tc24\_ID)$ are the signs of FE. But if there is appreciable increase in $H_S(src\_IP)$ as well as in $H_S(tc16\_ID)$ and $H_S(tc24\_ID)$, it means DDoS attack has happened because a large number of zombies send traffic from different parts of the Internet belonging to different network domains. The flowchart for detection of attack is given in figure 4.

4. FRAMEWORK

The system architecture of the proposed approach is given in the figure 5. Three ISP are shown and ISP_1 is the protected ISP domain. ISPs contain many PoPs. These PoPs actually consist of interconnected edge and core routers [8]. PoPs are connected to customer domains via edge routers and are attached with each other through high bandwidth links between their core routers. Moreover ISPs are joined with each through peering via their PoPs [8] . So these PoPs are entry and exit points of the ISPs. The legitimate and attack traffic from ISP_1, ISP_2, and ISP_3 are directed towards web server. Some of the customer domains have attack zombies as per figure 5. So customer domains generate legitimate, attack, or legitimate and attack traffic towards the web server. Through peering points legitimate and attack traffic from other ISPs enter protected ISP_1. The protected ISP_1 in the distributed framework shown in figure clearly indicates that at all the PoPs, we run traffic monitoring module, which not only separates source addresses from incoming packets but also classifies them into 16-bit and 24-bit traffic clusters. A time series analysis of this traffic is carried out at each PoP and computed entropy with total count of packets are sent to PoP connected to the server. Here anomaly based detection module runs which checks for presence of attacks using cumulative entropy computed using equation 3 and 4. The steps followed in distributed framework are given below:-

Step 1. Source IP address is detached from the incoming packet destined to protected web server at all the PoPs except at PoP Ps.

Step 2. Classification into 16-bit and 24-bit cluster is done at each PoP by using bit-wise AND operation of each source IP address with 255.255.0.0 and 255.255.255.0 respectively.

Step 3. A count is maintained for each source IP and 16-bit, 24-bit cluster in a time window $\{t-\Delta, t\}$.

Step 4. At the end of Δ seconds, source IP $H_i(src\_IP)$, 16-bit traffic cluster $H_i(tc16_ID)$and 24-bittraffic cluster $H_i(tc24\_ID)$ entropies are computed using equation 1 and 2 by all the PoPs where i=1 to N. Here N is number of PoPs.

Step 5. The computed entropies in step 4 are sent by all the PoPs to PoP Ps with sum S_i of all the packets received at respective PoP where i=1 to N.

Step 6. At PoP Ps cumulative source IP $Hs(src\_IP)$, 16-bit traffic cluster $Hs(tc16_ID)$and 24-bit traffic cluster $Hs(tc24\_ID)$ entropies are computed using equation 3 and 4.

Step 7. Source IP $Hs(src\_IP)$, 16-bit traffic cluster $Hs(tc16_ID)$and 24-bit traffic cluster $Hs(tc24\_ID)$ entropies computed in step 6 are compared with baseline respective entropies. The detection procedure flags normal, flash event, and DDoS attack.

Step 8. In case DDoS attack is detected in step 7 then either 16-bit traffic cluster or 24-bit traffic cluster is selected as anomalous cluster for attack source characterization depending upon their entropy variation from threshold.

Step 9. Anomalous cluster is analysed to find new clusters which have not appeared earlier before detection of DDoS attack as they contain source IP of all zombies which are used for attack.

Step 10. A packet having information of all abnormal traffic clusters is made by PoP Ps which is communicated to all the PoPs which share the multicast group with PoP Ps.

Step 11. All the PoPs detach information of all abnormal traffic clusters from the packet communicated in step 10 and store the same in filter database as attack signatures.

Step 12. Each packet destined to protected web server is allowed to pass only after comparing it with attack signatures stored in filter database.

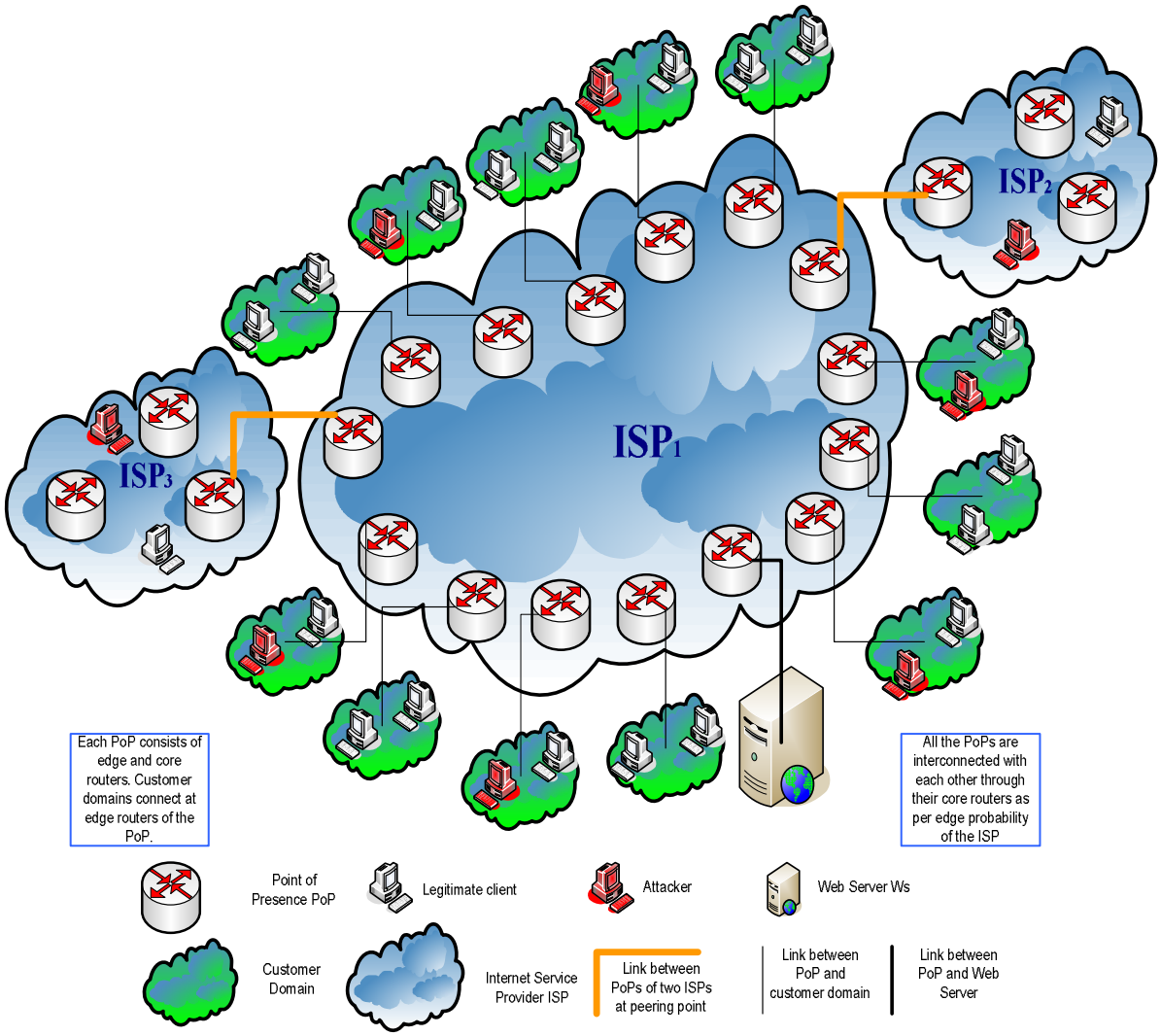


FIGURE 1: System Architecture

Now the traffic reaching at PoP Ps drops significantly as attack traffic is getting filtered at PoPs only. As the attack traffic is dropped at PoPs only so is also does not consume expensive inner bandwidth of protected ISP. The innocent traffic cluster which were mixed by attackers in a crafty manner so as to hide their zombies also do suffer as only those traffic clusters are punished which have attack zombies. Hence collateral damage is also minimum in our approach. It is worth mention here that the different operations of the approach are carried at different points and hence there is no single computational point which can be attacked by the attacker.

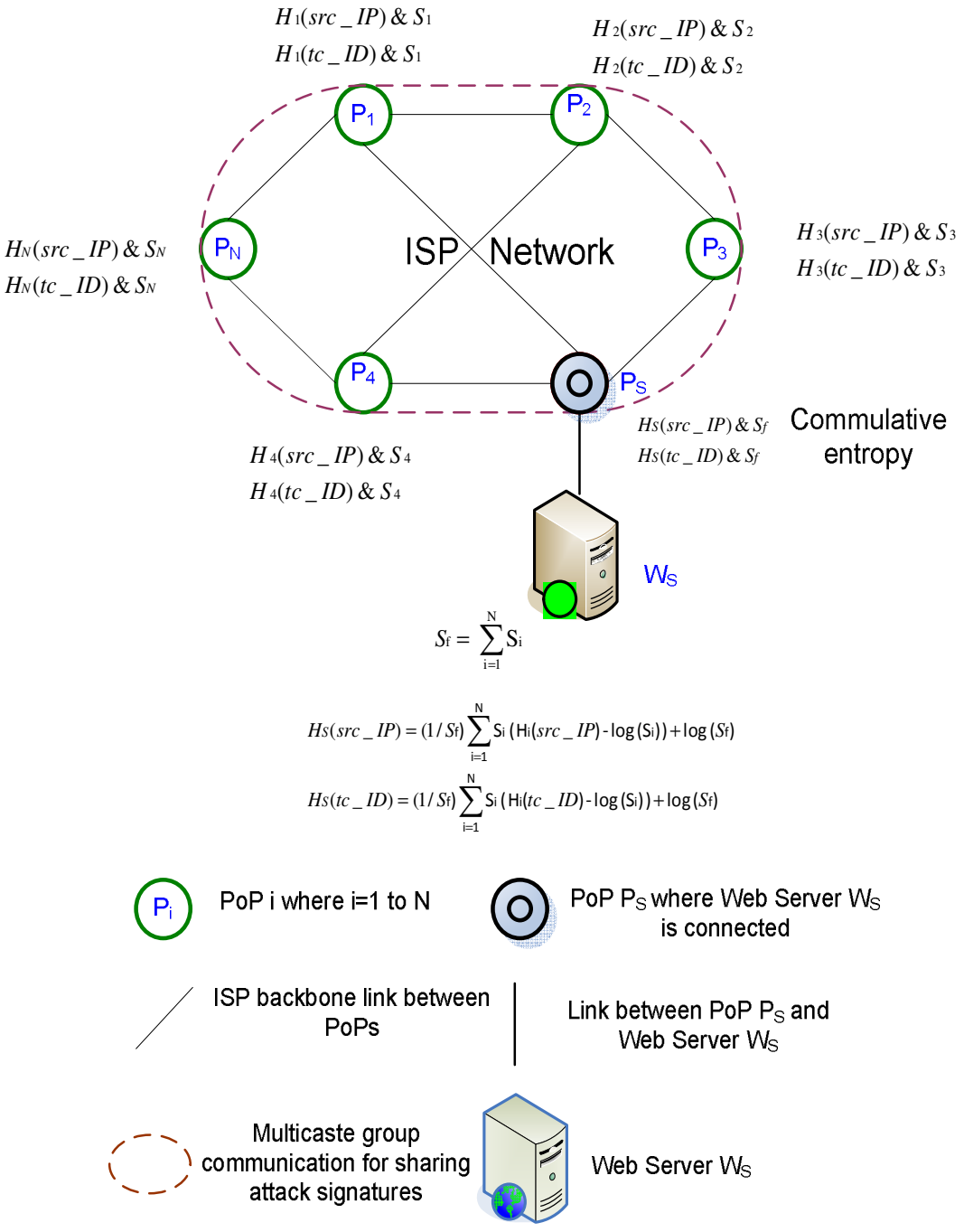


FIGURE 2: Distributed Framework

5. CONCLUSION AND FUTURE WORK

A distributed rather than centralized approach in ISP domain is the only pragmatic solution available against DDoS attacks as centralized approach suffers from single point failure bottleneck. Many defence schemes have used entropy but traffic cluster entropy combined with source address entropy is used to detect volume as well most of other intelligently crafted DDoS attacks. The proposed defence framework is comprehensive as it detects wide range of attacks, characterize attack sources and filter attack traffic. The computational burden is also distributed in such way as if amassed traffic is analysed at single point.

The future work of the paper is as below: -
An evaluation of traffic cluster entropy approach using NS-2 simulation test bed.
Implementation of distributed framework in NS-2 is there in line of sight.

6. REFERENCES

[1] M. Sachdeva, G. Singh, K. Kumar and K. Singh. "DoS Incidents and their impact: A review." The International Arab Journal of Information Technology IAJIT, ISSN: 1683-3198, Vol. 7, No. 1, January 2010, pp. 14-22.

[2] C. Douligeris and A. Mitrokotsa. "DDoS attacks and defense mechanisms: classification and state-of-the-art." Computer Networks, Vol. 44, No. 5, pp. 643–666, April 2004.

[3] J. Mirkovic and P. Reiher. "A Taxonomy of DDoS Attack and DDoS Defense Mechanisms." ACM SIGCOMM Computer Communications Review, Volume 34, No. 2, pp. 39-53, April, 2004.

[4] T. Peng, C. Leckie, and K. Ramamohanarao. "Survey of Network-Based Defense Mechanisms Countering the DoS and DDoS Problems." ACM Computing Surveys, Vol. 39, No. 1, Article 3, April 2007.

[5] M. Sachdeva, G. Singh, K. Kumar and K. Singh. "A Comprehensive Survey of Distributed Defense Techniques against DDoS attacks." International Journal of Computer Science and Network Security (IJCSNS), ISSN: 1738-7906, VOL.9 No.12, December 2009, pp. 7-15.

[6] J. Mirkovic. "D-WARD: Source-End Defense Against Distributed Denial-of-service Attacks." Ph.D. Thesis, University of California, Los Angeles, 2003

[7] K. Kumar, R.C. Joshi, and K. Singh. "An ISP Level Distributed approach to detect DDoS Attacks." Innovative Algorithms and Techniques in Automation, Industrial Electronics and Telecommunications, ISBN: 978-1-4020-6265-0 (Print) 978-1-4020-6266-7 (Online), Springer Netherlands, DOI 10.1007/978-1-4020-6266-7, Pages 235-240, September 04, 2007.

[8] K. Kumar. "Protection from Distributed Denial of Service (DDoS) Attacks in ISP Domain." Ph.D. Thesis, Indian Institute of Technology, Roorkee, India, 2007.

[9] A. D. Keromytis, V. Misra, and D. Rubenstein. "SOS: An Architecture For Mitigating DDoS Attacks." IEEE Journal on Selected Areas in Communication, Vol. 22, No.1, pp. 176-188, 2004.

[10] C. Papadopoulos, R. Lindell, J. Mehringer, A. Hussain, and R. Govindan. "CROSSACK: Coordinated Suppression of Simultaneous Attacks", Proceedings of DISCEX, pp. 2-13, 2003.

[11] X. Yang, D. Wetherall, and T. Anderson. "A DoS-limiting network architecture", Proceedings of ACM SIGCOMM, pp. 241-252, 2005.

[12] W. Shi, Y. Xiang and W. Zhou. "Distributed Defense Against Distributed Denial-of-Service Attacks", Proceedings of ICA3PP 2005, LNCS 3719, pp. 357-362,2005.

[13] G. Oikonomou, J. Mirkovic, P. Reiher, and M. Robinson. " A Framework for a Collaborative DDoS Defense", Proceedings of the 22nd Annual Computer Security Applications Conference, pp. 33-42, 2006.

[14] M. Robinson, J. Mirkovic, M. Schnaider, S Michel, and P. Reiher,"Challenges and principles of DDoS defense," ACM SIGCOMM, 2003.

[15] M. Caesar and J. Rexford. "BGP routing policies in ISP networks,"

[16] U. K. Tupakula and V. Varadharajan. "A controller agent model to counteract DoS attacks in multiple domains", Proceedings of Integrated Network Management, IFIP/IEEE Eighth International Symposium. pp.113-116, 2003

[17] S. Chen and Q. Song. "Perimeter-Based Defense against High Bandwidth DDoS Attacks." IEEE Transactions on Parallel and Distributed Systems, Vol. 16, No. 6, pp. 526-537, June 2005.

[18] G. Carl, G. Kesidis, R. R. Brooks, and S. Rai. "Denial-of-Service Attack - Detection Techniques." IEEE Internet Computing, Vol. 10, No. 1, pp. 82-89, Feb. 2006.

[19] C. E. Shannon and W. Weaver. The Mathematical Theory of Communication. University of Illinois Press, 1963.

[20] L. Feinstein, D. Schnackenberg, R. Balpuari, and D. Kindred. "Statistical Approaches to DDoS Attack Detection and Response" ,In Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX'03), Vol. 1, pp. 303-314, 2003.

Development of an Efficient Computing Multilingualism Model for Diacritical Marks in Arabic and Hindi

Abu Sarwar Zamani *sarwar_zamani@yahoo.com*
Lecturer/College of Science
Shaqra University
Riyadh, Kingdom of Saudi Arabia

Dr. Nassir Al Arifi *nalarifi@ksu.edu.sa*
Professor, Dean/College of Science
Shaqra University
Riyadh, Kingdom of Saudi Arabia

Md. Mobin Akhtar *jmi.mobin@gmail.com*
Lecturer/College of Science
Shaqra University
Riyadh, Kingdom of Saudi Arabia

Abstract

Language competence is a cognitive property of the individual speaker. There is a wide gap between commonly voiced representations of language, person, and place and actual practices of language use, identity assertion, and spatial occupation. It is noted that the one can focus on resolving related outstanding standardization issues in support of localization and multilingual requirements. This paper investigates the Diacritical marks and various typographic rules in Hindi and Arabic which are complex in multilingual documents. A computing Multilingualism model is developed which proposed a solution to the problem of position of Diacritical Marks in Multilingual documents. The developed model is found to be an efficient tool for solving the problem of positioning diacritical marks for multilingual fonts in True Type as well as Open Type format.

Keywords: Multilingualism, Diacritical Marks, Computing, Typographic.

1. INTRODUCTION

Language is a defining feature of human civilization and many languages and scripts have come into existence that is used by people around the world. Language plays unique role in capturing the breadth of human diversity. We are constantly amazed by the variety of human thought, culture, society and literature expressed in many thousands of languages around the world.
Literature is one of the vital components for the development of the society. With more than half of the worlds literature being published in languages other than English. Massive volumes of text in many languages are becoming available online. The documents may be created initially in digital form or could be converted from other media.
On the other hand rapid diffusion over the international computer networks of the world wide distributed document bases, the question of multilingual access and multilingual information retrieval is becoming increasingly important.
In a multilingual digital document, the principles of designing are risky by the likely conflict rules and mechanism that control each of the writing. Diacritics are an example.

Diacritical marks is a small mark added to a letter that changes its pronunciation such as (أ) hamza indicates upper of Alif, (إ) hamza indicates lower of Alif. Diacritics are often placed above the letter but they can be placed below, in or through, before or after or around a glyph.

Diacritical marks have common roles between the different languages of the world like:-

- Define playback
- Amend the phonetic value of a letter.
- Avoid ambiguity between two homographs.
- Etc.

However, Hindi is an Indo-European language spoken mainly in north, central, and western Indian. Hindi also refers to standardized register of Hindustani that was made one of the official languages of India. Hindi is written in Devnagri and has been partially purged of its Persian and Arabic vocabulary, which was replaced by words from Sanskrit.

This study focuses on to appropriate a resolution to the problem of positioning of diacritics:

We have taken some steps:
- We compare problem design of diacritical marks in the Hindi script with the design of diacritical for Arabic script.
- We spend the last part to problem of positioning diacritical marks.
- We identified strategies to solve this problem and examine their ability in the Hindi case.

2. GENERAL INFORMATION

A. History About Diacritic Sign

The first diacritics appeared in Ancient Greece and Rome, evolved and spread in subsequent European languages. While they were created to help in the pronunciation of letters and words.
Arabic is in the Semitic language group, which seems to have originated somewhere near modern Syria, Hebrew and to have spread from there through Lebanon, Israel, and Jordan down to the Arabian Peninsula. It is also cursive and written from right to left. The majority believes it has developed down writing Nebatean. Others believe it comes from Al-Musnad also known as Al Hamiri (writing of the former yemini). A small group believes that writing is a pure divine production. Until the time of Mohammed, in the 600's AD, Arabic was mainly spoken and not written. Still, there are some written records from the Arabian Peninsula from before the 600's AD. These are called Sabataean. But they are only short inscriptions in stone, not really literature. After the Islamic conquests of the late 600's AD, people soon began to speak Arabic all over the Islamic Empire, from Afghanistan to Spain, and people speak Arabic in even more places today (though not in Spain). By 1000 AD, people spoke Arabic in India. Many people began to write in Arabic. Among the first things to be written was the Quran, because the Quran played a key role in the development of Arabic script. But soon many scientific texts and medical books and math books were written in Arabic, and also stories like the Arabian Nights or the story of Aladdin. There were many Arab historians, geographers, philosophers, and poets. However, the most common solution is to add diacritical marks on the letters, often imitating the spellings of other languages [2].

Hindi is the third most widely-spoken language in the world (after English and Mandarin): an estimated 500-600 million people speak the language. A direct descendant of Sanskrit through Prakrit and Apabhramsha, Hindi belongs to the Indo-Aryan group of languages, a subset of the Indo-European family. It has been influenced and enriched by Persian, Turkish, Farsi, Arabic, Portuguese, and English. Hindi inherited its writing system from Sanskrit. Hindi can be traced back to as early as the seventh or eighth century. The dialect that has been chosen as the official language is Khariboli in the Devnagari script. Other dialects of Hindi are Brajbhasa, Bundeli, Awadhi, Marwari, Maithili and Bhojpuri.The general appearance of the Devanagari script is that of letters 'hanging from a line'. This 'line', also found in many other South Asian scripts, is actually a part of most of the letters and is drawn as the writing proceeds. The script has no capital letters.It was in the 10th century that authentic Hindi poetry took its form and since then it has been constantly modified. History of Hindi literature as a whole can be divided into four stages:

- **Adikal** (the Early Period),
- **Bhaktikal** (the Devotional Period),

- **Ritikal** (the Scholastic Period) and
- **Adhunikkal** (the Modern Period).

Adikal - The Early Period: Adikal starts from the middle of the 10th century to the beginning of the 14th century.

Bhakti Kal or the Devotional Period: Bhakti Kal or the Devotional Period stretched between the 14th and the 17th century. During this age Islamic customs were heaped upon the common people, and the Hindus were quite dejected at the effect on their culture.

Ritikal or The Scholastic Period: The poets of Ritikal or the Scholastic period can be classified into two groups on the basis of their subject: Ritibaddha (those wedded to rhetorics) and Ritimukta (free from rhetorical conventions).

Modern Hindi Literature: Modern Hindi literature has been divided into four phases; the age of Bharatendu or the Renaissance (1868-1893), Dwivedi Yug (1893-1918), Chhayavada Yug (1918-1937) and the Contemporary Period (1937 onwards).
Bharatendu Harishchandra (1849-1882) brought in a modern outlook in Hindi literature and is thus called the 'Father of Modern Hindi Literature'. Mahavir Prasad Dwivedi later took up this vision. Dwivedi was a reformist by nature and he brought in a refined style of writing in Hindi poetry, which later acquired a deeper moral tone.

Classification in Hindi
There are some kinds of Hindi Diacritic Marks which are below:

- Diacritics Above

FIGURE 1: Diacritics above in Hindi.

- Diacritics Below

FIGURE 2: Diacritics below in Hindi.

- Diacritics Through

FIGURE 3: Diacritics aesthetics in Hindi.

- Esthetics Diacritics

□□□□□ □□□□□□ , जय हिंद

FIGURE4: Diacritics aesthetics in Hindi.

- Explanatory Diacritics

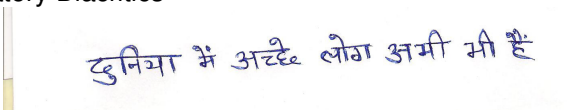


FIGURE5: Diacritics Expalantory in Hindi.

C. Classification in Arabic

Arabic Diacritics can be classified into three categories [1] :

- Language's diacritics: composed on:
  - Diacritics above
    it's a mark placed above a letter, as Fatha, Damma or Sukun.

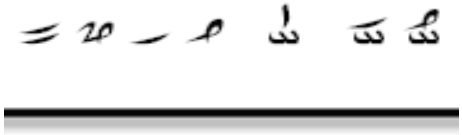


FIGURE 6: Arabic diacritics above.

  - Diacritics below

FIGURE7: Arabic diacritics below.

  - Diacritics Through

FIGURE:8: Jarrat wasl through Alef.

  - Esthetics Diacritics

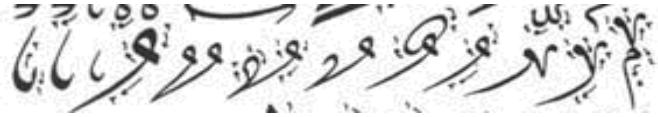


FIGURE9: Kasra and Kasrattan.

  - Explanatory Diacritics

FIGURE10: Explanatory diacritics [10].

3. DIACRITICAL MARKS WITH UNICODE

Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. The Unicode character set has the capacity to support over one million characters, and is being developed with an aim to have a single character set that supports all characters from all scripts, as well as many symbols, that are in common use around the world today or in the past. The Unicode character encoding treats alphabetic characters, ideographic characters and symbols in an equivalent manner, with the result that they can coexist in any order with equal ease. Unicode assigns to each of its character a unique numeric value and name.

There were four key original design goals for Unicode:

(i) To create a universal standard that covered all writing systems.

(ii) To use an efficient encoding that avoided mechanisms such as code page switching, shift- sequences and special states.

(iii) To use a uniform encoding width in which each character was encoded as a 16-bit value.

(iv) To create an unambiguous encoding in which any given 16-bit value always represented the same character regardless of where it occurred in the data.

However, Unicode provides other information crucial to ensure that the encoded text will be readable: the case of coded characters, their properties and their directionality letter. Unicode also defines semantic information and includes correspondence tables of breakage or conversions between Unicode and directories of other important character sets.

Bi-directional text is text containing text in both text directionalities, both right-to-left (RTL) and left-to-right (LTR). The bidirectional algorithm takes place in six steps:

- Determine the default direction of the paragraph;
- Process the Unicode characters that explicitly mark direction;
- Process numbers and the surrounding characters;
- Process neutral characters (spaces, quotation marks, etc.);
- Make use of the inherent directionality of characters;
- Reverse substrings as necessary.

4. DESIGINING, POSITIONING & MULTILINGULISM

In the time of lead type, the creation and design of letters was solely dependent on type foundries. These foundries, such as Monotype and Linotype, had all the copyrights of these fonts. Lots of concept underlies the field of design, as the balance, the rhythm, etc. The principles of design face in the case of mixture of different directions postings to change the rules of writing. It is in a somewhat similar situation when a multitude of styles in a monolingual Arabic text where the change of style indicates a title or section begins [1].

A. Language Variety Space

There are many varieties of language, text, discourse. These varieties are conditioned mainly by the functions of language in actual use. Positions in variety space define a language variety with specific forms, and specific conventions of meaning and use. Consequently, understanding and intelligent design of text and discourse, from conversation, through letters to hypertext, requires models of language functions as touchstones of quality. In Arabic, heights [1] and forms of letters vary depending on the context:

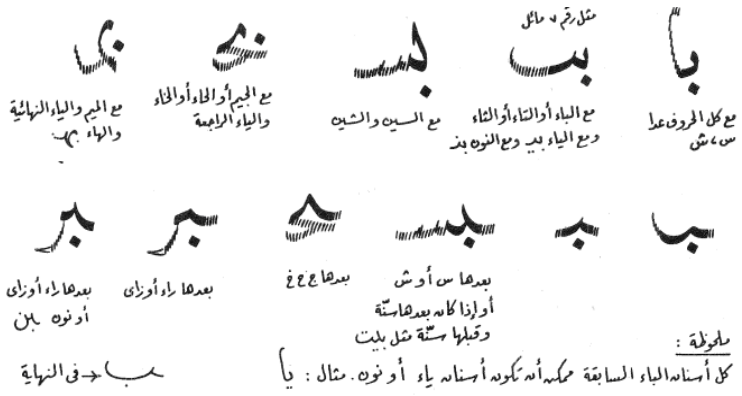


FIGURE 11: Arabic letter Beh

The spatial properties vary between Hindi and Arabic scripts. Arabic scripts start from right to left, which vary slightly depending on whether they are connected to another letter before or after them. The definition of “bold” depends, in Arabic, of style. The reduction in the density of letters is by layering or by reducing the body. Diacritics in the Thulut style, unlike the Naskh, by a Qalam, pen, different from that used for the body of letters base. The harmonization of multilingual document is therefore influenced by the multitude of scripts or styles in the same language [5].

B. **Justification of the Hindi Text**

The justification of the Hindi text makes itself while varying the space between the words and the characters, so that the line of text filled the inter-margin space. The value of the spacing varies between a minimal value and another maximal when the optimal value doesn't permit the justification of the text. The hyphenation permits to cut the word that arrives at the end of line in order to have a better visual within a text. The general appearance of the Devanagari (Hindi) script is that of letters 'hanging from a line'. This 'line', also found in many other South Asian scripts, is actually a part of most of the letters and is drawn as the writing proceeds. The script has no capital letters. Amongst its interesting features is a three-tier level of honorifics, allowing great subtlety in adjusting the level of communication to suit 'formal', 'familiar' and 'intimate' conversational contexts. Thus, the polite communicating of gratitude, etc, is an intrinsic part of the language itself and does not rely solely on separate words for 'please' and 'thank you'.

Problems related to the justification and literature of the text, especially a justification as well as literature of the kind made by processing software word processing, without correction by a human operator are potentially many. Basically there are two types of literature are used in Hindi Sculpture.

(1) Swars (□□□□ / Vowels)

DevaNagari vowels are not scattered in the 'Varnamala' (DevaNagari alphabet) but are arranged at the beginning of the alphabet.

अ आ इ ई उ ऊ ए ऐ ओ औ ऍ ऑ ऋ

These vowels were arranged according to a scheme [7]. This scheme is not completely scientific (phonetic), but definitely helpful in memorizing & reciting these vowels.

(2) Vyanjans (□□□□□□ / Consonants)

These are very logically arranged in following groups.

(i) **Sparsh:** Sparsh means touch. While speaking, along with vibrations in vocal cord and passage of air from mouth and nose, tongue and lips move. Particularly for pronouncing consonants the movements of tongue and lips are important [7]. In DevaNagari, most of the

consonants are arranged logically; depending upon the position of tongue (what it touches) and movements of lips. Like…

क to म
ka ma

(ii) **Antashth:** This is the middle set 'Antahsth' in Sanskrit means 'middle' or 'inner'. Theses are…

य र ल व
ya ra la va

(iii) **Ushm:** 'UShm' means hot! Isn't it amazing to know that terminologies developed separately? resulted in related terms- 'Friction' and 'Hot'(heat)!!! These are…

श ष स ह
sha sha‡ sa ha

C. **Justification of the Arabic Text**
In the Arabic writing, that is cursive, a word can be dilated by the kashida - specific to the Arabic writing - to cover much space [1] [8] and can be pressed by the use of the ligatures [1] [8]. It has other mechanisms of management of the Arabic line: graphic fillers (as the three points), reduction of the size of the characters, elongation of the letters, superposition of the letters, writing in the margin, etc. [1] [8]. These mechanisms influence on the measurements and the positioning of the Arabic diacritical marks [2].

5. DIACRITICS DESIGN

There are two problems in the design of Hindi Diacritics.

- They must concord with the Glyph.
- Do not cause problems with other basic glyphs;

In the Arabic case, there are aesthetic diacritics whose position depends on other diacritical marks. The interactive diacritics relationship with the mechanisms of justification requires resizing and repositioning diacritical word influenced by the effects of justification.

A. Asymmetry Problem

The balance is the stability resulting from the review of an image and a comparison with our ideas of the physical structure (such as mass, gravity, or the edges of a page). That is the arrangement of objects in a design specified according to their weight in the visual picture composition. The balance generally exists in two forms: symmetrical and asymmetrical.
The symmetrical balance occurs when the weight of a graphic composition is evenly distributed around a central axis vertical or horizontal. The symmetrical balance is also known as formal balance. The asymmetrical balance occurs when the weight of the graphic composition is not spread evenly around a central axis. The asymmetrical balance is also known as informal balance. The size of a Hindi diacritic and weight must be balanced with the glyph base with which it is used [9]. The horizontal alignment of diacritical glyph with the foundation should be such that there is balance the two views. For diacritic center symmetry with glyphs basic symmetrical, simply align the center of the bounding box of diacritic with the basic glyph [9]. If either one is asymmetrical other measures must be used. Follow, we present the main issues of design diacritics as they have been cited in [9].

1) Case of symmetrical basic Glyph

A glyph is a graphical symbol that represents a model component, such as an individual molecule. In some cases an attribute of the glyph is a function of the model component that it represents.
One solution is to align the optical center of the letter with the mathematical center of space. The optical center is estimated by the center of the contour.

2) Case of Asymmetrical base Glyph

In this case, the diacritic exchange up connection following the basic glyph. The optical alignment is not always used and other solutions are offered by new technologies such as OpenType and Graphite.

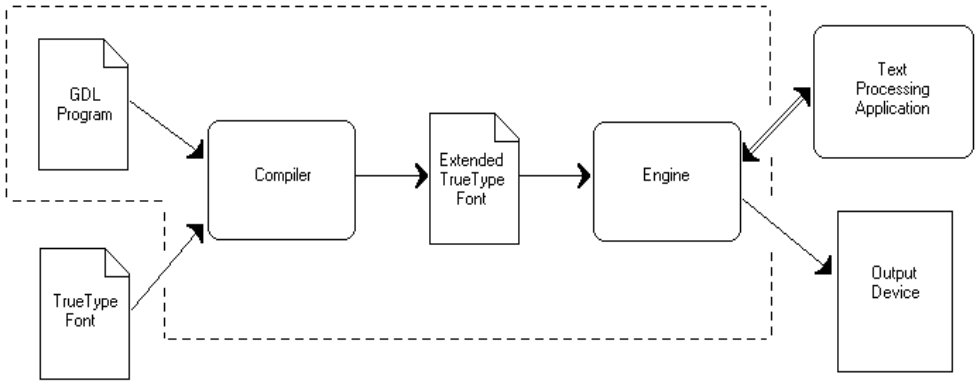


FIGURE 12: Graphite System Architecture.

Description: This system can be used to create "smart fonts" capable of displaying writing systems with various complex behaviors, such as:

- A rule-based programming language Graphite Description Language (GDL) that can be used to describe the behavior of a writing system
- A compiler for that language
- A rendering engine that can serve as the back end of a text processing application.

Graphite renders TrueType fonts that have been extended by means of compiling a GDL program [2].

B. Multiple Diacritics

Diacritics could cause multiple problems with the baseline or with other glyphs. Different techniques are used to solving this problem including: draw a glyph gathering all the diacritics multiple, etc.

C. Particular Issues to Arabic

Arabic diacritics role is to fill the void, white space, in the word that there are specific diacritical marks, for aesthetics. There are three mechanisms for creating void in the Arabic word: kashida, extension glyphs and the interconnection between glyphs. In each case, the void is filled in two steps:

- The first, by resizing the Fatha in proportionality with the white;
- The second, by placing the aesthetics' and explanatory diacritics.

Diacritical marks lead, according to the language's function, to repeat the characteristics common to many of the glyphs.
The concept of symmetry in Arabic design is related to the line writing where the extensions are to balance the masses of other glyphs.

Arabic diacritics have a relationship with the mechanisms of justification. The diacritical marks are cosmetic compared to other signs respecting fill the void and not obscure the gray.

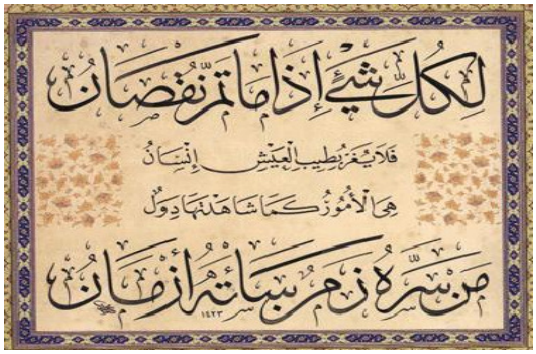


FIGURE 13: Arabic Diacritics Role.

6. NEW TECHNOLOGIES & DIACRITICAL POSITIONING

We are studying the three font's formats: TrueType, OpenType and Graphite.

1) True Type: TrueType fonts offer the highest possible quality on computer screens and printers, and include a range of features which make them easy to use. TrueType is an outline font standard originally developed by Apple Computer in the late 1980s as a competitor to Adobe's Type 1 fonts used in PostScript. The primary strength of TrueType was originally that it offered font developers a high degree of control over precisely how their fonts are displayed, right down to particular pixels, at various font heights [4].
2) Open Type: The Open Type font format is an extension of the TrueType font format, adding support for PostScript font data.
   Open Type fonts and the operating system services which support Open Type fonts provide users with a simple way to install and use fonts, whether the fonts contain TrueType outlines or CFF (PostScript) outlines [12].
   The Open Type font format addresses the following goals:

- broader multi-platform support
- better support for international character sets
- better protection for font data
- smaller file sizes to make font distribution more efficient
- Imported internet and PDF (Portable Document Format) publishing.

GPOS table manages the positioning of glyphs. We can put any diacritic on any glyph basic threw it [4]. Each diacritic has a base. Diacritics are divided into several classes according to their behavior. Each basic glyph as attachment points that diacritic class.

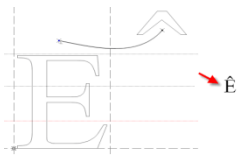


FIGURE 14:Diacritics Position.

A. Attachment and Cluster in Graphite

The positioning of glyphs is done by two simple operations: moving and kerning, a simple tool: the points of attachment. If two glyphs "A" and "B" are attached, one-by-example "B" is

attached to "A" and "A" is said base of "B". Another glyph "C" in turn can be attached to either "A" or "B", etc. [12].

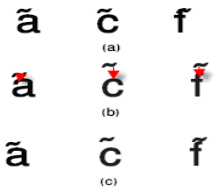


FIGURE 15: Diacritics attachment points.

The FIGURE 15 demonstrates the usefulness of attachment points. As shown in FIGURE 15 (a), a record of diacritics with a "not smart fonts" seems correct when they are attached to a tiny symmetrical centered as "a", but if not symmetric the diacritic is not centered correctly and comes into collision with the upper half of the glyph, or both. For Graphite font, stain is different: FIGURE 15 (b) shows the commitment indicated by small dots and arrows, and FIGURE 15 (c) shows the results with the correct record. The mechanism of base resolves the multiple diacritics problem, when the first diacritic is attached to a glyph base; it in turn is the basis of the following diacritic.
The basic glyph and diacritic form a cluster. Graphite includes the ability to calculate metrics cluster or sub-cluster glyph individual for use in operations positioning [12].

FIGURE 16: Multiple diacritics attachment points.

FIGURE 17: Examples of Arabic fonts.

B. Diacritics Positioning System

To place one or more diacritical marks relative to the base glyph, this system use a diacritic's bounding box and the base glyph's bounding box, in association with diacritic place data stored in the system[11]. The position data enables the diacritic positioning system to call associated functions that place multiple diacritics above and/or below a single base character without interfering with one another, e.g. to stack the diacritics. In addition, the information about the diacritic characters can be employed to prevent interference between a diacritic and the base character in special circumstances [11].

a. The Algorithm of Diacritic Position system

Start

Step – 1: Glyph Received

Step – 2: If Diacritical Then Step – 3 Else Step - 7

Step – 3: If Special base Then Step – 4 Else Step - 5

Step – 4: Retrieve diacritical mark and GOTO Step - 6

Step – 5: Retrieve base char mark orientation

Step – 6: Call H function

Step – 7: Call V function

Step – 8: Draw glyph

Step – 9: Continue……...

End

b. Description

When the system receives the information that the mark is to be placed over the base character, he looks up the orientation for this mark in the table that is stored in memory. This table [11] lists each diacritic by its name or their Unicode value. Based on this information in this step, the system calls a pair of functions H and V for properly positioning mark.

c. Commentary

Graphite and OpenType font formats have the advanced features to treat Arabic script. For this reason, we limit this study to the system for positioning diacritical mark in TrueType font format. In the Arabic script, the position and dimension of diacritical mark Fatha and Fathattan are related to form of base glyph and followed base glyph. So, to extend a system which operates under the same architecture as the diacritics positioning system three things to take into account:

- The functions H and V must have the ability to calculate the horizontal and vertical position of diacritic glyph relative to the base glyph and followed base glyph.
- The system must be able to substitute the diacritical mark if an extension takes place.

7. CONCLUSION

Most of the fonts used to write Arabic do not have a deep tables and technologies of different formats, but we believe that the resolution of problems of diacritical in the multilingual digital document affects a layout engines. These problems have link with the problems of design of Arabic basic letters as the superposition of letters, the reduction of body and ligatures.

8. REFFERENCES

[1] Vlad Atansiu, “Le phénomène calligraphique à l’époque du sultanat mamluk”, PhD Thesis, Paris, 2003.

[2] http://a1.esa-angers.educagri.fr/informa/, February 2009.

[3] Mohamed Hssini, Azzeddine Lazrek and Mohamed Jamal Benatia, “*Diacritical signs in Arabic e-document*”, CSPA’08, The 4th International Conference on Computer Science Practice in Arabic, Doha, Qatar, April 1-4, 2008 (in Arabic).

[4] R. Nicole, “Graphite Application Programmer’s Guide”, http://www.sil.org/.

[5] Mohamed Hssini and Azzeddine Lazrek," Design and Computer Multilingualism: case of Diacritical Marks, Department of Computer Science, Faculty of Sciences, University Cadi Ayyad - Marrakech, Morocco.

[6] Yannis Haralambus, "Fontes et codage", O'Reilly, Paris, 2004.

[7] J. C. Wells, "Orthographic diacritics and multilingual computing", Language problems & language planning ISSN, 2000, vol. 24, n° 3, pp. 249-272.

[8] Mohamed Jamal Eddine Benatia, Mohamed Elyaakoubi, Azzeddine Lazrek, "*Arabic text justification"*, TUGboat, Volume 27, Number 2, pp. 137-146, 2006.

[9] J. Victor Gaultney, "Problems of diacritic design for Latin script text faces", http://www.sil.org/, December 2008.

[10] H. Albaghdadi, "Korassat alkhat", Dar Alqalam, Beirut, 1980.

[11] Chapman, Christopher J., "*D*iacritic positioning system for digital typography*"*, http://www.freepatentsonline.com/WO2008018977.html, January 2009.

[12] http://www.typographie.org/, January 2009.

Black Box Backup System

Iyad Aldasouqi *iyad@rss.gov.jo*
Information Technology Center
Royal Scientific Society
Amman, 11941, Jordan

Arafat Awajan *awajan@psut.edu.jo*
The King Hussein School for Information Technology
Princess Sumaya University for Technology
Amman, 11941, Jordan

Abstract

Modern organizations from different sizes (Small, , Medium and Large) consider information as one of the most important of their assets that need to be secured against increasing number of threats. The importance of the information comes from its impacts on the main tasks performed by the organization. The evolution of Information Technology and Information Systems is changing permanently the characteristics and the components of such systems and the ways needed to protect them against any security risk.

Periodic data backup is a system administration task that has changed as new technologies have altered the fundamental structure of networks. These changes encourage rethinking of modern backup strategies and techniques. In addition, standard backup programs and specialized tools are often needed.

This paper provides an overview of issues to be considered for a long term, stable and secure backup system. A new approach (Hardware) called Black Box backup system is proposed based on current risk management plans and procedures used mainly in the aerospace industry.

Keywords: Black Box, Backup, Network Backup System, Mirroring, RAID

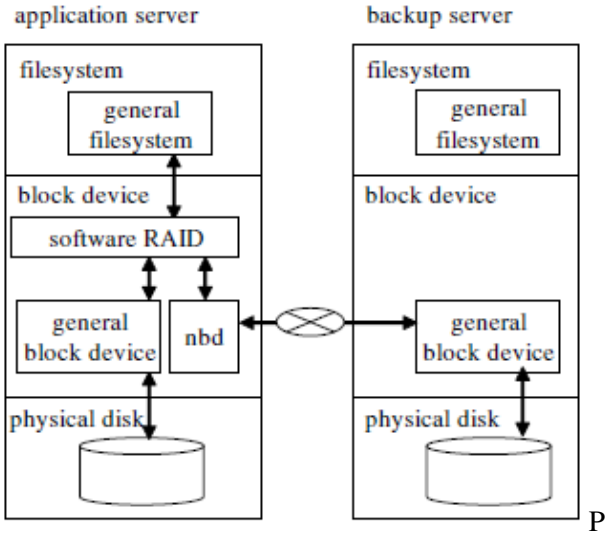
1. INTRODUCTION

Every organization tries to deliver value from information Technology (IT) while managing an increasingly complex range of IT-related risks. The best practice can help to avoid collisions, and reduce the occurrence of major IT risks, such as: project failures, security breaches, system crashes, and failures by service providers to meet the upon greed requirements. In addition, today's attacks aren't as likely threats never seen; therefore technologies which will be able to protect enterprise networks against these kinds of attacks should be chosen carefully and designed properly.

The purpose of this paper is to find and contribute in a stable secured back up system which can be resistant to any hazards, catastrophes, crises and natural disasters. The paper includes discussions on the features and limitations of the native backup and recovery available programs. It is hoped that the information presented here can help administrators consider tradeoffs in cost, performance, and reliability for different types of solutions. Furthermore, most of backing up techniques are utilizing for certain functions, such as the using of some techniques for backing up file-system e.g. In the snap shot backup, the technique which was LVM (Logical Volume Manager) and utilize the filesystem [1, 2].

Furthermore, backing up techniques have been used for the periodic backup, but cannot mirror data in real-time. In a real-time mirroring, RAID [3] has been frequently used to mirror data on a

local disk. The mirroring on a network has been often implemented as a function of the clustering system [4, 5]. Moreover, Network Block Device (NBD) makes a block device available on the network. The method to combine NBD with software RAID makes it possible to mirror on the network in real-time (Figure 1).



P

FIGURE 1: Conventional backup system with mirroring on device layer [20].

2. DATA BACKUPS SYSTEMS

Data backup is a necessary requirement for every organization. The well-known traditional reasons are system crash and disasters such as a flood and fire strike networks. Users may inadvertently delete files or overwrite existing files, hackers or disgruntled employees do the same purposely. Disk drives are inherently fragile devices. As a result, of that files become corrupted by bad disk sectors, magnetic fields, and improper system shutdown. In addition to the traditional threats, there are new threats such as thieves steal laptops, and the data contained on them. The threat posed by modern worms dwarfs those of older worms [6], and they are able to compromise every vulnerable machine on the Internet faster than any manual response can prevent [7].

Nowadays, organizations' computer and information systems are the most important assets and organizations depend on them more than ever; therefore, loss of data is more expensive than ever in terms of efforts spent and downtime and with increasing threats and increasing costs, backups are more crucial than ever.

Therefore, developing a backup strategy is needed for specific network, data, and organizational objectives (different strategies for different purposes). A survey of factors to consider is presented in [8]. It provides an excellent planning tool for developing backup strategies.

2.1. Properties of Good Backups

In a well-managed network, backup operations are performed on a regular predefined basis. Additionally, a good recovery system is essential. During both normal use and recovery, backup operations should be transparent to users. Backup operations should be automatic and not be the responsibility of users. Instead, a system administrator should centrally manage backup and recovery operations. Since backups are of high priority, they should be managed by a person who understands their importance, rather than a new hired or intern one.

Finally, the scale of modern networks is beyond what can be manually managed. Good management requires human intelligence supported by automated information gathering and management.

2.2. Methods Available for Data Backup and Recovery

Backup solutions can be divided into two major categories. The first category includes the native backup and recovery programs which backup volume data from file servers, also it is providing a backup application that is needed by data to achieve the goal of backup process.

The second category relies on file system. These categories take data from a file server in the same way that users access their data, which is very helpful in 24/7 environment.

2.3. Information storage strategies.

Having a data-backup-recovery strategy requires answering two questions:

1. How quickly must you recover the data before your business experiences serious setbacks?
2. How much are you willing to pay to implement a data-backup plan?

An organizations data should be backed up regularly on some type of removable medium, and then delivered to another location for protection purpose.

Other questions may be asked:

- How safe is the information in your computer?
- If a fire, flood, earthquake or even sabotage struck your office, would your electronic data survive?
- If you could access your data, how long would it take to get your information system up and running again?

To answer above questions, and to start designing such a strategy, we have to take into consideration two important issues:

- Downtime: How quickly must you recover the information before your business experiences serious setbacks?

- Cost: How much are you willing to pay to implement a data-backup plan?

While the questions are often difficult to answer, it is clear that large and small organizations need to prepare disaster-preparation strategies. We can conclude that faster recovery times equal lower downtime costs. However, strategies that speed recovery also can be expensive. This paper focuses on one area of disaster preparedness which is data backup.

2.4. Back up Practice / Cost

Storing the backed-up data in a secure place is called vaulting process, in which organizations copy computer files regularly on removable medium (Magnetic tape, CD or hard disk), and then delivered them to an off-site location for safekeeping. The timing of backups can vary depending on the organization's needs.

The off-site location called the renting a bank, with cost depending on the amount of space needed and its location. Another but the more expensive strategy is the redundant computer hardware, where if one component fails, a backup device keep the system running. An example of this approach is to use a technology known as redundant array of independent disks (RAID). There are many kinds of RAID systems, all of them designed to provide different levels of error recovery and fault tolerance.

One RAID choice is disk mirroring—a process in which data are simultaneously duplicated on one or more disks within the same system. The only additional hardware you need to set up a mirroring system is an additional disk drive that is the same size as your current drive. A typical 60-Gb drive, for example, costs about $200. For additional protection you also might consider buying a separate disk controller card for the new drive for around $80. The dual disk drive/controller card option results in a special type of mirroring protection called disk duplexing [9].

Another choice is disk striping; while there are many variations of striping, the most common is to set up an array of at least three and usually five disk drives. Disk striping does not store redundant data across the disk array; rather, it uses a system of parity checks—or hash totals—to rebuild lost data should one drive in the array fail. If you are running Windows NT Server, Windows 2000 Server or Novell NetWare, your computer is capable of handling disk mirroring and striping.[9]

If you are running your business on a single computer, it's easy to add a RAID configuration. However, if your business runs on two or more networked computers, a network administrator will be the responsible person to maintain the system.

3. OUR CONTRIBUTION

The idea of using Black Box Backup System (BBBS) came from Black box (as in figure 2) system used in aircraft. It is a generic term used to describe the computerized flight data recorders carried by modern commercial aircraft. This device is typically used in conjunction with a second black box known as the Cockpit Voice Recorder (CVR), which documents radio transmissions and sounds in the cockpit, such as the pilots' voices and engine noises. In the event of a mishap, the information stored in these black boxes can be used to help determining the cause of the accident.

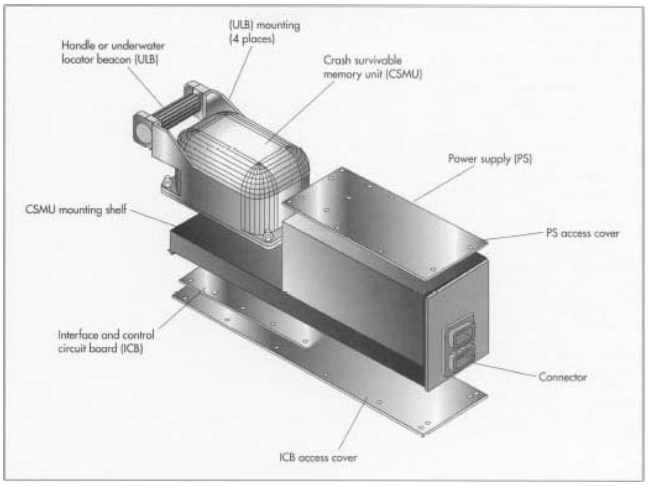


FIGURE 2: Black box model.

The proposed BBBS can be used in IT risk management plan, as a real time back up system which can be connected to a SCASI adapter or to Ethernet HUB, and it can depend on a separate processor (come up with processor) or used the server processor the most important feature is the possibility to use it against natural disasters (Like Earthquakes), water flood resistance and anti-fire. Therefore, we can be sure that availability, survivability and security issues are achieved, which can help saving all kinds of organization information assets with different sizes with reasonable cost.

3.1. Functionality

Storing data may uses several techniques such as a sniffer technique via the Ethernet adapter or regular backup technique with differential option and small time differences between backing up processes, or add sensing software to trigger the black box to start backup process. Therefore, when a crisis occurred (fire, earthquake or attack), the whole data can be recovered. This unit should be shielded and fixed to the ground. So this black box can be considered as a hybrid implementation technique; since it might be used for security and safety purposes.

Furthermore, early warning system can be used as an early warning system such as “Earthquake Monitoring System Using Ranger Seismometer Sensor” as in figure 3[10], which can forecast the occurrence of natural disasters 20 seconds in prior; so a small software program similar to the one used in Stream Processing Environmental Applications in Jordan Valley [11], that can receive the output of [10] and start shutdown the system before the disaster occurred as in figure 4.

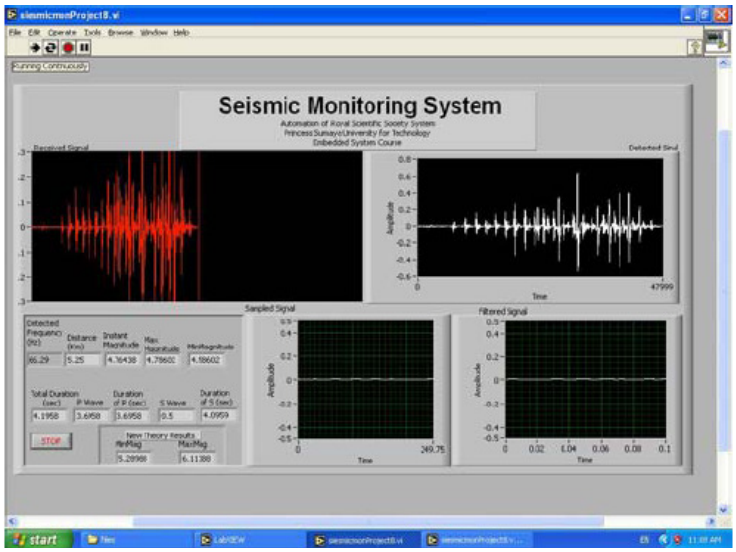


FIGURE 3: Early warning system

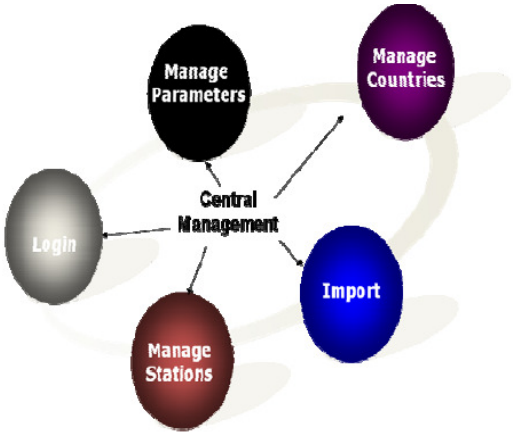


FIGURE 4: Stream processing architecture

In addition, the system used in Cluster-based scalable network services [4], can be send a statues update of the log files in addition to the control function to turn off services and shutdown the servers since it can talk with the hardware.

3.2. System block diagram

Figure 5 shows a general overview of the BBBS diagram and its position in the overall information security system.

The interior proposed black box consists of the eight following components [figure 6]:

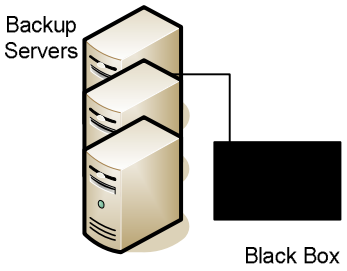


FIGURE 5: Black-Box overview.

a) Computer Interface Board (Data Acquisitions Card): it is a translator between the computer and machines or hardware as National Instruments products [19].
b) Audio Compressor Board: it is an audio recording system that is used to record any voice or movements around BBBS such as operators' discussion or thieves' talks.
c) High Temperature Insulation: it is a protection layer that can protect BBBS in case of fire occurred.
d) Stainless Steel Shell: it is another layer of protection to secure BBBS internal component from damage and tampering.
e) Under construction Locator Beacon: it is a device that gives a specific sound (like a buzzer or siren) to tell the rescue team about BBBS in case of a disaster occurred.
f) Stacked Memory Boards: it is the place where the data saved.
g) Memory Interface Cable: it is a cable that is used as a media between the storage area and the computer.
h) Acquisition Processor Board: it is a board similar to computer's motherboard that has different kind of interfaces to connect all BBBS components together.

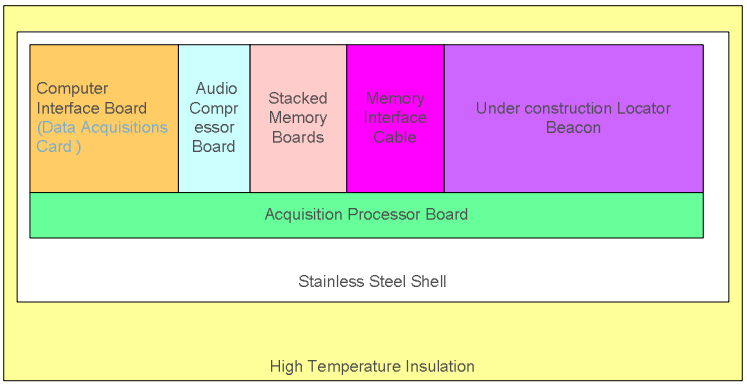


FIGURE 6: Suggested Black box block diagram

4. RELATED WORK

Backup and recovery systems are hot topics covered by many researches in the past and represent one of the most important security issues in the present. Theses research are based on different approaches varying from performance oriented, security oriented to accessibility oriented research. Our approach considers mainly the disaster recovery factor in addition to all other prospective.

A more comprehensive description of backup system issues and examples are in as in [12, 13]. Amanda (Advanced Maryland Automated Network Disk Archiver) is an early example of freely available backup management software [14, 15]. It uses a combination of full and incremental backups to concurrently backup networked clients to a single designated backup server and uses configuration files to determine the type of backup to perform. Multiple commercial systems [16,17, 18] now provide Amanda-like functionality; however, none deal gracefully with wireless hosts.

RAID [3] can protect systems against the failure of individual components. It provides no protection against unintentional/unauthorized modification of data, nor from catastrophic failure. Traditional RAID systems are impractical to field for mobile systems.

5. CONCLUSION AND FUTURE WORK

Most of backup strategies so far assume company data are located in easily identifiable and accessible places, but more of data are being stored on personal desktops, laptops and personal digital assistants, therefore dispersing important information into disparate isolated pockets. There are three basic users of data to be considered in this work: the central office worker, remote office staff member and the traveler. Office users and remote office should back up their files to local storage devices such as a CD, Zip or tape drives as well as to a network frequently.

The most important data are the organization databases and applications, which should be backed up almost every day or every moment; since it is the most important asset. Furthermore theses data should be located in at least in two different locations, and should be proceed and treated carefully using up to date secured techniques and technologies.

The strategy discussed in this paper is based on the use of a “Black box” model for conducting the backup activities and improving the disaster recovery planning in the organization. From that prospective, and since the components of the original black box were build based on IT technology, we suggest to implement this strategy in to save all kind of data and transactions including human being data.

As a future work, we suggest to start implementing and testing this system at Small Medium Enterprise (SME) organizations and to simulate via accepted estimating of the results comparing with traditional systems.

6. REFERENCES

[1] M. Rosenblum and J. K. Ousterhout. The design and implementation of a log-structured file system. ACM Trans. Comput. Syst., pages 26–52, 1992.

[2] S. Shim, W. Lee, and C. Park. An efficient snapshot technique for ext3 file system in linux 2.6. realtime linux foundation(RTLW), Nov. 2005.

[3] D. A. Patterson, G. Gibson, and R. H. Katz. A case for redundant arrays of inexpensive disks (raid). Proceedings of the 1988 ACM SIGMOD international conference on Management of data, pages 109–116, jun 1988.

[4] A. Fox, S. D. Gribble, Y. Chawathe, E. A. Brewer, and P. Gauthier. Cluster-based scalable network services. Symposium on Operating Systems Principles, pages 78–91, 1997.

[5] V. S. Pai, M. Aron, G. Banga, M. Svendsen, P. Druschel, W. Zwaenepoel, and E. Nahum. Locality-aware request distribution in cluster-based network servers. SIGOPS Oper. Syst. Rev., pages 205–216, 1998.

[6] Spafford, Eugene H., ‘‘An Analysis of the Internet Worm,’’ Proc. European Software Engineering Conference, September 1989.

[7] Staniford, Stuart, Vern Paxson, and Nicholas Weaver, "How to Own the Internet in Your Spare Time," USENIX Security Symposium, August 2002.

[8] Frisch, Æleen, Essential System Administration Third Edition, O'Reilly & Associates, 2002.

[9] www.adaptec.com

[10] Iyad Aldasouqi, Adnan Shaout, Earthquake Monitoring System Using Ranger Seismometer Sensor, INTERNATIONAL JOURNAL of GEOLOGY, Issue 1, Volume 3, 2009

[11] Iyad Aldasouqi, Jalal Atoum, Stream Processing Environmental Applications in Jordan Valley, Computer Science Journals, 2010.

[12] Preston, W. Curtis, Unix Backup and Recovery, O'Reilly and Associates, 1999.

[13] Frisch, Æleen, Essential System Administration Third Edition, O'Reilly & Associates, 2002.

[14] The AMANDA Homepage, http://www.amanda.org .

[15] da Silva, J., and O. Guomundsson, "The Amanda Network Backup Manager," Proceedings of the Seventh Large Installation Systems Administration Conference (LISA), November 1993.

[16] Dantz, Dantz Retrospect – Intelligent Backup and Restore, http://www.nwfusion.com/whitepapers/dantz/whitepaper.html , June 2004.

[17] IBM Software, IBM Storage Management Solutions, http://www.nasi.com/tivoli_backuprecovery.htm , 2004.

[18] Legato Software, Legato Networker, http://www. legato.com/products/networker/

[19] www.ni.com/dataacquisition/

[20] NISHIMURA Satoshi, SANO Mutsuo, IKEDA Katsuo, The design and implementation of an extensible network backup system in real-time, Proceeding ICUIMC '09 Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication

Face Recognition Using Neural Network Based Fourier Gabor Filters & Random Projection

Anissa Bouzalmat *anissabouzalmat@yahoo.fr*
Faculty of Technical Sciences /Computer Sciences Department
Sidi Mohamed Ben Abdellah University/
Fez, 30 000, Morocco

Naouar Belghini *n.belghini@ieee.ma*
Faculty of Technical Sciences /Computer Sciences Department
Sidi Mohamed Ben Abdellah University/
Fez, 30 000, Morocco

Arsalane Zarghili **a***.zarghili@ieee.ma*
Faculty of Technical Sciences /Computer Sciences Department
Sidi Mohamed Ben Abdellah University/
Fez, 30 000, Morocco

Jamal Kharroubi *jamal.kharroubi@yahoo.fr*
Faculty of Technical Sciences /Computer Sciences Department
Sidi Mohamed Ben Abdellah University/
Fez, 30 000, Morocco

Aicha Majda *aicha_majda@yahoo.fr*
Faculty of Technical Sciences /Computer Sciences Department
Sidi Mohamed Ben Abdellah University/
Fez, 30 000, Morocco

Abstract

Face detection and recognition has many applications in a variety of fields such as authentication, security, video surveillance and human interaction systems. In this paper, we present a neural network system for face recognition. Feature vector based on Fourier Gabor filters is used as input of our classifier, which is a Back Propagation Neural Network (BPNN). The input vector of the network will have large dimension, to reduce its feature subspace we investigate the use of the Random Projection as method of dimensionality reduction. Theory and experiment indicates the robustness of our solution.

Keywords: Face Recognition, Fourier Transform, Gabor Filter, Neural Network, Sparse Random Projection.

1. INTRODUCTION

Human face detection and recognition is an active area of research spanning several disciplines such as computer vision and pattern classification. A robust face recognition system is a system based on good feature extractor method and good classifier. Neural network have been successfully applied to many pattern classification problems. And among the new techniques used in the literature for feature extraction, it is proven that Gabor filters can extract the maximum information from local image regions [1][2] and it is invariant against, translation, rotation, variations due to illumination and scale [3][4][5].

In[6][7]Gabor wavelets & neural network was presented for face detection, A. Khatun et al [8] propose a hybrid neural network solution for face recognition trained with Gabor features. The

neural network employed is based on BAM for dimensionality reduction and multi-layer perception with backpropagation algorithm for training the Gabor features.
P. Latha et al [9] use Gabor wavelet to present face, and applied neural network to classify views of faces. The dimensionality is reduced by the Principal component analysis.

In this study, we present an intelligent neural network system for face recognition. We use Gabor filters and Fourier transform for feature selection as they present desirable characteristics of spatial locality and orientation selectivity. Theses feature vectors are used as input of our Back Propagation Neural Network (BPNN), it was chosen as classifier for the proposed system because of its simplicity and its capability in supervised pattern recognition [10]. The input vector of the network will have large dimension, to reduce its feature subspace, we use Random Projection (RP)that has emerged as a powerful and efficient method for dimensionality reduction that preserves the structure of the data without introducing very significant distortion.[11][12].

This paper is organized as follows: Description of our solution is presented in section II, in section III, we discuss the experimental results and section IV gives conclusion and future works.

2. THE PROPOSED SOLUTION

2.1 System Architecture

The system proposed in this research is designed for facial face recognition. The system consists of three modules: a) Facial feature extraction using Gabor filter b) dimensionality reduction using sparse random projection. Finally, the obtained feature vectors are fed up into BPNN for classification.
The overall system architecture is shown in Figure 1.

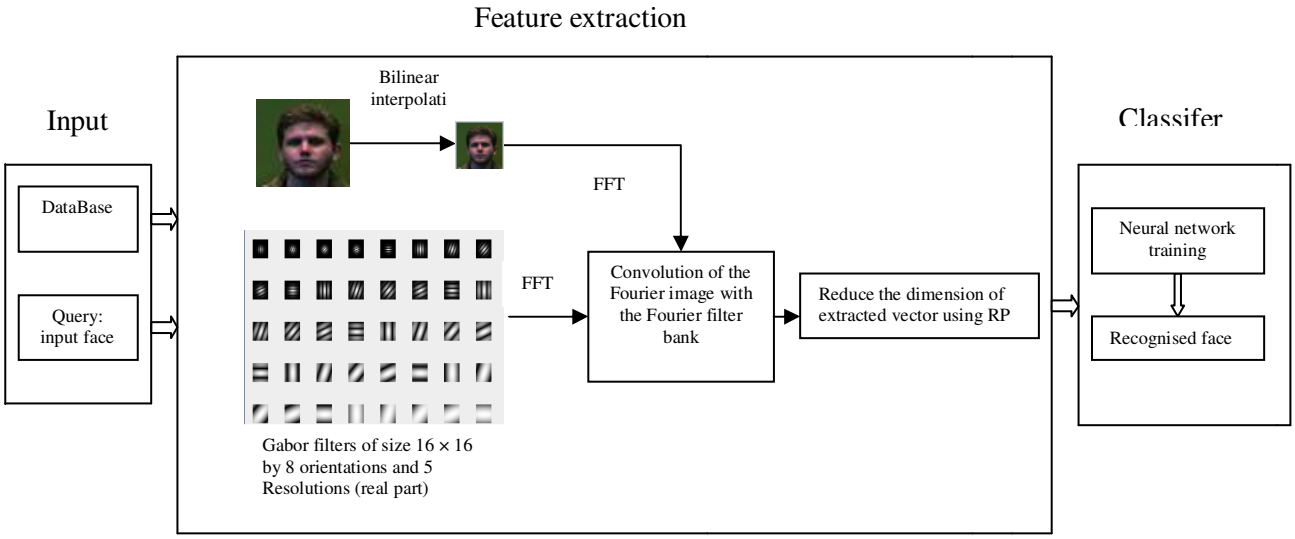


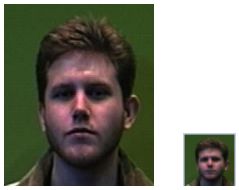
FIGURE 1: Architecture description of the proposed approach.

2.2 Extracting Feature Vectors

Several works [13] [14] have also shown that the Gabor filters representation and extraction of face images is robust. However, the high dimensional Gabor feature vectors caused the method to be computationally very expensive. Hence, the necessity to resize the original image (bilinear interpolation) and to apply a reduction dimensionality method (RP).

2.2.1 Bilinear Interpolation

The original image was reduced in size 32×32 by bilinear interpolation. This is necessary to reduce the computation time. The result is shown in (Figure 2) and then the Fourier transformed is applied to the image.



(a) (b)

FIGURE 2: Resizing of the face: a) input image, b)Bilinear interpolation resizing.

2.2.2 Fourier Transformed Image

Fourier Transformed image is the image I in the frequency domain as in this field every point represents a particular frequency contained in the image space of square image of size N × N.

$$Fourier(m,n,I)=\frac{1}{N^2}\sum_{a=0}^{N-1}\sum_{b=0}^{N-1} I(a,b)\quad e^{-i2\pi\left(\frac{ma+nb}{N}\right)} \qquad (1)$$

2.2.3 Gabor Filters

Gabor is a function that satisfies certain mathematical requirements extraction information is based on the use of a bank of Gabor filters [15], 8 orientations and 5 resolutions. The 2D Gabor filter is formed by modulating a complex sinusoid by a Gaussian function where each filter is defined by:

$$Gabor(x, y, \mu, \nu) = \theta(x, y, \mu, \nu)\left(\alpha - \beta\right) \qquad (2)$$

Where:

$$\theta(x, y, \mu, \nu) = \frac{\left\|k_{\mu\nu}\right\|^2}{\sigma^2}\ \exp\left(\frac{-\left\|k_{\mu\nu}\right\|^2\ \ \left(x^2+y^2\right)}{2\sigma^2}\right)$$

$$\alpha = \exp\left(ik_{\mu\nu} * (x,y)\right)$$

$$\beta = \exp\left(\frac{-\sigma^2}{2}\right)$$

Where (x,y) represents a 2-dimensional input point. The parameters μ and v define the orientation and scale of the Gabor kernel. ||.|| indicates the norm operator, and σ refers to the standard deviation of the Gaussian window in the kernel.

The wave vector K μυ is defined as:

$$\mathrm{k}_{\mu\nu} = k_{\nu}\exp^{i\varphi_{\mu}} \qquad (3)$$

Where: $\mathrm{k}_{\nu} = \frac{k\max}{f^{\nu}}$, $\varphi_{\mu} = \frac{\pi\mu}{8}$

if 8 different orientations are chosen.Kmax is the maximum frequency, and f υ is the spatial frequency between kernels in the frequency domain. In our configuration, 5 different scales and 8 orientations of Gabor wavelets are used, e.g. v € {0, ..., 4} and μ€ {0, ..., 7}. Gabor wavelets are chosen with the parameters:

$$\text{kmax}=\frac{\pi}{2} \quad , \quad f=\sqrt{2} \quad , \quad \sigma=\pi$$

The collection of all 40 Gabor kernels is called a filter bank. An example can be found in Figure 3.

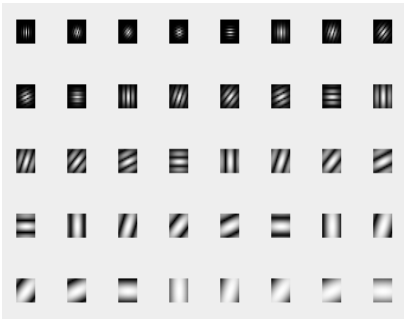


FIGURE 3: Gabor filters of size 16 × 16 by 8 orientations and 5 Resolutions (real part).

The Fourier Gabor wavelet representation of an image is the convolution of the Fourier image with the Fourier filter bank. The convolution of Fourier image F(I) and a Fourier Gabor kernel F(ψμ,v (x,y)) is defined as follows:

$$O_{\mu\nu}(x,y) = F\left(I(x,y)\right) * F\left(\psi_{\mu\nu}(x,y)\right) \quad (4)$$

and called Fourier Gabor feature. As the response Oμ,v(x,y) to each Fourier Gabor kernel is a complex function with a real part : Real{ Oμ,v(x,y)} and an imaginary part : Imag{ Oμ,v(x,y)},we use its real Real{ Oμ,v(x,y)} to represent the Fourier Gabor features.The complete set of Gabor wavelet representations of the image I(x,y) is:

$$G(I) = \{O_{\mu\nu}(x,y) : \mu \in \{0,..,7\}, \nu \in \{0,..,4\}\} \quad (5)$$

The resulting features for each orientation, scale are referred to as Fourier Gabor feature vector.The following algorithm shows the steps of respectful representation of the face with Fourier-Gabor filters.

- Algorithm 1:

1. Prepare 5 × 8 matrix Gabor each of size 16 × 16 as shown (Figure 3).

2. Apply the Fourier transform to each matrix Gabor.

3. Apply Fourier to each image in the training set of size 32×32.

4. Convolution of the Fourier transform of the image size 32×32 by each image of the Fourier transformed Gabor size 16 × 16 (8 orientations and 5 scales) .

5. Construct the image Fourier_Gabor_IMG (5×8×32×32) from the sub images (32×32 obtained in step 4) (Figure 4).

FIGURE 4: Fourier_Gabor_IMG (5×8×32×32) Results Convolution of Fourier transformed image (32 × 32) for the Fourier transformed of each Gabor filter 16× 16.

The use of Gabor filters is very expensive in computing time, due to the convolution of the whole image with filter size 16 × 16. For this reason, we limit the use of the image size of 32 × 32 convolved with 40 Gabor filters: 8 orientations and 5 scales, then we resize the image result (Fourier_Gabor_IMG) to 100*100,and Finally we reduce the vector of features by applying the method of random projection.

2.2.4 Sparse Random Projection

In the computer vision literature, many schemes have been investigated for finding projections that better represent data in lower-dimensional spaces. One benefit of feature extraction, which carries over to the proposed sparse representation framework, is reduced data dimension and computational cost. The choice of feature transformation is considered critical to the success of the algorithm.

Random Projection has been applied on various types of problems like machine learning [16]. Its power comes from the strong theoretical results that guarantee a very high chance of success [11].

Bingham. et al [17] present experimental results on using RP as a dimensionality reduction tool, their application areas were the processing of both noisy and noiseless images, and information retrieval in text documents. They show that projecting the data onto a random lower-dimensional subspace yields results comparable to conventional dimensionality reduction methods such as PCA and RP is computationally significantly less expensive than it.

Let $X \in R^n$ .The method multiplies X by a random matrix $RP \in R^{n*k}$: $Y^k = RP * X$

The idea is to preserve as much the “structure” of the data while reducing the number of dimensions it possesses; Projections are based on the Johnson-Lindenstrauss lemma [19] that states that a set of n points in a high dimensional Euclidean space can be mapped down onto a $k\rangle = o\left(\frac{\log(n)}{\varepsilon^2}\right)$ dimensional subspace and provided that RP has i.i.d. entries with zero mean and unit variance[18].

Initially, random projections were done with a normal matrix, where each entry rij was an independent, identically distributed N(0, 1) variable with not orthogonal subspace.

Achlioptas provided the sparse matrix projection that refer to a powerful concentration bounds(s=3 and s=1)[19]

Recently, Li et al. generalize Achlioptas’ result by providing the very-sparse projection matrix, they show that s>>3 can be used (for example $s = \frac{n}{\log(n)}$ [20]).

$$r_{ij} = +\sqrt{s} \begin{cases} +1 & p = \frac{1}{2s} \\ 0 & p = 1 - \frac{1}{s} \\ 1 & p = \frac{1}{2s} \end{cases} \qquad (5)$$

The Johnson-Lindenstrauss lemma proof that we can reduce to k> O(log(n)/ϵ2) dimension in order to approximately preserve pairwise distances up to a factor of (1±ϵ). Practically it is interested to get some explicit formula for k.

A series of simplifications to the original proof of Johnson and Lindenstrauss, culminating showed

that: $k\rangle = \frac{Log(n) \ * \ 4}{\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}}$

This is not a strict lower-bound but deduce that the pairwise distance is probably preserved with the Johnson-Lindenstrauss guarantees[11]. We will test in the experimental section how the rate of recognition is affected with different values of k when using sparse random projection with s=1. After the generation of vector features with reduced dimension, back propagation neural network is applied for recognition.

2.3 Back propagation Neural Network

An artificial neural network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems process information. It is configured for a specific application through a specific learning process. The most commonly used family of neural networks for pattern classification tasks is the feed-forward network, which includes multilayer perceptron and Radial-Basis Function (RBF) networks.

Back propagation is a feed forward supervised learning network. The general idea with the backpropagation algorithm is to use gradient descent to update the weights to minimize the squared error between the network output values and the target output values. The update rules are derived by taking the partial derivative of the error function with respect to the weights to determine each weight's contribution to the error. Then, each weight is adjusted. This process occurs iteratively for each layer of the network, starting with the last set of weights, and working back towards the input layer, hence the name "backpropagation". The network is trained to perform its ability to respond correctly to the input patterns that are used for training and to provide good response to input that are similar.
The general overview used for the recognition task is as follow:

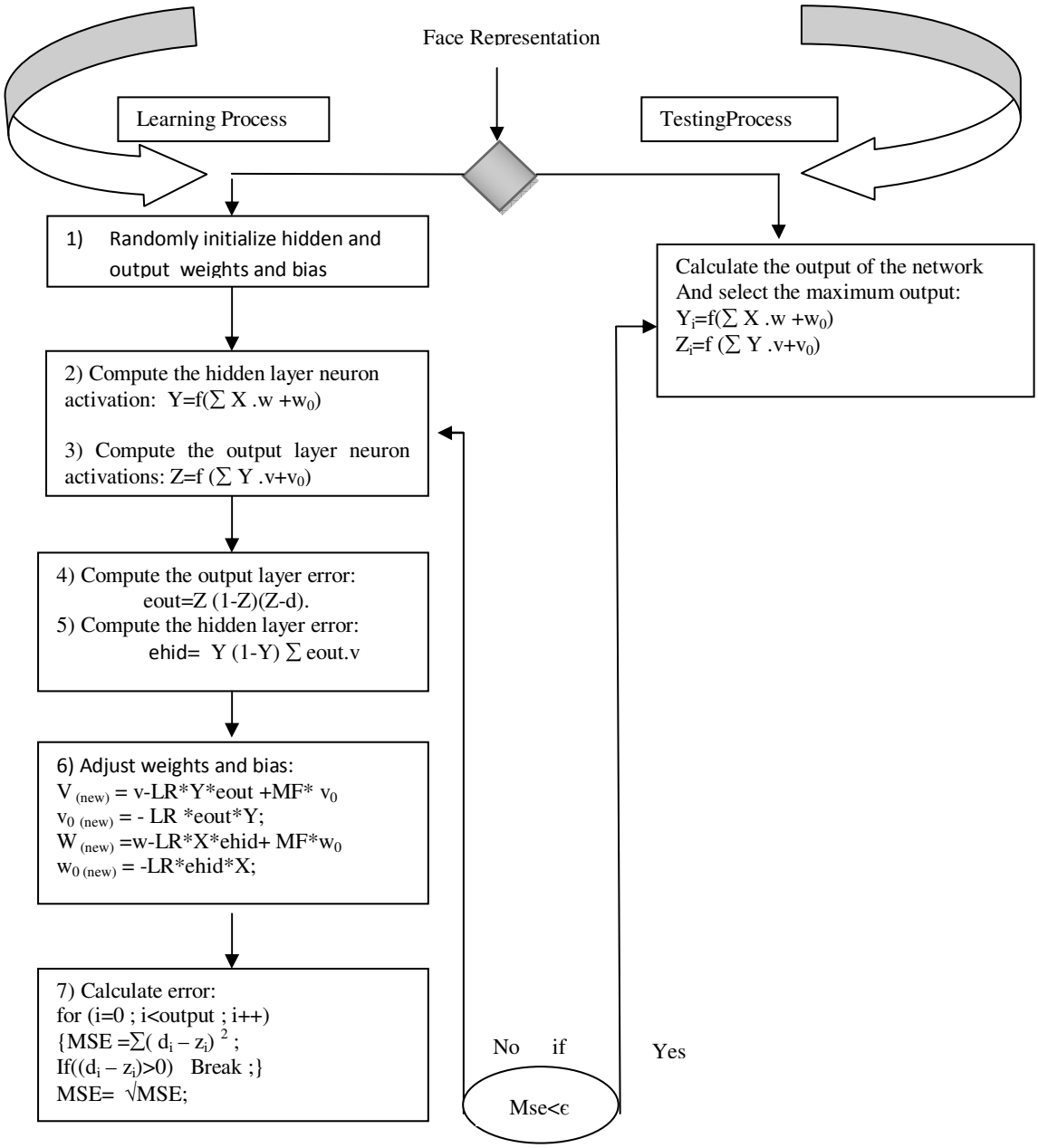


FIGURE 5: The recognition process.

Where X the vector of input layer neurons, Y is the vector of the hidden layer neurons, and Z represents the output layer neurons. w is the weight matrix between the input and the hidden layer. w_0 is the bias of the hidden layer neurons. v is the weight matrix connecting the hidden and the output layers, and v_0 is the bias of the output layer neurons. eout is the error vector for output neurons and ehid is the error vector of each hidden layer neuron and d is the desired output vector.

LR and MF are learning rate and momentum factor.

The sigmoid activation function is defined by: $f(x) = \frac{1}{(1+\exp(-x))}$

3. EXPERIMENTATION AND RESULTS

In [22] authors compare RP with PCA, their results show that PCA performs better than RP mostly for low dimensions (20-30). This result is consistent with previous studies where it has been reported that RP compares favorably with PCA for moderate or higher number of dimensions. It is clear as shown in Figure 6, that in the high dimension RP become more effective than ACP.

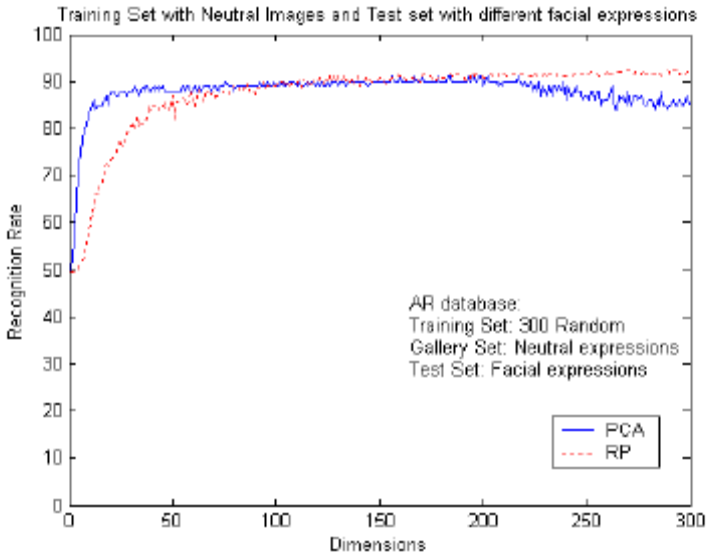


FIGURE 6: Experiments using the AR database and majority voting. The subjects in the training set were chosen randomly. The proportion of subjects in the gallery and test sets varies (neutral versus facial expressions). The blue line corresponds to PCA using closest match, while the red line corresponds to RP using majority voting [21].

In our experiments, the face image database used is a collect of 20 Persons from database [22]. These face images varies in facial expression and motion. Each person is represented by 20 samples, 10 are used for training and 10 for test.

We implement the algorithm described above and we evaluate how the rate of recognition is affected using sparse random projection (s=1) and using the lower-bound value proposed by Johnson-Lindenstrauss[11] .

$$k \rangle K_0 = \frac{Log(n) \quad * \quad 4}{\frac{\varepsilon^2}{2} \quad - \quad \frac{\varepsilon^3}{3}}$$

Then we compare the obtained results with results obtained without applying the random projection i.e. using the original data.

Some scenarios of Training are presented in the following: The error is set to 0.0009 for stopping condition.

Case1: length of the original feature vector n=10000

Case2: length of the feature original vector n=1000

Case3: length of the original feature vector and applied sparse RP with s=1 and k=260

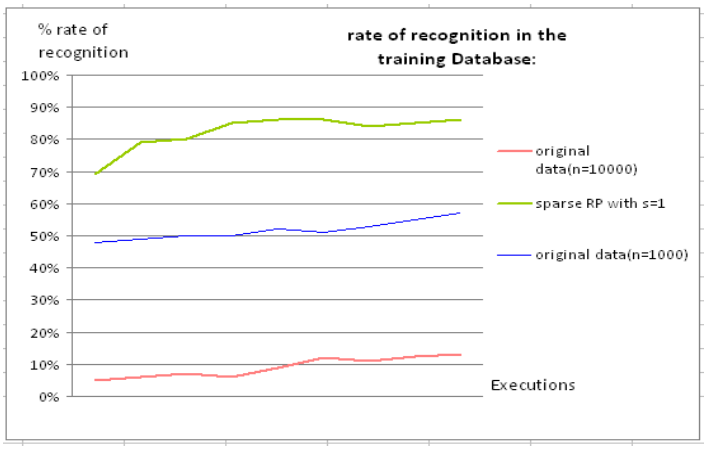


FIGURE 7: Curve of rate of recognition with original data and when applying sparse RP.

We remark that when n=10000, the rate of recognition with original data do not surpass 10%, with n=10000 is between 40% and 60% whereas it attends 87% when introducing RP(s=1 and k=260). Is seems clearly that original data cannot be adopted to train de neural network.

Some query and recognized image are shown in Figure 8.

a) Query images b) Recognized images

FIGURE 8: Example of input and output of our system

Since our features extraction vectors have high dimension, we assume that random projection is an adequate method of dimensionality reduction. In the case of our study, obtain a higher FR rate depends on the choice of the random projection matrix and the dimension of the feature vector of original data.

4. CONCLUSION AND PERSPECTIVES

This paper develops technique to extract the features vector of the whole face in image DB by using Gabor filters which are known to be invariant to illumination and facial expression.

We introduce 8 different orientations and 5 different resolutions to extract the maximum of information, to reduce the dimension of the result vector, we apply sparse random projection, it provides many advantages: it is easy to implement, fast and more effective when compared to other methods. BPNN is then applied to perform the recognition task. Our network achieves higher recognition rate and better classification efficiency when the feature vectors have low-dimensions. This solution was implemented using Java environment. The effectiveness of the proposed method is demonstrated by the experimental results.

In the future, local feature extracted methods will be investigated for classification.

5. REFERENCES

[1] H. Deng, L. Jin, L. Zhen, and J. Huang. "A new facial expression recognition method based on local gabor filter bank and pca plus lda". International Journal of Information Technology, vol.11, pp.86-96, 2005.

[2] L. Shen and L. Bai. "Information theory for gabor feature selection for face recognition", Hindawi Publishing Corporation, EURASIP Journal on Applied Signal Processing, Article ID 30274, 2006.

[3] Z. Y. Mei, Z. Ming, and G. YuCong. "Face recognition based on low diamensional gabor feature using direct fractional-step lda", In Proceedings of the Computer Graphics, Image and Vision: New Treds, IEEE Computer Society, 2005.

[4] B. Schiele, J. Crowley,"Recognition without correspondence using mul-tidimensional receptive field histograms", International Journal on Computer Vision, 2000.

[5] A. Bouzalmat, A. Zarghili, J. Kharroubi, "Facial Face Recognition Method Using Fourier Transform Filters Gabor and R_LDA", IJCA Special Issue on Intelligent Systems and Data Processing, pp.18-24, 2011.

[6] C.Sharma, "face detection using gabor feature extraction technique", Journal of Global Research in Computer Science, vol.2 (4), pp.40-43, April 2011.

[7] A.Kaushal and J P S Raina, "Face Detection using Neural Network & Gabor Wavelet Transform", International Journal of Computer Science and Technology, Vol. 1, Issue 1, September 2010.

[8] A.Khatun and Md.Al-Amin Bhuiyan, "Neural Network based Face Recognition with Gabor Filters", IJCSNS International Journal of Computer Science and Network Security, vol.11 No.1, January 2011.

[9] P.Latha, L.Ganesan, N.Ramaraj, "Gabor and Neural based Face Recognition", International Journal of Recent Trends in Engineering, Vol 2, No. 3, November 2009.

[10] C.M Bishop, Neural Networks for Pattern Recognition, London, U.K: Oxford University Press, 1995.

[11] A.K Menon, "Random projections and applications to dimensionality reduction", Phd thesis, School of Information Technologies, The University of Sydney, Australia, 2007.

[12] N. Belghini, A. Zarghili, J. Kharroubi and A. Majda, Sparse Random Projection and Dimensionality Reduction Applied on Face Recognition, in The Proceedings of International Conference on Intelligent Systems & Data Processing, January 2011,pp.78-82.

[13] R.Rao and D.Ballard. "An active vision architecture based on iconic representations", Artificial Intelligence, pp.461-505,1995.

[14] B.Schiele and J.Crowley. "Recognition without correspondence using multidimensional receptive field histograms". On Computer Vision, 2000.

[15] J Essam Al Daoud, "Enhancement of the Face Recognition Using a Modified Fourier-Gabor Filter", International Journal Advance. Software Computer. Applications, Vol. 1, No. 2, 2009.

[16] D. Fradkin and D. Madigan, “Experiments with random projection for machine learning,” in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[17] E.Bingham and H.Mannila. “Random projection in dimensionality reduction: Applications to image and text data”, In Proc. of KDD, San Francisco, CA, 2001.

[18] R.Arriaga and Santosh Vempala. “An algorithmic theory of learning: Robust concepts and random projection”, In Proc. of FOCS, 1999.

[19] D.Achlioptas, “Database-friendly random projections:Johnson-Lindenstrauss with binary coins”, Journal of Computer and System Sciences, 2003.

[20] P.Li, T.J.Hastie, and K.W.Church. “Very sparse random projections”. In KDD ’06: Proceedings of the 12th ACM SIGKDD .international conference on Knowledge discovery and data mining, 2006, p 287–296.

[21] N.Goel, G.Bebis, and A.Nfian. “Face recognition experiments with random projection”. In Proc. of SPIE, 2005.

[22] Face Recognition DataBase from University of Essex (UK) at http://cswww.essex.ac.uk/mv/allfaces/index.html

Finding Relationships Between the Our-NIR Cluster Results

N.Sudhakar Reddy *sudhakar.n@svcolleges.edu.in*
Professor in CSE Department of CSE
S.V.College of Engineering
Tirupati, India

K.V.N.Sunitha
Department of CSE
G.Narayanamma Institute of Technology and Science
Hyderabad, India

Abstract

The problem of evaluating node importance in clustering has been active research in present days and many methods have been developed. Most of the clustering algorithms deal with general similarity measures. However In real situation most of the cases data changes over time. But clustering this type of data not only decreases the quality of clusters but also disregards the expectation of users, when usually require recent clustering results. In this regard we proposed Our-NIR method that is better than Ming-Syan Chen proposed a method and it has proven with the help of results of node importance, which is related to calculate the node importance that is very useful in clustering of categorical data, which is for evaluating of node importance by introducing the probability distribution which will be better than by comparing the results .That is detects drifting concepts and try to show the evolving clustering results in the categorical domain. This scheme is based on the cosine measure that analyzes relationship between clustering results at different time stamps using Our-NIR method

Keywords: Clustering, Weather Prediction, Drifting, Our-NIR.

1. INTRODUCTION

Extracting Knowledge from large amount of data is difficult which is known as data mining. Clustering is a collection of similar objects from a given data set and objects in different collection are dissimilar. Most of the algorithms developed for numerical data may be easy, but not in Categorical data [1, 2, 11, 12]. It is challenging in categorical domain, where the distance between data points is not defined. It is also not easy to find out the class label of unknown data point in categorical domain. Sampling techniques improve the speed of clustering and we consider the data points that are not sampled to allocate into proper clusters. The data which depends on time called time evolving data. For example, the buying preferences of customers may change with time, depending on the current day of the week, availability of alternatives, discounting rate etc. Since data evolve with time, the underlying clusters may also change based on time by the data drifting concept [10, 15]. The clustering time-evolving data in the numerical domain [1, 5, 6, 9] has been explored in the previous works, where as in categorical domain not that much. Still it is a challenging problem in the categorical domain.

As a result, our contribution in modifying the frame work which is proposed by Ming-Syan Chen in 2009[8] utilizes any clustering algorithm to detect the drifting concepts. We adopted sliding window technique and initial data (at time t=0) is used in initial clustering. These clusters are represented by using Chen NIR and Our-NIR [8, 19], where each attribute value importance is measured. We find whether the data points in the next sliding window (current sliding window) belongs to appropriate clusters of last clustering results or they are outliers. We call this clustering result as a temporal and compare with last clustering result to drift the data points or not. If the concept drift is not detected to update the Our-NIR otherwise dump attribute value based on importance and then reclustering using clustering techniques [19]. In this paper mainly

concentrating on the inter-similarity of adjacent clusters from time to time based similarity measure that is easy to find the drifts are occurred or not.

The rest of the paper is organized as follows. In section 2 discussed related works, in section 3 vector representation provided, in section 4 cosine measure for relation analysis among the clusters discussed and also contains results with comparison of Ming-Syan Chen method and Our-NIR method and finally concluded with section 5.

2. RELATED WORK

In this section, we discuss various clustering algorithms on categorical data with cluster representatives and data labeling. We studied many data clustering algorithms with time evolving. Cluster representative is used to summarize and characterize the clustering result, which is not fully discussed in categorical domain unlike numerical domain. In K-modes which is an extension of K-means algorithm in categorical domain a cluster is represented by 'mode' which is composed by the most frequent attribute value in each attribute domain in that cluster. Although this cluster representative is simple, only use one attribute value in each attribute domain to represent a cluster is questionable. It composed of the attribute values with high co-occurrence. In the statistical categorical clustering algorithms [3,4] such as COOLCAT and LIMBO, data points are grouped based on the statistics. In algorithm COOLCAT, data points are separated in such a way that the expected entropy of the whole arrangements is minimized. In algorithm LIMBO, the information bottleneck method is applied to minimize the information lost which resulted from summarizing data points into clusters. However, all of the above categorical clustering algorithms focus on performing clustering on the entire dataset and do not consider the time-evolving trends and also the clustering representatives in these algorithms are not clearly defined.

The new method is related to the idea of conceptual clustering [9], which creates a conceptual structure to represent a concept (cluster) during clustering. However, NIR only analyzes the conceptual structure and does not perform clustering, i.e., there is no objective function such as category utility (CU) [11] in conceptual clustering to lead the clustering procedure. In this aspect our method can provide in better manner for the clustering of data points on time based. The main reason is that in concept drifting scenarios, geometrically close items in the conventional vector space might belong to different classes. This is because of a concept change (drift) that occurred at some time point. Our previous work [19, 20] addresses the node importance in the categorical data with the help of sliding window. That is new approach to the best of our knowledge that proposes these advanced techniques for concept drift detection and clustering of data points.

After scanning the literature, it is clear that clustering categorical data is un touched many ties due to the complexity involved in it. A time-evolving categorical data is to be clustered within the due course hence clustering data can be viewed as follows: there are a series of categorical data points D is given, where each data point is a vector of q attribute values, i.e., $p_j=(p_j^1,p_j^2,...,p_j^q)$. And $A = \{A_1,A_2,..., A_q\}$, where A_a is the a^{th} categorical attribute, $1 \leq a \leq q$. The window size N is to be given so that the data set D is separated into several continuous subsets S^t, where the number of data points in each S^t is N *shown in figure 1*. The superscript number t is the identification number of the sliding window and t is also called time stamp. Here in we consider the first N *data points of data set D* this makes the first data slide or the first *sliding window* S^1 *or S1*. The intension is to cluster every data slide and relate the clusters of every data slide with previous clusters formed by the previous data slides. Several notations and representations are used in our work to ease the process of presentation. In the previous work we considered the sample data set for the clustering of concept drift categorical data in that paper initially clustering done by standard algorithm that result shown in figure 1 and finally concluded with the updated Our-NIR results respect to sliding window and clusters as shown in figure 2[20] .Based on the relationship analysis, the evolving clusters will provide clues for us to catch the time evolving trends in the data set. This can achieve by introducing vector model and cosine measure, the similarity measure is most efficient for the vector representation.

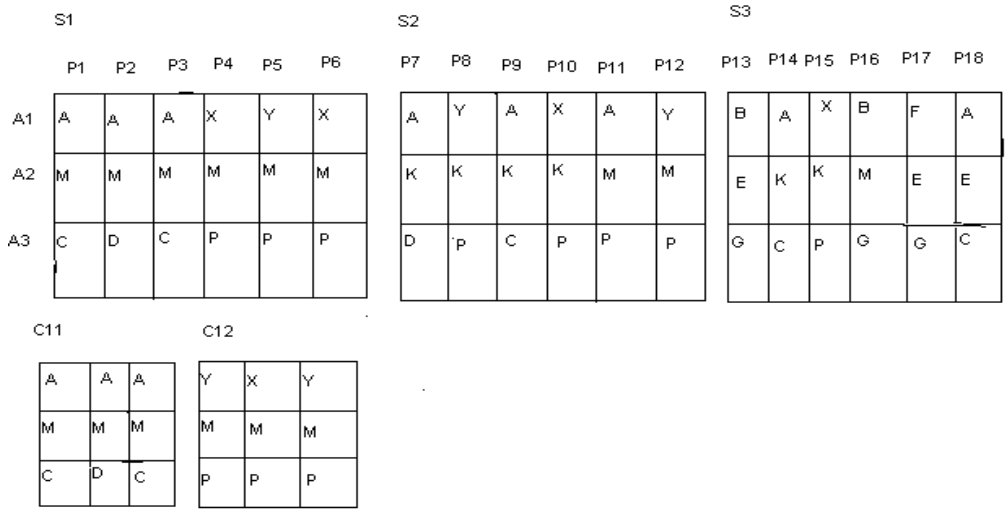


FIGURE 1. Data set with sliding window size 6 where the initial clustering is performed

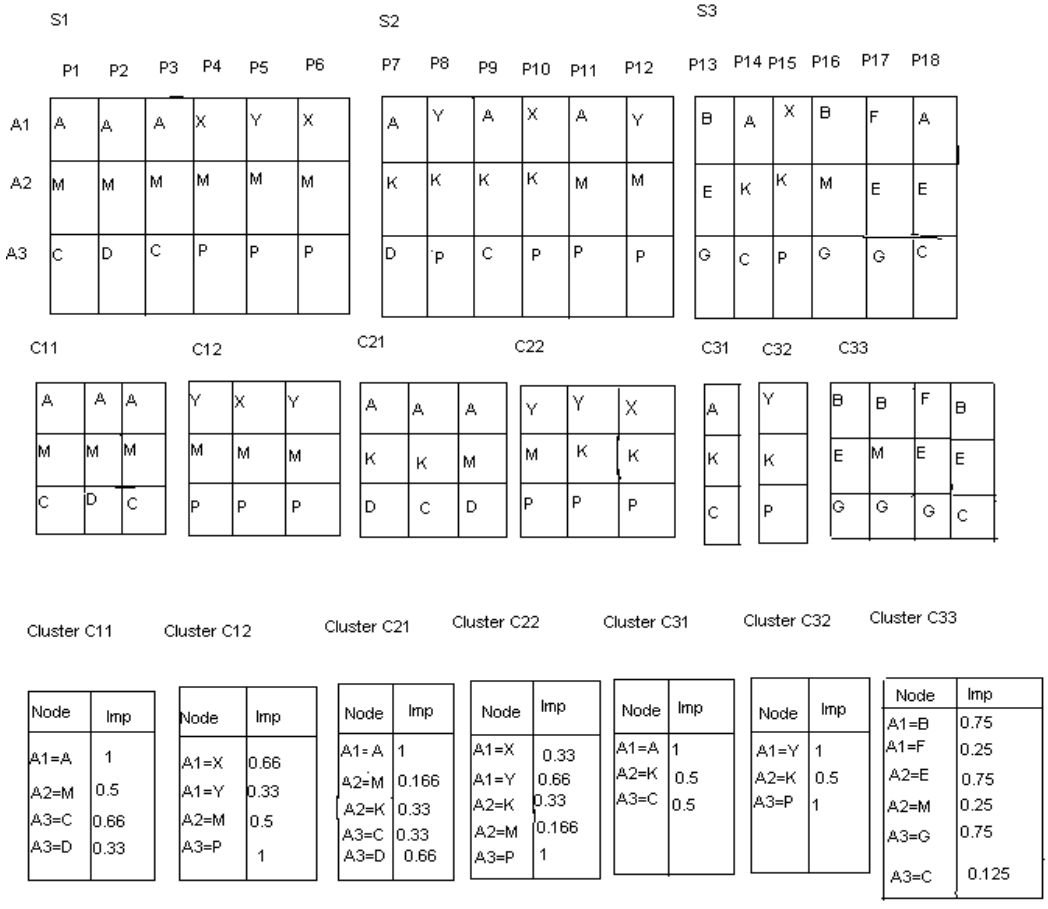


FIGURE 2: Final clustering results as per the data set of fig 1 and output Our-NIR Results

3. VECTOR REPRESENTATION

The vector model is a view of the representative to contain the domain of nodes. The size of vector is based on the total number of nodes in entire data set. A cluster in this space is a vector, and an each index of vector is the value of importance by Our-NIR method in that node domain. Based on the node vector representation, the node Our-NIR Vector of cluster C_i is shown as follows:

$C_{i=}$ $(w_i(I_1),w_i(I_2),.....w_i(I_i),.....w_i(I_z))$,

Where $W_i(I_r)=o$,

$$Wi(I_i)=w(C_i,i_{ir}), \begin{cases} \text{if } I_r \text{ does not occur in } C_i, \\ \text{if } I_r \text{ occur in } C_i. \end{cases}$$

All the nodes in entire data set can be represented in this model with the following calculations:

1. The weight of each node across the entire data set needs to be calculated based on sliding window data set and Our-NIR method [20]. This gives how important the node is in the sliding window of data set.
2. The weight of every node within a given sliding window needs to be calculated for all slidings. This obtains how important the node is within a single sliding window.
3. Every two adjacent vectors of the sliding window clusters are compared

The value in the vector C_i on each node domain is the Our-NIR value of this node in cluster C_i, i.e., $W (c_i, N_{[i, r]})$. If the node does not occur in cluster C_i, the value in the vector C_i on this node domain is zero. Here contains all distinct nodes that occur in the entire data set, not just in cluster C_i based on the domain of attribute values. Therefore, the dimensions of all the vectors C_i are the same.

	A	B	C	D	E	F	G	K	M	P	X	Y	|Cij|
C11	1	0	0.66	0.33	0	0	0	0	0.5	0	0	0	1.33
C12	0	0	0	0	0	0	0	0	0.5	1	0.66	0.33	1.33
C21	1	0	0.33	0.66	0	0	0	0.33	0.166	0	0	0	1.2965
C22	0	0	0	0	0	0	0	0.33	0.166	1	0.33	0.66	1.2965
C31	1	0	0.5	0	0	0	0	0.5	0	0	0	0	1.2247
C32	0	0	0	0	0	0	0	0.5	0	1	0	1	1.5
C33	0	0.75	0.125	0	0.75	0.25	0.75	0	0.25	0	0	0	1.352

FIGURE 3: Our-NIR Vectors C1, C2 and C3 of the clustering results C1, C2 and C3 In fig 2

Example: In the example data set shown in fig 1, in that figure there are totally 12 distinct nodes in the entire data set and the Our-NIR results of C11 and C12 are shown in fig 3 based on this

figure 2 the vector space defined as said above in this section the vector of cluster C11 and similarly for the remaining clusters as shown in figure 3.

The clusters C_i and C_j are represented by the Our-NIR vectors C_i and C_j . We studied several similarity measures for the finding of similarity of clusters, finally concluded among them the cosine measure is often used to compare documents in text mining. In addition, it is used to measure cohesion within clusters in the field of Data Mining.

4. COSINE MEASURE

The cosine treats both vectors as unit vectors by normalizing them, it calculates the cosine of the angle between the two vectors. It does provide an accurate measure of similarity but with no regard to magnitude. But magnitude is an important factor while considering similarity. It is popular measure of similar in the vector representation [14] . The cosine measure between vectors C_i and C_j is calculated as the shown equation 1.

$$\text{Similarity} = \cos\emptyset = \frac{\sum_{i,j=1}^{n} Ci \times Cj}{\sqrt{\sum_{i=1}^{n} Ci^2} * \sqrt{\sum_{i=1}^{n} Cj^2}} \qquad \text{---------> 1}$$

Consider the clustering results C11 and C22 in fig 3. The Our-NIR vectors of the clustering results C11 and C12 are shown in fig 4. The similarity between vectors C11 and C21 is 0.8933 and similarly calculated for the other clusters.

In addition, cosine measure of C22 and C32 is 0.900, which is larger the C11 and C21. Therefore, cluster C22 is said to be more similar to C12 than to cluster C11.

	$\overline{C11}$	$\overline{C12}$	$\overline{C21}$	$\overline{C22}$	$\overline{C31}$	$\overline{C32}$	$\overline{C3}$
$\overline{C11}$			0.88	0.048			
$\overline{C12}$			0.048	0.88			
$\overline{C21}$					0.8376	0.084	0.023
$\overline{C22}$					0.01039	0.9384	0.023

FIGURE 4: cosine similarity table between the clustering results c1 and c2 and between the c2 and c3 by Our-NIR results in fig 3.

	$\overline{C11}$	$\overline{C12}$	$\overline{C21}$	$\overline{C22}$	$\overline{C31}$	$\overline{C32}$	$\overline{C3}$
$\overline{C11}$			0.9296	0			
$\overline{C12}$			0	0.9296			
$\overline{C21}$					0.178	0	0
$\overline{C22}$					0	0.186	0

FIGURE 5: Cosine similarity table between the clustering results c1 and c2 and between the c2 and c3 By CNIR results

In figure 4 the similarity of each pair of adjacent clustering results, where t^b is the time stamp that different concepts happens, is measured by the cosine measure. Based on this measure, it provides for us to catch the time-evolving trend in the data set and also it could help for how to link the clusters at different time stamps.

Comparison of CNIR and Our-NIR

The cosine similarity of each pair clustering results of both the CNIR and Our-NIR shown in figure 5. As per the observation in that figure some of the inter-clusters may get zero similarity by CNIR where as in Our-NIR getting different. That shows the relationship between the clustering results at different time stamps. At same time when we are looking into the sample data set in figure 1 there it could be different with the CNIR result that means Our-NIR showing the better performance.

5. CONCLUSION

In this paper, a frame work proposed by Ming-Syan Chen Node Importance Representative (CNIR) in 2009[8] which is modified by new method that is Our-NIR to find node importance by us [19]. We analyzed by taking same example in this find the differences in the node importance values of attributes [19] in same cluster which plays an important role in clustering. The representatives of the clusters help improving the cluster accuracy and purity and hence the Our-NIR method performs better than the CNIR method [8].The pairing of each adjacent clusters similarity is based on Our-NIR method better than the CNIR in terms of cluster distribution. The future work improves the performance of precision and recall of DCD by introducing the leaders-subleaders algorithm for reclustering.

REFERENCES

[1] C. Aggarwal, J. Han, J. Wang, and P. Yu, "A Framework for Clustering Evolving Data Streams," *Proc. 29th Int'l Conf.Very Large Data Bases (VLDB) ,*2003.

[2] C.C. Aggarwal, J.L. Wolf, P.S. Yu, C. Procopiuc, and J.S. Park, "Fast Algorithms for Projected Clustering," *Proc. ACM SIGMOD" 1999,* pp. 61-72.

[3] P. Andritsos, P. Tsaparas, R.J. Miller, and K.C. Sevcik, "Limbo: Scalable Clustering of Categorical Data," *Proc. Ninth Int'l Conf. Extending Database Technology (EDBT),* 2004.

[4] D. Barbará, Y. Li, and J. Couto, "Coolcat: An Entropy-Based Algorithm for Categorical Clustering," *Proc. ACM Int'l Conf. Information and Knowledge Management (CIKM),* 2002.

[5] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-Based Clustering over an Evolving Data Stream with Noise," *Proc. Sixth SIAM Int'l Conf. Data Mining (SDM),* 2006.

[6] D. Chakrabarti, R. Kumar, and A. Tomkins, "Evolutionary Clustering,"*Proc. ACM SIGKDD" 2006,* pp. 554-560..

[7] H.-L. Chen, K.-T. Chuang and M.-S. Chen, "Labeling Unclustered Categorical Data into Clusters Based on the Important Attribute Values," *Proc. Fifth IEEE Int'l Conf. Data Mining (ICDM),* 2005.

[8] H.-L. Chen, M.-S. Chen, and S-U Chen Lin "Frame work for clustering Concept –Drifting categorical data," *IEEE Transaction Knowledge and Data Engineering v21 no 5 ,* 2009.

[9] D.H. Fisher, "Knowledge Acquisition via Incremental Conceptual Clustering," Machine Learning, 1987.

[10] Fan, W. *Systematic data selection to mine concept-drifting data streams*. in *Tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. 2004. Seattle, WA, USA: ACM Press: p. 128-137.

[11] MM Gaber and PS Yu "Detection and Classification of Changes in Evolving Data Streams," *International .Journal .Information Technology and Decision Making,* v5 no 4, 2006.

[12] M.A. Gluck and J.E. Corter, "Information Uncertainty and the Utility of Categories," Proc. Seventh Ann. Conf. Cognitive Science Soc., pp. 283-287, 1985.

[13] G Hulton and Spencer, "Mining Time-Changing Data Streams" *Proc. ACM SIGKDD,* 2001.

[14] AK Jain MN Murthy and P J Flyn "Data Clustering: A Review," *ACM Computing Survey,* 1999.

[15] Klinkenberg, R., *Learning Drifting Concepts: Example Selection vs. Exam- ple Weighting* Intelligent Data Analysis, Special Issue on Incremental Learn- ing Systems Capable of Dealing with Concept Drift, 2004. **8**(3): p. 281-200.

[16] O.Narsoui and C.Rojas,"Robust Clustering for Tracking Noisy Evolving Data Streams" SIAM I*nt. Conference Data Mining ,* 2006.

[17] C.E. Shannon, "A Mathematical Theory of Communication," Bell System Technical J., 1948.

[18] .Viswanadha Raju, H.Venkateswara Reddy andN.Sudhakar Reddy," A Threshold for clustering Concept – Drifting Categorical Data", IEEE Computer Society, ICMLC 2011.

[19] S.Viswanadha Raju,H.Venkateswara Reddy and N.Sudhakar Reddy " Our-NIR:Node Importance Representation of Clustering Categorical Data ", IJCST June 2011.

[20] S.Viswanadha Raju, N.Sudhakar Reddy and H.Venkateswara Reddy," Clustering of Concept Drift Categorical Data using Our-NIR Method, IJEE 2011

INSTRUCTIONS TO CONTRIBUTORS

The *International Journal of Computer Science and Security (IJCSS)* is a refereed online journal which is a forum for publication of current research in computer science and computer security technologies. It considers any material dealing primarily with the technological aspects of computer science and computer security. The journal is targeted to be read by academics, scholars, advanced students, practitioners, and those seeking an update on current experience and future prospects in relation to all aspects computer science in general but specific to computer security themes. Subjects covered include: access control, computer security, cryptography, communications and data security, databases, electronic commerce, multimedia, bioinformatics, signal processing and image processing etc.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCSS.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 5, 2011, IJCSS appears in more focused issues. Besides normal publications, IJCSS intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

IJCSS LIST OF TOPICS

The realm of International Journal of Computer Science and Security (IJCSS) extends, but not limited, to the following:

- Authentication and authorization models
- Computer Engineering
- Computer Networks
- Cryptography
- Databases
- Image processing
- Operating systems
- Programming languages
- Signal processing
- Theory
- Communications and data security
- Bioinformatics
- Computer graphics
- Computer security
- Data mining
- Electronic commerce
- Object Orientation
- Parallel and distributed processing
- Robotics
- Software engineering

CALL FOR PAPERS

Volume: 6 - **Issue:** 1 - February 2012

i. Paper Submission: November 30, 2011 **ii. Author Notification:** January 01, 2012

iii. Issue Publication: January / February 2012

CONTACT INFORMATION

Computer Science Journals Sdn BhD
B-5-8 Plaza Mont Kiara, Mont Kiara
50480, Kuala Lumpur, MALAYSIA

Phone: 006 03 6207 1607
006 03 2782 6991

Fax: 006 03 6207 1697

Email: cscpress@cscjournals.org

CSC PUBLISHERS © 2011
COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA

PHONE: 006 03 6207 1607
006 03 2782 6991

FAX: 006 03 6207 1697
EMAIL: cscpress@cscjournals.org