



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

ISSN (ONLINE): 2279-0055

ISSN (PRINTED): 2279-0047

Organised by:
Centre for Development of Advanced Computing (CDAC), Noida, Uttar Pradesh, INDIA



International Association of Scientific Innovation and Research (IASIR)

(An Association Unifying the Sciences, Engineering, and Applied Research)

STEM International Scientific Online Media and Publishing House

B-126, Freedom Fighter Enclave, Neb Sarai, IGNOU Road, New Delhi-110 016, India

Offices Overseas: USA, Canada, Australia, Germany, Netherlands

Website: www.iasir.net, E-mail (s): iasir.journals@iasir.net, iasir.journals@gmail.com, ijetcas@gmail.com

**Editorial Board Members of National Conference on Information Technology,
Electronics and Management (NCITEM-2017), July 20-21, 2017**

Chief Editor

Mr. V. K. Sharma,
Patron, NCITEM-2017 &
Director,
C-DAC, Noida

Editors:

Dr. Arti Noor,
Convener, NCITEM-2017 &
Joint Director,
C-DAC, Noida

Ms. Priti Razdan,
Convener, NCITEM-2017 &
Joint Director,
C-DAC, Noida

Ms. Kriti Saroha,
Co-Convener, NCITEM-2017 &
Joint Director,
C-DAC, Noida

Mr. Sanjay Ojha,
Co-Convener, NCITEM-2017 &
Asst. Professor,
C-DAC, Noida

TABLE OF CONTENTS

(Special Issue, 2017)

Special Issue

Paper Code	Paper Title	Page No.
IJETCAS 17-S101	Smart Notice Board Using ARDUINO Aakash Banerjee, Srinatha Mishra, Dr. Kanika Kaur	01-05
IJETCAS 17-S102	Implementation of 4-Bit Shift Registers using Diode Free Adiabatic Logic Shazia Pervez, Manisha Sahoo, Dr. Arti Noor	06-10
IJETCAS 17-S103	Implementation of Low Power Noise-Aware Power Gated Circuit Ankita Gupta and Arti Noor	11-16
IJETCAS 17-S104	Traffic Counting and Classifier using Single Loop Method for Non-lane Based, Mixed Traffic Flow Condition Hemant Jeevan Magadum and Ravikumar P	17-22
IJETCAS 17-S105	Layout techniques & matching strategies for CMOS analog integrated circuits Chetali Yadav, Sunita Prasad, M. Bharath Reddy and Manoj Kuma	23-28
IJETCAS 17-S107	FPGA Implementation of Sensor Fusion Technique for Obstacle Detection Danish, Dr. Sunita Prasad	29-34
IJETCAS 17-S108	Implementation of Cognitive Neuroscience Neuron Cell using Adaptive Velocity Threshold Particle Swarm Optimization 4-Bit Addition Divya Singh, Dr Sunita Prasad	35-40
IJETCAS 17-S109	Interfacing of ADC 0809 with FPGA Development Board Tushar Puri, Hemant Kaushal	41-46
IJETCAS 17-S110	Design and Implementation of Population Based Ant Colony Optimisation Algorithm Namrata Prakash and Dr. Sunita Prasad	47-51
IJETCAS 17-S111	VHDL Implementation of AES with Random Delay to Resist the Power Attack and To Confuse With DES Arvind Kumar Singh, BM Suri ² and SP Mishra	52-56
IJETCAS 17-S113	Design and Implementation of FPGA Based USB 2.0 Controller Bhavya, Niharika	57-62
IJETCAS 17-S201	Dimensionality Reduction in Big Data: A Survey Suhani and Nidhi Jain	63-68
IJETCAS 17-S202	Water Quality Monitoring using Data Mining Techniques Mitali Kathpal, Kriti Saroha	69-74
IJETCAS 17-S203	Secure Virtualization Environment with The Aid of SELinux Soumya Bhowmik	75-79
IJETCAS 17-S204	Analysis on Road Accidents Data to Improve Road Safety Nidhi Kalra, Kriti Saroha	80-83
IJETCAS 17-S206	Security of Data Store for E-Commerce Portal Nidhi Gahlot and Sanjay Ojha	84-87
IJETCAS 17-S207	An Efficient Plagiarism Detection System using Boyer Moore Algorithm Harsha Gupta and Sanjay Ojha	88-91
IJETCAS 17-S209	Text Extraction and Recognition from Images Priti Gangania ¹ and Tushar Patnaik	92-95
IJETCAS 17-S210	Big Data Fusion: A Survey SumedhaSeniaray ¹ and Nidhi Jain	96-100
IJETCAS 17-S211	Weather Prediction (Rainfall) Using Multiple linear regression along with adjusted R² Gunpreet Singh and Kriti Saroha	101-105

IJETCAS 17-S212	A Brief Survey on Detection of Wormhole Attack in MANET Jyoti shokhanda and Rekha saraswat	106-109
IJETCAS 17-S213	Efficient Detection of Black Hole in Mobile Adhoc Networks Deepak Sharma and Munish Kumar	110-115
IJETCAS 17-S214	Emotion Detection through Facial Expressions Sanika Singh, Tushar Patnaik	116-120
IJETCAS 17-S215	Hand Gesture Recognition Anushka Sharma and Tushar Patnaik	121-124
IJETCAS 17-S216	Recognition of Fruits and Vegetables from Images Shanam Afzal and Tushar Patnaik	125-127
IJETCAS 17-S217	Brief Overview & Comparison of Various Energy Aware Routing Protocols in MANET Sameeksha Kukreti and Rekha Saraswat	128-130
IJETCAS 17-S218	Sentiment Analysis of Social Media and Web Data using Machine Learning ShivaKarthik S, Lovey Joshi, Krishnanjan Bhattacharjee, Swati Mehta, Ajai Kumar	131-137
IJETCAS 17-S219	A Brief Survey on Various Spin Protocols In Wireless Sensor Network Sunita Kumari and Rosy Verma	138-140
IJETCAS 17-S220	A Survey of Centralized Key Management Schemes for Secure Multicast Communication Ekta Garg and Vinod Kumar	141-146
IJETCAS 17-S221	PairTester: Pairwise Generation of Test Cases in presence of constraints Reetika Gupta and Neha Bajpai	147-152
IJETCAS 17-S222	EFSM Slicer: Generation of Amorphous Slices based on Method Dependencies using EFSM Aakash Gautam and Neha Bajpai	153-158
IJETCAS 17-S223	XSS-VDetector: XSS VULNERABILITY DETECTION IN WEB APPLICATIONS Tannu Rohela and Neha Bajpai	159-163
IJETCAS 17-S224	Web Application Security: An Integral part of Web Application Development Life Cycle Pragya Sharma, Priyesh Ranjan, Praveen Kumar Srivastava	164-169
IJETCAS 17-S225	BIG DATA: A Survey paper on Recommendation System Alok Barddhan ¹ and Nidhi Jain	170-174
IJETCAS 17-S226	Restoration of Mural Images Gunjan Mishra and Tushar Patnaik	175-179



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Smart Notice Board Using ARDUINO

Aakash Banerjee¹, Srinatha Mishra², Dr. Kanika Kaur³

^{1,2}Final Year Students, ECE Department, KIIT, Gurgaon-122102, Haryana-India

³Associate Professor, ECE Department, KIIT, Gurgaon-122102, Haryana-India

Abstract: Notice board is used in institution or organization or public utility places like College campus, school campus, office areas, railway stations, etc. But the traditional ways of giving or sending various notices day to day is a hectic and tiring process. This paper deals with an advanced E-Notice Board. Our proposed system will enable authorized people to transmit notices on a notice board wirelessly using GSM with their 2g phone and users get notification about the notice. Its operation is based on Arduino Uno having micro controller ATMEGA328P and programmed in C language. When the user sends notice via registered phone simultaneously that message will get display on the notice board and also through the parse cloud other users get auto notification on their smart phone. We can also make the system compatible with more than one wireless technology and can be displayed to large dot matrix displays.

Keywords: Arduino, GSM module, authorized personnel

I. INTRODUCTION

To understand the proposed E-Notice Board on cloud platform, we have developed following prototype. It consists of Arduino board, controller ATmega 328P, GSM Module, GSM module and LCD display board. LCD display board is used for testing the proposed model. The interfacing of an Arduino board with GSM module is quite easy with help of the terminal pin, read/write pin. Hence we employ Atmel ATmega328p microcontroller [1]. The complexity of coding of our proposed system is less as compared with PC, but once programmed the micro controller works at its best. The design procedure involves identifying the different components and assembling all of them and it makes proper communication. Then coding process has to be done, which has to take care of the difference between two successive communications and most important of authentication of the senders number.

II. Problem Definition

It is a long process to put up notices on the notice board. This wastes a lot of resources like paper, printer ink, man power and also loss of time. In this paper we have proposed a system which will enable people to wirelessly transmit notices on notice board using GSM. Here we have proposed a system by which only authenticated person can handle the notice board. It require less time due to fast data transmission through GSM turn into cost effective device and save the resources like paper.

In this paper we have discussed about the design and Implementation of Digital notice board by using ARDUINO. The main objective of this system is to develop a wireless notice board that display message sent from the user and to design a simple, easy to install, user friendly system, which can receive and display notice in a particular manner with respect to date and time which will help the user to easily keep the track of notice board every day and each time he uses the system.

III. Scope of Proposed Work

The main aim of this project is to save time and provide information as early as possible by displaying the message. It can be used for multiple purposes like we can share live share market news, we can display trains time table, we can even show time table of college and important information for students and for teacher in school and college etc.

By using multiple screens for displaying the big size advertising purpose and the contents on the screen is made up of several images files and broadcasting display information and also remotely control it. The broadcasting information such as road highways, subways, buses and bus station, train and train station, shopping malls, city squares, hospital, conference hall, colleges and schools for displaying notice for student information and displaying all institutional information for visitors and this same application in industry for displaying notices or useful information which has want to giving employees.

IV. HARDWARE REQUIREMENT

A. ARDUINO BOARD

The Arduino Uno is an ATmega328p microcontroller board. This board has 14 digital input/output pins (6 as a PWM outputs), 6 analog inputs, a 16 MHz crystal oscillator, a USB connection, a power jack, an ICSP header, and a reset button, as displayed in fig.1.

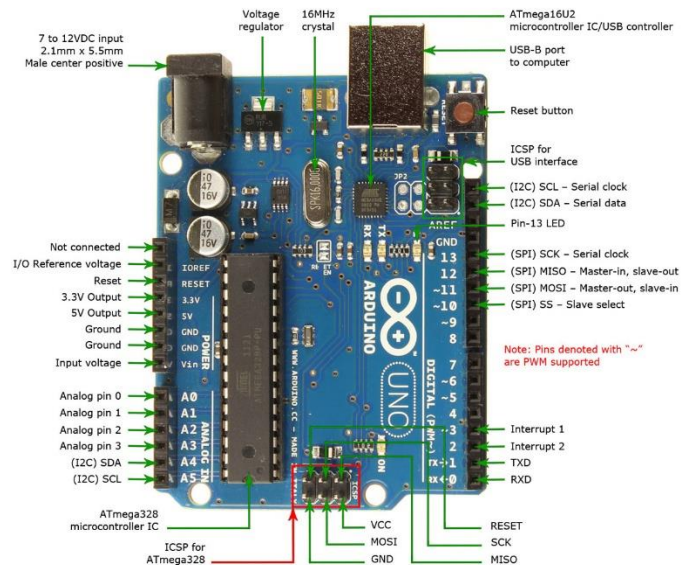


Fig. 1 Pin Diagram of ARDUINO Board

B. GSM MODEM

The GSM shield by Arduino is used to send/ receive messages and make/receive calls just like a mobile phone by using a SIM card by a network provider, displayed in fig. 2. We can do this by plugging the GSM shield into the Arduino board and then plugging in a SIM card from an operator that offers GPRS coverage. The shield employs the use of a radio modem by SIMComm. We can communicate easily with the shield using the AT commands. Mentioned in fig.3 The GSM library contains many methods of communication with the shield.

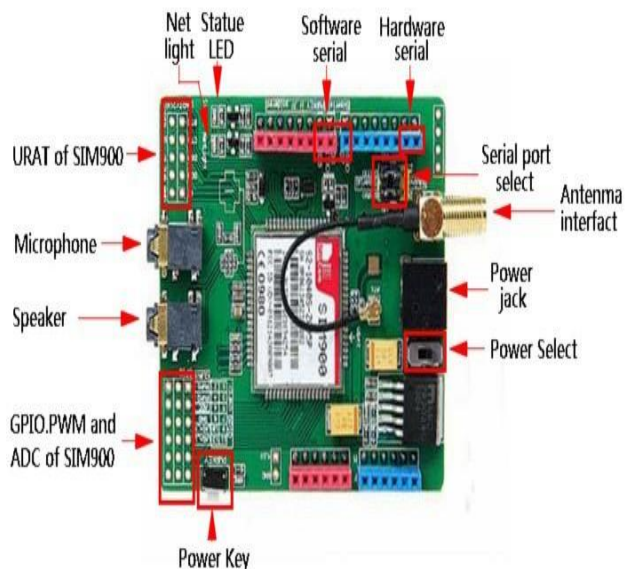


Fig.2 SIM900 GSM Modem

AT Command	The Functions of AT Command
ATD	Dial
AT+CGMS	Send SMS Message
AT+CMSS	Send SMS Message from storage
AT+CMGL	List SMS Messages
AT+CMGR	Read SMS Messages
AT+CSCA?	Service Centre Address
AT+CPMS	To choose storage from ME or SM
AT+IPR=0	To choose auto baud rate
AT+CMGF=	To choose PDU Mode or Text Mode

Figure 3 AT Commands

C. LCD MODULE

LCD stands for Liquid Crystal Display which is used to display text or Characters. We are using 14 pins LCD which are given below as per fig 4 :

- Pin 7 to Pin 14 All 8 pins are responsible for the transfer of data.
- Pin 4 This is RS i.e. register select pin. 5 This is R/W i.e. Read/Write pin.
- Pin 6 This is E i.e. Enable pin.
- Pin 2 This is VDD i.e. power supply pin.
- Pin 1 This is VSS i.e. source pin.
- Pin 3 This is short pin.

PIN NO	Symbol	Fuction
1	VSS	GND
2	VDD	+5V
3	V0	Contrast adjustment
4	RS	H/L Register select signal
5	R/W	H/L Read/Write signal
6	E	H/L Enable signal
7	DB0	H/L Data bus line
8	DB1	H/L Data bus line
9	DB2	H/L Data bus line
10	DB3	H/L Data bus line
11	DB4	H/L Data bus line
12	DB5	H/L Data bus line
13	DB6	H/L Data bus line
14	DB7	H/L Data bus line
15	A	+4.2V for LED
16	K	Power supply for BKL(0V)

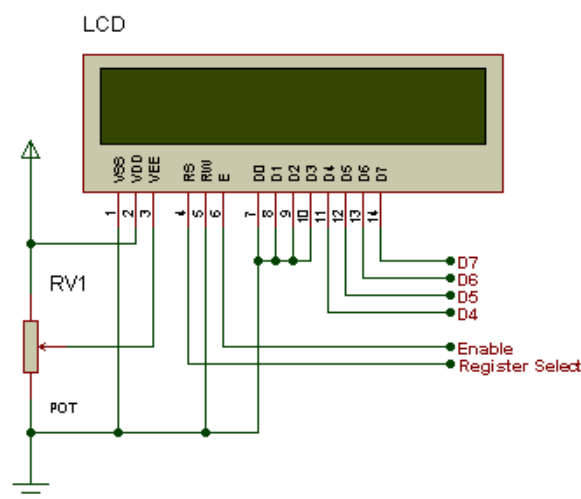


Fig.4. LCD Module

V. SOFTWARE REQUIREMENT

A. Arduino IDE

The open-source Arduino Software (IDE) makes it easy to write code and upload it to the board. It runs on Windows, Mac OS X, and Linux. The environment is written in Java and based on Processing and other open-source software.

This software can be used with any Arduino board. Embedded c programming is required to configure the Arduino board. This is a simple language and mostly used in most embedded system coding. Most of the new comers and well practiced coders use this language.

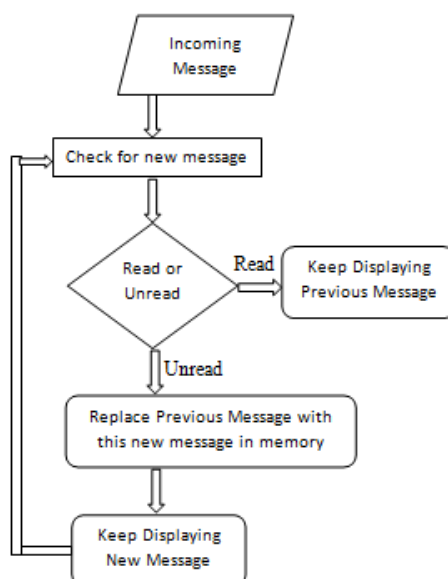
B. Proteus Design Suite

The Proteus Design Suite is an EDA tool, windows application for schematic capture, simulation, and PCB layout design. It can be purchased in many configurations, depending on the size of designs being produced and the requirements for microcontroller simulation. All PCB Design products include an auto router and basic mixed mode SPICE simulation capabilities.

VI. DATA FLOW DIAGRAM & WORKING

For making the notice boards easy to use and more technically advance, we have use this prototype of wireless notice board. We can display the message on LCD board by simply sending the message through our any 2G phone. These display systems are very accurate, easy to control, cheaply available and the most important thing is that they can be operated on low Voltage (Up to 12 Voltage).

A GSM module is used here for the wireless notice board to receive the information or message to display on board.



A. ARCHITECTURE

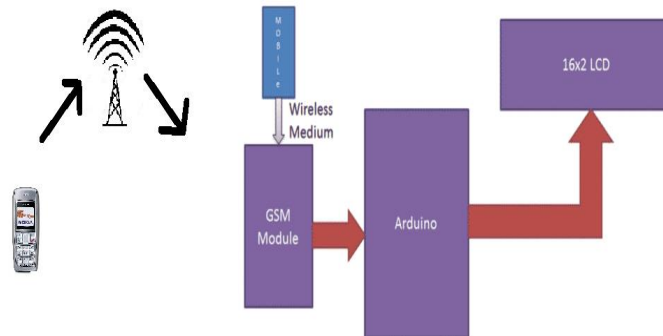


Fig.5 Notice Board Architecture

Fig 5 Describes the functioning of Smart Notice board, when the message to be display on the LCD is send from the mobile over the 900/1800 band, the GSM module receives the message.

Now the SIM900 GSM controller reads the message from the GSM modem and displays it on LCD.

When message is sent from the mobile, GSM modem sends the below command serially to the Arduino UNO and indicates the reception of the new unread text message.

+CMTI: "SM",3

In +CMTI: "SM",3 command, 3 indicates the location of the new message which is to be read and displayed on LCD. The command to read the message from GSM modem is

at+cmgr=3

Here 3 indicates the location of the message to be read. Once the above command is sent to the modem, the module acknowledge the command and in response ti this, it send the below command serially to the microcontroller ATmega328p.

+CMGR: "REC UNREAD","MOBILE NUMBER",,"DD/MM/YYYY, HH:MM:SS+34"

Urgent Announcement

The above command consists of "REC UNREAD" which tells that the message is unread, the sender's mobile number, date and time at which the message has been received.

"**Urgent Announcement**" is the text message.

So we need to extract the message from the above command in order to display it on the LCD.

VII. CONCLUSION

Arduino is a cheap and user friendly device and easy to handle and configure. Its IDE makes the process easier. LCD is used here for prototype, instead of it 8x8 LED matrix can also be used. LCD boards are used to display messages in Railway stations, shopping malls for displaying advertisement, Educational institution and organizations, managing traffic in smart cities and other public utility places. This prototype proves to be best and cost effective device.

REFERENCES

- [1] Li, J., Da-You, L., and Bo, Y., "Smart home research," Proceedings of the Third International Conference on Machine Learning and Cybernetics, vol. 2, pp. 659–663, Shanghai, 26–29 August 2004.
- [2] Choi, J., Shin, D., and Shin, D., Research on Design and Implementation of the Artificial Intelligence Agent Smart Home Based on Support Vector Machine, Berlin/Heidelberg: Springer, p. 417, 2005.
- [3] Li, B., Hathaipontaluk, P., and Luo, S., "Intelligent oven in smart home environment," International Conference on Research Challenges in Computer Science (ICRCCS '09), pp. 247–250, Shanghai, 28–29, December 2009.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

IMPLEMENTATION OF 4-BIT SHIFT REGISTERS USING DIODE FREE ADIABATIC LOGIC

Shazia Pervez¹, Manisha Sahoo², Dr. Arti Noor³

School of Electronics,

CDAC, Noida, India

¹spervez29@gmail.com, ²manishasahoo98@gmail.com, ³artinoor@cdac.in

Abstract. In digital circuits, shift registers are used as the basic memory units. This paper presents a low power adiabatic 32 bit serial in serial out (SISO) and serial in parallel out (SIPO) shift registers. Since power dissipation is a critical factor and most of the existing power reduction methods have too many trade-offs, Adiabatic logic technique can be considered as a promising method when compared to the conventional CMOS technique. In this paper, a type of adiabatic technique, DFAL (Diode Free Adiabatic Logic), has been studied and NOR gate, NAND gate, XOR gate, D Flip-Flop, SISO and SIPO shift registers have been designed using this configuration. DFAL circuits are analyzed based on transistor count, power dissipation and delay. All the circuits are simulated in Pyxis (Mentor Graphics) 180nm technology at 1.8V.

Keywords: adiabatic; low power; DFAL; shift register; SISO

I. Introduction

The ever increasing number of portable systems, the reduction in size of devices and the widespread usage of battery operated systems are emphasizing on the development of power and area efficient circuits. Low power techniques are becoming popular in circuits and systems nowadays [1]-[3]. Adiabatic Logic Circuits recycle the energy by returning energy to the source using a variable power clock supply instead of the fixed voltage, to reduce the overall power consumption.

Generally, two types of adiabatic circuits are studied i.e. Fully Adiabatic and Partially Adiabatic logic circuits [4]. In ideal situation, the occurrence of losses does not occur in case of fully adiabatic circuits. The energy that is being dissipated at the time of charging can be described in terms of T i.e. time period, Stored Charge ($C_L V_{DD}$), and Load capacitance (C_L) [10]. Yibin Ye and Kaushik Roy have compiled a detailed analysis and modelling of adiabatic-logic technique [4] and have observed a significant reduction in power dissipation using this technique. They have also introduced Quasi-static energy recovery logic family (QSERL) using the principle of adiabatic switching [5] where two sinusoidal supply clocks (complementary) are used. P. Sasipriya, V.S. KanchanaBhaaskaran have presented a comprehensive analysis and evaluation of ASL (Adiabatic Static logic), QSERL (Quasi Static Energy Recovery Logic), CEPAL (Complementary Energy Path Adiabatic Logic) and QSECR (Quasi Static Efficient Charge Recovery Logic) logic [6]. Due to diode connected MOSFET in QSERL there is degradation of output amplitude. This problem was overcome by Sanjay Singh, K. Srinivasarao [7]. They have introduced Diode Free Adiabatic Logic (DFAL) which also reduces the delay and circuit complexity. Different circuits have been designed using DFAL [8], [9]. This paper presents 32 bit SISO and SIPO Shift Registers. SISO shift register is a type of register that can perform shift left operation and shift right operation. SIPO shift register performs serial to parallel operation. It is commonly used in applications where conversion from serial to parallel interfaces is required.

The remaining part of this paper is described as follows. Section II is a detailed study on adiabatic logic and section III describes the shift registers. In section IV the various schematics and their analysis are presented and the conclusion is stated in section V.

II. Adiabatic Logic

A. Adiabatic Technique

Adiabatic technique is a low power technique which promises no power dissipation in ideal situations (asymptotically zero power dissipation). The main feature of this technique is that the used supply voltage is not a constant voltage. The supply is also called the power clock. The variable power supply is implemented by a resonant LC circuit that reclaims the energy which is stored in the capacitor and reuses it in the succeeding cycle. The power clock can be sinusoidal, trapezoidal or triangular. According to the type of the power clock the circuit operates in different modes.

To charge a given node which is associated with a capacitance C_L from 0 to V Volts in conventional circuit (CMOS), $VQ (=C_L V^2)$ of energy from the supply is being derived (as shown in Fig. 1). 50% of energy, $0.5C_L V^2$, is stored in the capacitance, and the remaining is dissipated through resistance in the path. Energy dissipates with a scale of $i\Delta V$ (instantaneous power being dissipated) whenever there is a voltage drop ΔV (stands for current). This type of dissipation in energy can be decreased to a great extent by considering adiabatic technique.

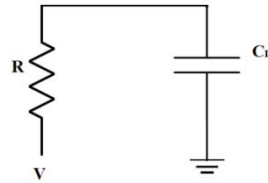


Fig. 1. Adiabatic Switching

Hence, the power dissipation in the static CMOS and adiabatic circuits are described as follows:

$$\text{Static CMOS: } C_L V^2 \quad (1)$$

$$\text{Adiabatic Logic: } (2RC_L/T) C_L V^2 \quad (2)$$

From the above equations it is obvious that the power dissipation in case of adiabatic circuits depends on the ratio RC_L/T (R stands for resistance, T denotes clock phase period). So if R , C_L and T are adjusted such that $RC_L \ll T$ then the total power dissipation will be

$$(RC_L/T) C_L V^2 \ll 0.5 C_L V^2 \quad (3)$$

It is suggested that the logic gates designed using adiabatic technique show that at 1MHz frequency of operation approximately 90% of the total power consumed can be reused[4]. Adiabatic circuits show better results in the low frequency range. In this paper we have focused on DFAL (Diode Free Adiabatic Logic).

B. DFAL (Diode Free Adiabatic Logic)

DFAL circuits do not have any diode connected MOSFETs in the charging or discharging path. Adiabatic circuits such as the Complementary Energy Path Adiabatic Logic, Two phase clocked static CMOS logic, and quasi-static energy recovery logic have the following problems:

- Delay
- Complex circuit
- Degradation of output amplitude.

DFAL can be used to overcome the above shortcomings [7]. In DFAL, split level sinusoidal complementary power clock supply V_{PC} and V_{PCbar} are used. The voltage level of V_{PC} is taken such that it is $V_{PC}/2$ greater than that of V_{PCbar} ; this tends to decrease power dissipation. The DFAL inverter shown in Fig. 2 is similar to the conventional CMOS logic; however its operation differs as it operates in an adiabatic way.

The M3 transistor of the pull down path is used to replace the diode of the discharging path. V_{PC} controls the turning ON and OFF of M3. M3 recycles the charges at the output node therefore the adiabatic losses are further recovered. In evaluation phase, when the p MOS tree is ON and output node is at logic 0, load capacitance charges through transistor M1 so output goes to logic 1. When n MOS tree turns ON and output node is at logic 1, M2 and M3 discharge and recycle the charges towards the power clock (V_{PC}), resulting in the output to be at logic 0. During hold phase, no transitions occur which reduces dynamic switching and hence energy dissipation.

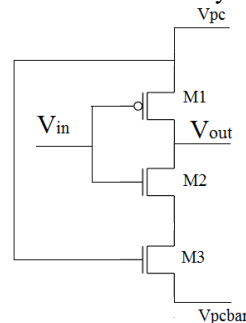


Fig. 2. DFAL Inverter

III. Shift Register

A. Serial In Serial Out (SISO) and Serial In Parallel Out (SIPO) Shift Registers

In SISO Shift registers the data inputs provided and the outputs obtained are both serial in nature. An n bit positive edge triggered synchronous SISO is depicted in Fig. 3. It is constructed using n D- flip flops. All the flip flops are initially in the clear state such that all outputs become zero. The data input is provided one bit at a time to the input D0 of the flip flop F1. The output of each flip flop is given as the input to the next flip flop. Finally the stored data can be retrieved from the last flip flop. All the flip flops are controlled synchronously by the same clock. One bit of data is obtained at the output at every positive edge of the clock. A SIPO shift register is similar to the SISO shift register as the data is shifted into the individual flip flops and consequently the output is obtained serially from the output port. On the other hand it is different from a SISO register as the outputs of all the intermediate flip flops are made available. Finally the stored data can be retrieved as an entire parallel word through the outputs of each flip flop. Hence, serial in parallel out registers are used for serial to parallel transformation of the data. Time delays in various digital circuits are provided by the serial in – serial out shift registers.

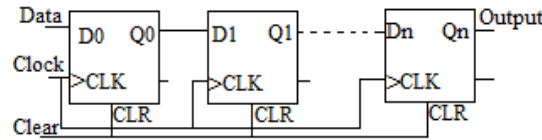


Fig 3. SISO Shift Register

IV. Observations And Results

The DFAL inverter is designed according to the circuit diagram described above and its waveform is shown in Fig.4. It can be clearly analyzed from the waveform that the output follows V_{pc} in the evaluation period and V_{pcbar} in the hold phase. The comparative study is depicted in Table 1. The percentage reduction in power consumption of the DFAL inverter is approximately 95%.

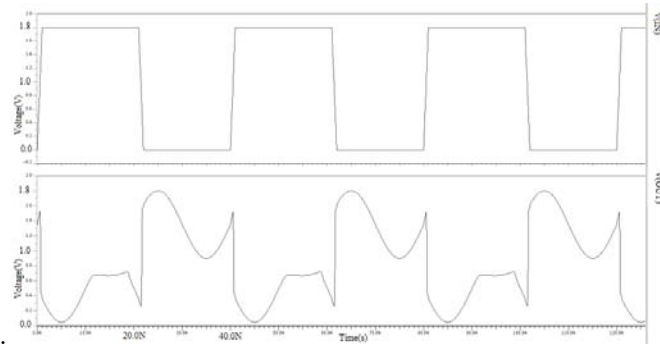


Fig.4. DFAL Inverter Waveform (X axis:Time in seconds, Y axis:Voltage in Volts)

Table 1.Comparison Of Inverters

Type of Logic	CMOS	DFAL
Power(pW)	15.58	0.73
Delay(pS)	24.10	176.81
PDP(J)	0.375e-21	0.129e-21
No of Transistors	2	3

The DFAL 2-input NOR is also designed and the comparative study is recorded in Table 2. The power consumed by the DFAL 2 input NOR is approximately 95.30% less than the conventional CMOS NOR gate.

Table 2.Comparison Of NOR

Type of Logic	CMOS	DFAL
Power(pW)	31.16	1.46
Delay(pS)	87.11	63.92
PDP(J)	2.714 e-21	0.093 e-21
No of Transistors	4	5

The waveform for DFAL 3 input NAND gate is depicted in Fig.5. The comparative analysis is depicted in Table 3. The power consumed by the DFAL 3 input NAND gate is approximately 90.64% less than the conventional CMOS 3 input NAND gate at 1.8V.

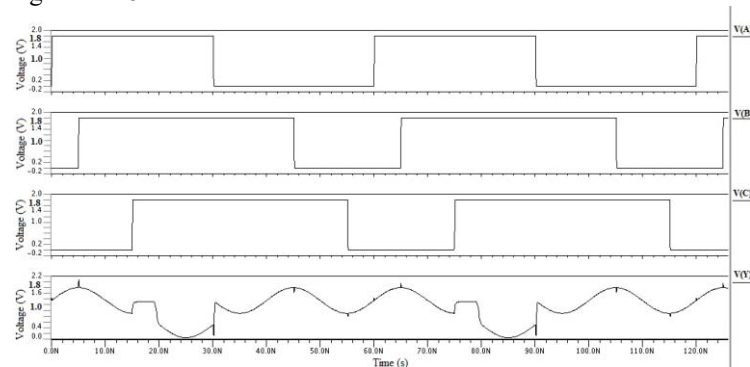


Fig.5. DFAL 3-Input NAND Waveform (X axis:Time in seconds, Y axis:Voltage in Volts)

Table 3.Comparison Of 3-Input NAND

Type of Logic	CMOS	DFAL
Power(pW)	4.14	0.39
Delay(pS)	76.86	192.50
PDP(J)	0.318 e-21	0.074 e-21
No of Transistors	6	7

The DFAL 2 input XOR is designed and its comparative analysis is also done in a similar way. The power consumed by the DFAL 2 input XOR is observed to be approximately 34.32% less than the conventional CMOS XOR gate.

The DFAL D Flip Flop is designed as shown in Fig. 6 and its waveform is shown in Fig. 7. The comparison is recorded in Table 4. The power consumed by the DFAL D Flip Flop is approximately 82.59% less than the conventional CMOS D Flip Flop.

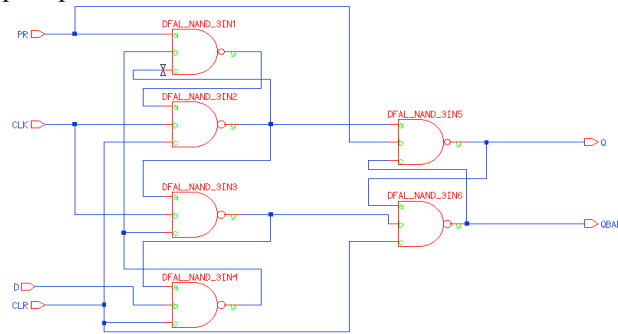


Fig. 6.DFAL D flip flop Schematic

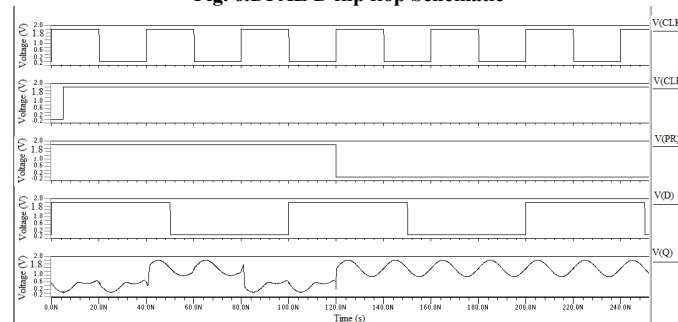


Fig.7.DFAL D flip flop Waveform (X axis:Time in seconds, Y axis:Voltage in Volts)

Table 4.Comparison Of D flip flop

Type of Logic	CMOS	DFAL
Power(pW)	86.83	15.11
Delay(pS)	395.88	1239.80
PDP(J)	34.375 e-21	18.733 e-21
No of Transistors	36	42

The DFAL 32 bit SISO and SIPO Shift Register are also designed and their waveforms are obtained and verified. The waveform for 32 bit SISO is shown in Fig. 8. The comparison is recorded in Table 5. The power consumed by the 32 Bit SISO and SIPO Shift Register is approximately 82.33% less than the conventional CMOS 32 Bit SISO Shift Register at 1.8V.

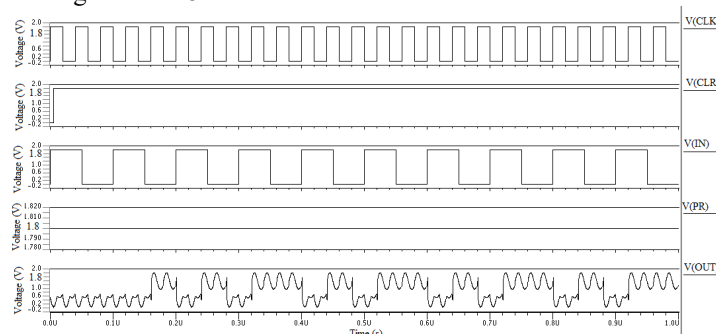


Fig.8. DFAL32 Bit SISO Shift Register Waveform

Table 5.Comparison Of 32 bit Shift Registers

Type of Logic	CMOS	DFAL
Power(pW)	2798	494.34
Delay(pS)	395.87	820.38
PDP(J)	1107.644 e-21	405.553 e-21
No of Transistors	1152	1344

V. Conclusion

The above results show that the different logic gates, D Flip Flop, SISO and SIPO shift registers implemented using DFAL logic consume much less power as compared to the conventional circuits (CMOS circuits). However, the delay and transistor count based comparison of the circuits designed using DFAL and the corresponding conventional circuits show that these parameters have larger values in case of DFAL circuits. Hence DFAL can be preferred in areas where low power is the main concern as the difference in power consumption is much more when compared to the difference in transistor count and delay. It was observed that in all the circuit the PDP is also less in the DFAL circuits.

Shift registers are used in computers as memory elements. A large amount of data has to be stored in all types of digital circuits and systems that too in an efficient manner so, there is the need to use storage components like RAM and different registers.

Adiabatic switching is an emerging low power technique; however it has not been implemented in the main stream VLSI domain [8]. This technique is suitable for operation in the low frequency range. Other Adiabatic logics may be developed using the primitive logics. The adiabatic techniques can also be embedded into MEMS and thereby explore the technology of AdiaMEMS and Nanotechnology in the near future.

References

- [1] D. Soudris, V. Pavlidis, and A. Thanailakis, "Designing low-power energy recovery adders based on pass transistor logic," ICECS 2001. The 8th IEEE International Conference on Electronics, Circuits and Systems, 2001. Volume 2, pp. 777-780, Sept 2-5, 2001.
- [2] Hiroaki Suzuki, WoopyoJeong, and Kaushik Roy, "Low-Power Carry-Select Adder Using Adaptive Supply Voltage Based on Input Vector Patterns," ISLPED'04 Proceedings of The 2004 IEEE/ACM International Symposium on Low Power Electronics and Design, 2004. pp.313-318, Aug 9-11, 2004.
- [3] N V Vijaya Krishna Boppana, Saiyu Ren, Henry Chen, "Low-Power and High Speed CPL-CSA Adder," NAECON 2014. IEEE National Aerospace and Electronics Conference, 2014. pp. 346-350, June 24-27, 2014.
- [4] Yibin Ye and Kaushik Roy, "Energy Recovery Circuits Using Reversible and Partially Reversible Logic," IEEE Transactions On Circuits And Systems-I: Fundamental Theory And Applications, vol. 43, no. 9, pp. 769-778, September 1996.
- [5] Yibin Ye and And Kaushik Roy, "QSERL: quasi-static energy recovery logic," IEEE Journal of Solid-State Circuits, vol. 36, no. 2, pp. 239-248, 2001.
- [6] P.Sasipriya, V.S. KanchanaBhaaskaran, "Two Phase Sinusoidal Power Clocked Quasi-Static Adiabatic Logic Families," IEEE Eighth International Conference on Contemporary Computing (IC3), pp. 503-508, Aug 20-22, 2015.
- [7] Sanjay Singh, K. Srinivasarao, "Implementation of 4-bit carry select adder using Diode Free Adiabatic Logic (DFAL)," IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), pp. 481-484, July 9-11, 2015.
- [8] Nitish, N. Pandey, R. Pandey and K. Gupta, "DFAL based implementation of frequency divider-by-3," 2015 Annual IEEE India Conference (INDICON), New Delhi, 2015, pp. 1-6.
- [9] K. Srinivasarao and G. Kumar, "Implementation of barrel shifter using diode free adiabatic logic (DFAL)," 2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE), Coimbatore, 2014, pp. 1-5.
- [10] Gary K. Yeap, "Practical Low Power Digital VLSI Design", KAP, 2002.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Implementation of Low Power Noise-Aware Power Gated Circuit

Ankita Gupta¹ and Arti Noor²

School of Electronics (SOE),
Centre for Development of Advance Computing, Noida,
Uttar Pradesh, India,

Abstract: Due to technology scaling, critical dimensions of semiconductor devices continue to shrink. Power gating technique is well known technique to suppress leakage current in logic circuits during standby mode. MTCMOS, also known as power gating technique is very well known leakage power reduction strategy. However, conventional MTCMOS technique inserted ground bounce noise during mode transition from SLEEP to ACTIVE mode which degrades circuit reliability. Thus, ground bounce noise becomes a major hurdle in standard MTCMOS circuits. In this paper we analyzed ground bounce noise and explored different noise minimization techniques. An intermediate relaxation mode is introduced between SLEEP to ACTIVE mode to suppress ground bounce noise. The comparison of ground bounce noise based on sleep transistor size, controlling transistor size and temperature of different MTCMOS techniques are explained in this paper. All these techniques are applied on True Single Phase Clock (TSPC) D flip flop. Parametric analysis based on different parameters such as controlling transistor size, sleep transistor size, temperature, with conventional, trimode, dual-switch and tri-transistor MTCMOS techniques are evaluated with 180-nm CMOS technology.

Keywords: Ground Bounce Noise, Power Gating, mode transition, intermediate mode, MTCMOS, flip flop.

I. INTRODUCTION

High power utilization in versatile hardware gadgets is an issue of significant concern. Shortening of battery life and extra bundling and cooling prerequisites are related with high power utilization. Static power consumption because of standby leakage currents is a vital element of total power consumption. Wide-ranging electronics devices contain differing kinds of element of that several stay idle throughout a selected operation. Static power consumption occurring within these idle elements and leakage power consumption in active element represent an enormous proportion of total power consumption in the circuit. The reduction of this leakage power becomes critical for fruitful power management. As a consequence of continued scaling of MOS circuits, a sensational improvement in the performance of MOS circuit has been accomplished. This has prompted increment in power consumption because of leakage current. Till now, sub-threshold leakage current is the root cause of the total power dissipation [1].

Multi-threshold CMOS is very well known leakage power reduction strategy [2]. In order to achieve high performance, power gating exploits the multi-threshold voltage which incorporates nMOS channel and pMOS channel with two distinct threshold voltages in a single integrated circuit. Low threshold voltage (low V_{th}) circuits are used on critical delay paths to achieve high speed and high-threshold-voltage (high V_{th}) devices are used on non critical paths to reduce leakages. Either a high V_{th} pMOS transistor or a high V_{th} nMOS transistor can be used to implement a sleep switch. A pMOS sleep transistor (header) is added in the middle of an actual power rail and a virtual power rail and an nMOS sleep transistor (footer) is added in the middle of an actual ground rail and a circuit ground rail, as shown in Fig.1. The pMOS and the nMOS sleep transistors are switched off to minimize the sub-threshold leakage currents during the period of inactivity.

II. RELATED WORK

This section deals with previous research works that were used to reduce ground bounce noise that is produced during mode switching and to lower leakage power consumption are discussed below.

In [3], this paper is concerned with the novel power gating techniques to lower the ground bounce noise. In this paper author employed the holistic integrated device circuit-architecture approaches for power gating techniques. It elaborates that during mode transition, sleep transistor controls the amount of current in the intermediate mode and sustain the power supply voltage.

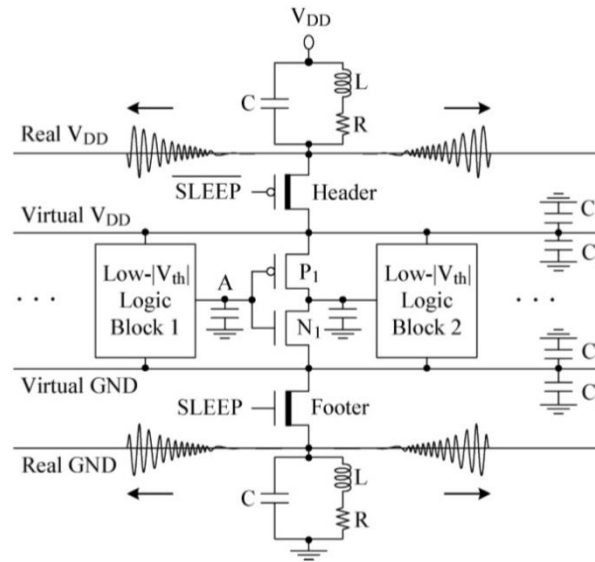


Figure 1. Conventional MTCMOS circuit [5]. Thick line highlights the High- V_{th} sleep transistors in the channel.

This novel power gating structure facilitates in reduction of ground bouncing noise by switching the sleep transistor on, in a non uniform stepwise and pseudorandom fashion.

In [4], new sequential MTCMOS structure with data holding sleep mode is used. It also contain smaller integrated sleep transistor to minimize the ground bounce noise. Comparison of different sequential MTCMOS structure according to leakage power, area overheads, ground bounce noise is tabulated.

III. DIFFERENT NOISE MINIMIZATION MTCMOS TECHNIQUES

In this section different noise minimization techniques are discussed. Section III-A discusses about trimode MTCMOS techniques. The dual switch MTCMOS technique is reviewed in Section III B. The tri transistor MTCMOS technique is introduced in Section III C.

A. Trimode MTCMOS Technique

The trimode power gating structure [5] is examined here. In the middle of the SLEEP and the ACTIVE modes an additional intermediate PARK mode is investigated to suppress the ground bouncing noise.

As shown in Fig. 2, the footer sleep transistor (N_1) is connected in parallel with high- V_{th} pMOS transistor (Parker).

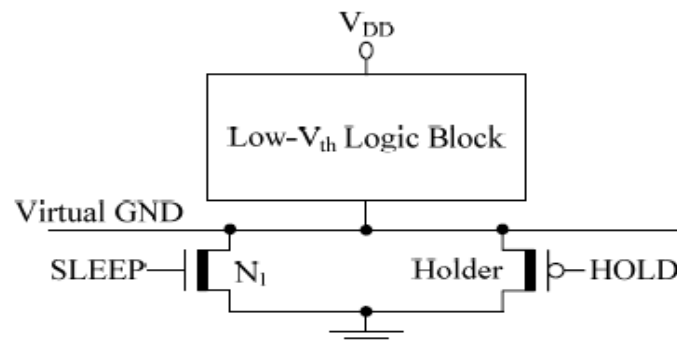


Figure 2. Tri-mode power gating technique [5]

When circuit is inactive Parker and N_1 are switched off to minimize sub-threshold leakage current. When Parker is turned on circuit changed to Parker mode and virtual ground line shifts from $\sim V_{DD}$ to threshold voltage of Parker (V_{tp}). By targeting oscillation of voltage, it provides a means to regulate ground bounce noise as shown in Fig. 2. The footer is switched on to attain the fully awakening process. After that the Parker is switched off. Circuit is fully activated when virtual ground line is expelled to $\sim V_{gnd}$. In this technique, there is no need of complex circuitry for regulating the operation of the sleep transistors.

B. Dual Switch MTCMOS Technique

As shown in Figure 3 in the middle of the actual power supply and the virtual power rail a high V_{th} nMOS transistor (N_2) and a header sleep transistor (P_1) are connected in parallel Likewise, N_1 and P_2 are connected in

parallel in the middle of the actual ground and the virtual ground rail (same as the PARK mode in [5]) an intermediate HOLD mode is switched on however the footer and header are remained cut-off. At the time of SLEEP mode P_1 , P_2 , N_1 , N_2 are switched off to minimize the subthreshold leakage currents. Before the triggering of the circuit, P_2 and N_2 are turned on. The circuit switches from the SLEEP mode to an in-between HOLD mode. As a result, differential voltage of $V_{DD} - V_{tn} - V_{tp}$ is generated in the middle of the virtual lines. Later, P_1 and N_1 are triggered to switch from the HOLD mode to the ACTIVE mode. The virtual power line is excited to $\sim V_{dd}$. The virtual ground line is expelled to $\sim V_{gnd}$. As circuit switches from SLEEP mode to the ACTIVE mode through intermediate HOLD mode, the ground bounce noise is lowered because of maximum reduction in the voltage swings on the virtual lines.

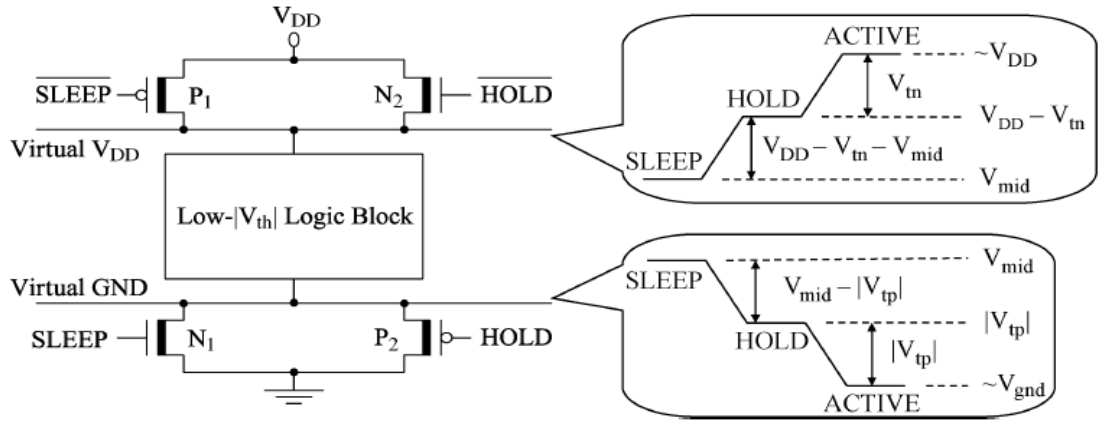


Figure 3. Dual-switch power-gating structure proposed in [5].

C. Tri Transistor-Controlled MTCMOS Technique

An additional improved form of the standard gated- V_{DD} & ground MTCMOS structure is initiated to reduce the ground bouncing noise during the mode transitions. Another high $|V_{th}|$ pMOS sleep transistor (Dozer) and footer are connected in parallel to execute an intermediate DOZE mode, as shown in Fig. 3 [5].

The header and Dozer are switched on during DOZE mode. The footer is remained off. In the ACTIVE mode, the footer and header are both switched on. Subsequently, the DOZE mode would be initiated as an intermediate state to reduce ground bouncing noise during the transition from SLEEP to ACTIVE mode. During the switching from the SLEEP mode to the ACTIVE mode, the virtual ground line is expelled in two steps as shown in Fig.4 [6].

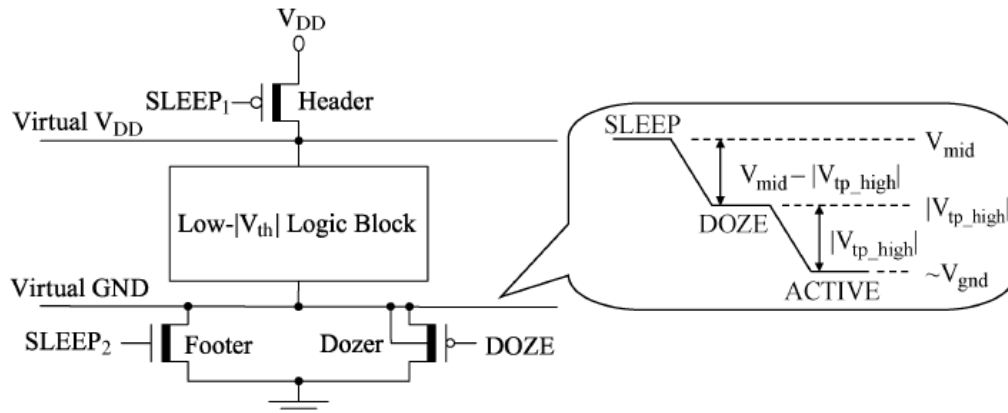


Figure 4. Tri-transistor-controlled MTCMOS circuit technique presented in [6].

IV. SIMULATION AND RESULTS

The TSMC 180-nm multithreshold-voltage CMOS technology with low- $V_{thOP} = -300$ mV, high- $V_{thOP} = -500$ mV, low- $V_{thON} = 300$ mV, high- $V_{thON} = 500$ mV and $V_{DD} = 1.8V$ is used for the parametric analysis of ground bounce noise, active power dissipation and leakage power consumption with various noise minimization MTCMOS techniques. True single phase clock (TSPC) D flip flop is designed based on following techniques as shown in Fig 5. Simulation of MTCMOS circuit is done with Mentor graphics tool.

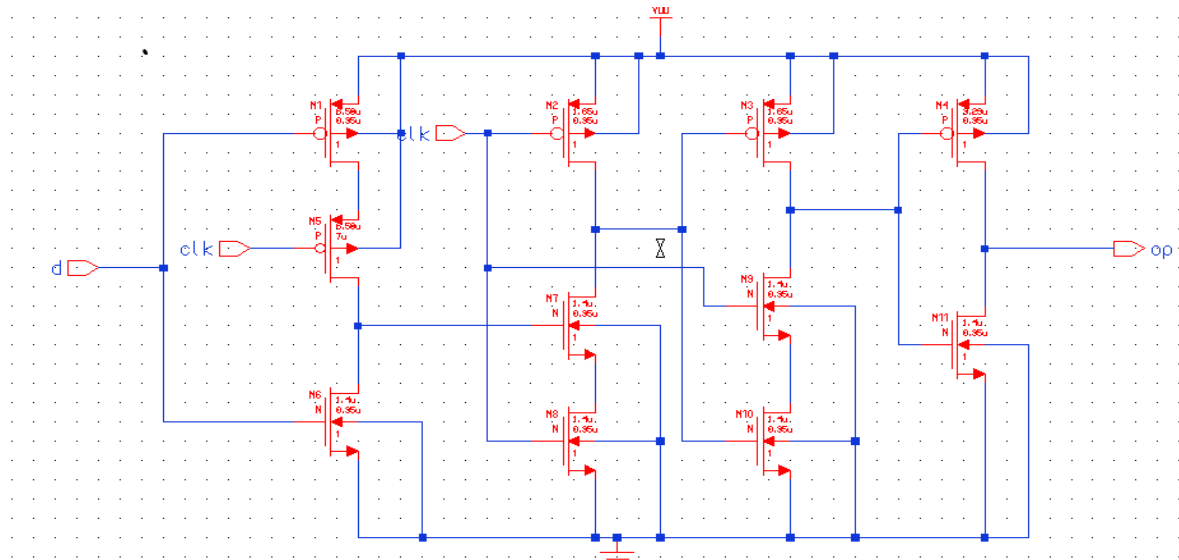


Figure 5. TSPC D FLIP FLOP

A. GROUND BOUNCING NOISE

The peak amplitude of ground bounce noise produced by various techniques while mode transition is induced on real ground line. The percentage reduction in ground bounce noise generated with trimode, dual-switch and tri-transistor structure as compared to conventional MTCMOS at 25°C, 70°C and 110°C. Ground bounce noise plot of conventional MTCMOS circuit is shown in figure 6.

Furthermore, by up to 94.14%, 94.69%, 94.34% and 94.99% reduction in ground bounce noise is achieved by trimode, dual switches, TTH and TTL respectively while compared to conventional MTCMOS circuit during sleep to active mode at 70°C. On the other hand, approximately 93.98%, 94.31%, 94.46%, 94.84% and 95.87% reductions in ground bouncing noise are acquired by trimode, dual switches, TTH and TTL respectively D flip flop as compared to conventional MTCMOS circuit at 110°C as shown in Table I and Table II.

TABLE I COMPARISON OF PEAK AMPLITUDE OF THE GROUND BOUNCING NOISE (mV)

Mode Transition	Sleep Mode to Active Mode					
Stored Data	0					
Temperature	25°C		70°C		110°C	
Transistor size(μm)	0.12	28.8	0.12	28.8	0.12	28.8
Conventional	441.59		422.93		407.80	
Trimode	29.399	25.078	29.240	25.176	29.583	25.596
Dual-switch	26.080	23.463	24.782	23.267	26.677	24.642
TTH	26.736	26.525	26.084	25.163	25.176	23.655
TTL	23.729	21.298	23.925	21.228	24.112	21.323

TABLE II COMPARISON OF PEAK AMPLITUDE OF THE GROUND BOUNCING NOISE (mV)

Mode Transition	Sleep Mode to Active Mode					
Stored Data	1					
Temperature	25°C		70°C		110°C	
Transistor size(μm)	0.12	28.8	0.12	28.8	0.12	28.8
Conventional	457.98		409.00		410.97	
Trimode	21.739	21.611	21.815	21.185	21.739	21.611
Dual-switch	20.181	20.061	20.328	19.930	22.465	19.382
TTH	21.274	21.031	21.447	21.413	21.492	20.328
TTL	19.551	18.655	19.728	18.435	20.346	18.699

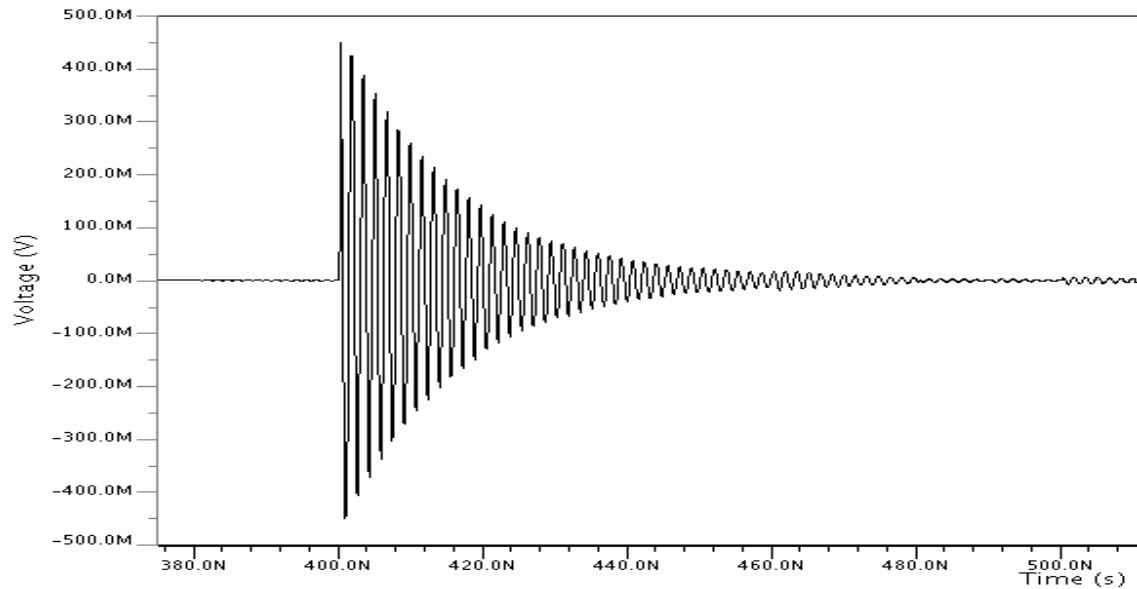


Figure 6. Ground bounce noise in TSPC D flip flop using conventional MTCMOS.

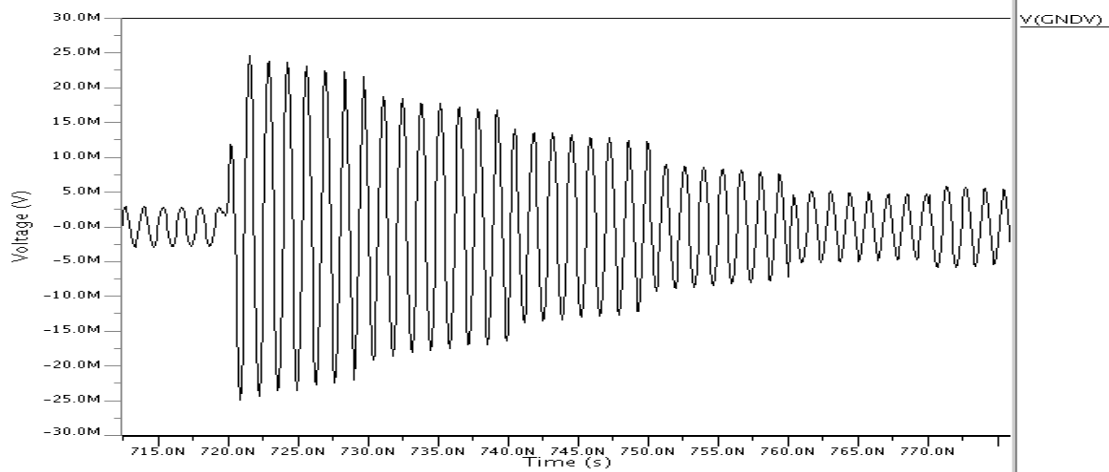


Figure 7. Ground bouncing noise in TTL logic.

a) IMPACT OF TRANSISTOR SIZE

Controlling transistor i.e. Parker in trimode, hold in dual-switch and dozer in tri-transistor influenced the charging and discharging speed and voltage of virtual line. With increase in controlling transistor width, it results in reduction in effective resistance of controlling transistor and reduction in voltage of the virtual ground line. Therefore, overall reduction in ground bounces noise as listed in Table I and Table II.

b) IMPACT OF TEMPERATURE ON GROUND BOUNCING NOISE

Temperature has lesser effect over noise than that of transistor size. However, at larger temperature, peak amplitude of ground bouncing noise shows inverse effect as its mobility is reduced. There is fluctuation due to other overriding factors.

V. CONCLUSION

Ground bouncing noise is major hurdle in conventional MTCMOS technique. This paper tried to solve the vary challenge related to conventional MTCMOS circuit. Various noise minimization MTCMOS structures are used to minimize the ground bouncing noise. In these techniques, a two step activation scheme is used to reduce the ground bouncing noise. A relaxation intermediate mode is introduced between SLEEP mode and ACTIVE mode due to which the range of voltage swing on the virtual rail is reduced to a greater extent. Because of this lower range of the voltage swing the magnitude of ground bouncing noise is reduced. TTL technique is proving to be the best technique among all noise minimization techniques to reduce ground bounce noise.

REFERENCES

- [1] K. Shi and D. Howard, "Challenges in sleep transistor design and implementation in low-power designs," in *Proc. ACM/IEEE Des. Autom. Conf.*, Jun. 2006, pp. 113–116.
- [2] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, "1-V power supply high-speed digital circuit technology," *IEEE J. Solid-State Circuits*, vol. 30, no. 8, pp. 847–854, Aug. 1995.
- [3] M. Pattanaik, V. L. Varaprasad and F. R. Khan, "Ground Bounce Noise Reduction of Low leakage 1-bit Nano-CMOS based Full Adder Cells for Mobile Applications," *IEEE International conference on electronics device, system and application*, pp. 31-36, 2010.
- [4] H. Jiao and V. Kursun, "Sleep Transistor Forward Body Bias: an Extra Knob to Lower Ground Bouncing Noise in MTCMOS Circuits," *IEEE International Symposium on Low Power Electronics Designs*, pp. 216-219, 2009.
- [5] H. Jiao and V. Kursun, "Ground-Bouncing-Noise-Aware Combinational MTCMOS Circuits," *IEEE Transaction on Circuits and System*, VOL. 57, NO. 8, pp. 2053-2065, August 2010.
- [6] H. Jiao and V. Kursun, "Sleep Transistor Forward Body Bias: an Extra Knob to Lower Ground Bouncing Noise in MTCMOS Circuits," *IEEE International Symposium on Low Power Electronics Designs*, pp. 216-219, 2009.

Traffic Counting and Classifier using Single Loop Method for Non-lane Based, Mixed Traffic Flow Condition

Hemant Jeevan Magadam¹ and Ravikumar P.²

Intelligent Transportation and Networking System (ITNS), CDAC

Vellayambalam, Thiruvananthapuram, India

Abstract: Classified volume counts are important for almost all traffic and transportation related applications and planning. Volume counts are normally collected through manual enumeration which is costly, time consuming and provides only limited samples. All weather permanent traffic counters are expensive. Dual loop counters performing length based classification can function well in lane following and homogeneous traffic condition. But it is highly complex in non-lane based mixed traffic flow condition. This paper presents an algorithm for classified volume counts in non-lane based, mixed traffic flow condition using single loop method. This algorithm was tested in real field at different locations to count and classify automobiles of different chassis lengths such as 2 wheeler, 3 wheeler, Car/ Jeep, LCV, bus and truck with moderate accuracy under heterogeneous and non-lane disciplined traffic condition. Experiments on single loop vehicle counting and classifiers are also being taken place worldwide. These methods give acceptable results under ideal condition. The proposed method can be easily implemented for permanent counting and classification of vehicles required for modelling and planning of traffic applications.

Keywords: Lane Following; Heterogeneous; Inductive Loop; Vehicle Count; Classified Count;

I. Introduction

Classified vehicle count is an important parameter for various ITS applications and traffic planning. Permanent traffic counting / classifiers are required for applications like tolling. It is comparatively easy to implement such systems as the vehicles cross toll points one after another at a very slow speed. Also, there will not be too many classes of vehicles crossing the toll point. Even camera based counting / classifiers are possible at these points as the lighting conditions can be controlled; cameras fail during different weather conditions. Traditionally, Inductive loop vehicle detectors are considered reliable in all weather and lighting conditions for vehicle counting and classification. With dual loop speed traps they give pretty good results provided the traffic is lane following and more or less homogeneous. Heterogeneous traffic refers to vehicles having different static and dynamic parameters such as two-wheelers, three-wheelers, cars, LCV, buses and trucks using the same road space. They are normally classified based on length. The traffic lanes are considered 3.6mtrs wide. Typically, one or more inductive loops of 2mtr x 2mtr size are used in every lane for vehicle counting and classification. An inductive loop is formed by three to four turns of 1.5mm sq. specially insulated copper cable laid in carriageway at a depth of 20-30mm in diamond cut slots. These loops are connected to the roadside electronics through lead-in cable. Presence or passage of the metallic parts of a vehicle over the loop will vary the inductance of the loop which is identified by the roadside electronics for various applications such as traffic signaling, access control, vehicle counting and classification etc.

In lane following traffic each vehicle follows the other intercepting the inductive loop one at a time. In such cases it is easy to find out the axle length of every vehicle by deploying speed trap. In the case of highly heterogeneous traffic having poor lane discipline it is possible that one loop can be intercepted by one or more vehicles; and one or more loops can be intercepted by one vehicle.

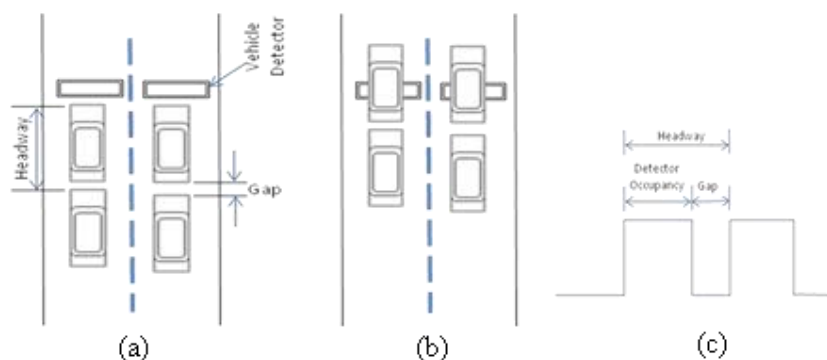


Figure 1: Traffic movement and detector output in disciplined traffic

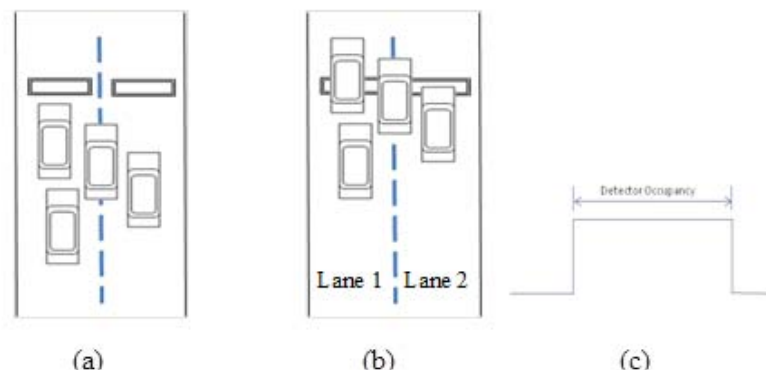


Figure 2: Traffic movement and detector output in poor lane disciplined traffic

Counting and classifying of vehicle under such conditions is a challenge. It is obvious that 100% accuracy cannot be achieved in such scenario. However, automatic vehicle counting and classifiers can be used in highly heterogeneous traffic having poor lane discipline where the accuracy can be compromised. Applications such as traffic planning, verification of pavement life expectancy, computing traffic growth rates on specific routes, identifying traffic pattern during various time slots, days, season etc. can use such vehicle counting and classifiers. This paper discusses a method for real time vehicle counting and classification using inductive loops for non-lane based mixed traffic flow conditions. The vehicle counting / classifier was field implemented at the JV Link Road, one of the major arteries in Mumbai, India and results verified for more than 100 samples of 10-15 minutes each

II. Inductive loop based Vehicle Counting and Classifier

Length based vehicle classification on freeways from single loop detector is followed by many agencies, including the USDOT. However, additional parameters are required to classify vehicles on urban streets where the mix of traffic is high. The parameter considered for our experiment is the width of the vehicle. As discussed earlier, the typical width of inductive loops used on roadways is 2mtr x 2mtr. Such loop can accommodate two motor cycles or scooters. These loops when kept apart by 1mtr can probably miss a two-wheeler or one bus can take two loops at a time, leading to errors in vehicle count.

The experiment conducted by the Intelligent Transportation and Networking Section (ITNS) of the Centre for Development of Advanced Computing (C-DAC), India used inductive loops of 1mtr x 1mtr placed in a row, each loop 0.5mtrs apart. This ensures no more than one vehicle occupy one inductive loop. However, one vehicle can occupy one or more loops depending on its width. The method assumes uniform average speed for all vehicles when moving in the platoon.

III. Test Site

Testing of the Inductive loop based real time vehicle counting and classifier was done at the Kowdiar, Thiruvananthapuram and Jogeswari-Vikroli Link Road (JVLR), Mumbai, India near the Indian Institute of Technology Bombay (IIT B) campus, Lake side gate. JVLR is an important arterial road connecting the Western Express Highway and Eastern Express Highway, Mumbai.

Eight inductive loops of size 1mtr x 1mtr were made on one direction of the road to cover the entire road width. The loops were kept at 0.5mtrs away from the shoulder and the median. Also a distance of 0.5mtrs was maintained between the loops. Four turns of 1.5mm Sq. copper cable were used to form each loop. The lead-in cables were brought to the nearby kiosk where the vehicle detector electronics are installed. A PTZ camera also was mounted at the test site to record video of the moving traffic over the inductive loops for the purpose of verifying the reported classified counts by the inductive loop based vehicle counting / classifier.



Figure 3: Installed Inductive loops at Kowdiar, Thiruvananthapuram

The inductive loop vehicle detectors used for access control and traffic signal operation typically have response time of 120-160mSec. Accuracy of these detectors was found poor in the counting and classifier application. Hence, vehicle detectors of 10mSec response time were used for the experiment at Kowdiar and JVL R.

IV. Algorithm

With the Traffic counting and classifier (TRAC2), the detector outputs are verified every 1mSec for presence or absence of a vehicle on the loop. A motorcycle or scooter will make detection on one loop only, as its maximum width is only 0.8mtrs. However, it can influence two loops if the vehicle is passing in between the loops. A car or three-wheeler can influence two adjacent loops, whereas a LCV, truck or bus can make change in three adjacent loops. Presence and absence of the vehicle in all the loops are monitored and the changes in the adjacent loops are grouped and filtered. Length of vehicles considered for classification is given in Table 1.

Table 1: Vehicle length considered for Classified Count

Sr. No.	Vehicle Type	Length (m)
1	Truck (L_T)	10.21
2	Bus (L_B)	11.54
3	LCV (L_L)	6.1
4	Car / Jeep (L_C)	4.4
5	Three Wheeler (L_3)	3.2
6	Two Wheeler (L_2)	2.4

The mean vehicle length (L_M) for all the classes of vehicle under consideration is computed as 6.31mtrs. This is a tuning parameter for the RTCC. The difference between the vehicle length and mean is tabulated as below for the purpose of finding out the variance.

Table 2: Difference between vehicle length and Mean

Sr. No.	Difference between Vehicle length and Mean	Length (m)
1	$\delta_T = \text{Truck } (L_T) - L_M(\text{Mean})$	3.90
2	$\delta_B = \text{Bus } (L_B) - L_M(\text{Mean})$	5.23
3	$\delta_L = \text{LCV } (L_L) - L_M(\text{Mean})$	-0.21
4	$\delta_C = \text{Car / Jeep } (L_C) - L_M(\text{Mean})$	-1.91
5	$\delta_3 = \text{Three Wheeler } (L_3) - L_M(\text{Mean})$	-3.11
6	$\delta_2 = \text{Two Wheeler } (L_2) - L_M(\text{Mean})$	-3.91

$$\text{Variance } \sigma^2 = \frac{(\delta_T^2 + \delta_B^2 + \delta_L^2 + \delta_C^2 + \delta_3^2 + \delta_2^2)}{\text{Number Of vehicle types } N}$$

$$\text{Standard Deviation } \sigma = \sqrt{\sigma^2}$$

Mean speed of the platoon is computed in every 20Sec from the average occupancy of the inductive loop detectors and the mean vehicle length. Individual detector occupancy is then filtered out to derive the classified count as per the length chart in Table 1. Correction factors are also derived for each class of vehicles which were found to be holding for all the data sets.

Step 1: Finding effective vehicle length

The Average length of the vehicles is the average of length of all vehicle types.

Average Vehicle Length (L_M) = Mean

Effective Vehicle length is derived as Average vehicle length x Classification constant / 100.

$$\text{Effective Vehicle Length } (L_E) = \frac{(L_M * C_c)}{100}$$

Step 2: Finding Classification Constant

Number of ways for crossing 0 loops = 1/4

Number of ways for crossing 1 loops = ${}^8C_1 = 8! / (1! \times (8-1)!) = 8$

Number of ways for crossing 2 loops = ${}^8C_2 = 8! / (2! \times (8-2)!) = 28$

Number of ways for crossing 3 loops = ${}^8C_3 = 8! / (3! \times (8-3)!) = 56$

There are 4 possibilities - You hit 0 Loops, 1 Loop and so on up to 3 Loops.

So each of these has 4 chances to occur.

So Probability = (4)* (1/no of ways a vehicle can hit simultaneous n Loops)

Probability for crossing 0 loops = $P_0 = 4 / {}^8C_0 = 4/4 = 1$

Probability for crossing 1 loops = $P_1 = 4 / {}^8C_1 = 4/8 = 1/2$

Probability for crossing 2 loops = $P_2 = 4 / {}^8C_2 = 4/28 = 1/7$

Probability for crossing 3 loops = $P_3 = 4 / {}^8C_3 = 4/56 = 1/14$

Gap Between 2 loops = $L_G = 0.5$ meter

Maximum Loops a vehicle can occupy = $\mu = 4$

$$\text{Classification Constant } (C_c) = e^{\frac{((N+L_G) \times \pi / \mu) + \cos(N))}{(P_0+P_1+P_2+P_3)}}$$

$$\text{Classification Constant } (C_c) = e^{3.81} \approx 45.5$$

Step 3: When a vehicle drives over a loop, it is counted, and the time that the vehicle spends over the loop is measured and recorded.

Step 4: From the average detection time, speed computed at every 20 second.

$$s = u \times t$$

Where

$$s = \sqrt{((34.45 - (0.56 \times \sigma^2)) + L_1)}$$

$$s = (L_S + L_1)$$

$$\text{Where } L_S = \frac{\sqrt{L_M}}{2} = \frac{\sqrt{(34.45 - (0.56 \times \sigma^2))}}{2}$$

$$\text{where } \sqrt{L_M} = \sqrt{(34.45 - (0.56 \times \sigma^2))} \text{ derived}$$

or $L_S = 70\%$ of Standard Deviation σ

L_1 = Length Of Loop

t_v = Individual recorded vehicle detection time

Speed for individual recorded detection time (m/s) =

((Square root of Average vehicle Length + L) x 1000) / Individual Detection Time

$$u_v[i] = \frac{s \times 1000}{t_v[i]}$$

Where

$u_v[i]$ = Speed of individual recorded detection time in meter per second

$t_v[i]$ = Detection time

Speed for individual recorded detection time (KMPH) =

Speed for individual recorded detection time (m/s) x 3.6

$$u_s[i] = u_v[i] \times 3.6$$

Average speed for 20 second (KMPH) = Speed for individual recorded detection time (KMPH) / Total count of speed

$$u_m = \frac{\sum_{i=0}^M u_s[i]}{M}$$

Where $i = 1$ to M - Number of detection time recorded

u_m = Average speed for 20 second duration in KMPH

$$u_a = \frac{\sum_{i=0}^M u_v[i]}{M}$$

Where $i = 1$ to M - Number of detection time recorded

u_a = Average speed for 20 second duration in m/s

$$\text{Weight (Speed Estimation factor)} = \alpha = \left(\frac{t}{1000} \right) \times \left(\frac{u_a}{N} \right)$$

Average Speed m/s for 20 second calculated based on Weight (Speed Estimation factor) is

$$u = u_a \times \alpha$$

Average Speed KMPH for 20 second

$$u_k = u \times 3.6$$

Step 5: Compute Vehicle length correlation factor from the computed speed by 20Sec data

Vehicle Length Correlation factor (ϕ_2) for 2 wheeler =

((Effective Vehicle Length (L_E) - Vehicle Length) / Effective Vehicle Length (L_E)) + 1

$$\phi_2 = \left(\frac{L_E - L_2}{L_E} \right) + 1$$

Vehicle Length Correlation factor (ϕ) for other vehicles =

(Vehicle Length - (Effective Vehicle Length (L_E)) / Vehicle Length) + 1

$$\phi_3 = \left(\frac{L_3 - L_E}{L_3} \right) + 1$$

$$\phi_C = \left(\frac{L_C - L_E}{L_C} \right) + 1$$

$$\phi_L = \left(\frac{L_L - L_E}{L_L} \right) + 1$$

$$\varphi_B = \left(\frac{(L_B - L_E)}{L_B} \right) + 1$$

$$\varphi_T = \left(\frac{(L_T - L_E)}{L_T} \right) + 1$$

Step 6: For individual class of vehicles length limits are computed based on the correction factor for each class and the average computed platoon speed.

Length limit Based on type of vehicle (ϑ) = (Vehicle Length *3.6/ Average speed for 20 second (KMPH)) x1000 x Vehicle Length Correction factor (φ)

$$\vartheta_2 = \left(\frac{(L_2 \times 3.6)}{u_k} \right) \times 1000 \times \varphi_2$$

$$\vartheta_3 = \left(\frac{(L_3 \times 3.6)}{u_k} \right) \times 1000 \times \varphi_3$$

$$\vartheta_C = \left(\frac{(L_C \times 3.6)}{u_k} \right) \times 1000 \times \varphi_C$$

$$\vartheta_L = \left(\frac{(L_L \times 3.6)}{u_k} \right) \times 1000 \times \varphi_L$$

$$\vartheta_B = \left(\frac{(L_B \times 3.6)}{u_k} \right) \times 1000 \times \varphi_B$$

$$\vartheta_T = \left(\frac{(L_T \times 3.6)}{u_k} \right) \times 1000 \times \varphi_T$$

Where u_k divided by 3.6 to convert speed in KMPH to Speed in meter per second.

Step 7: From the individual vehicle class length limit (ϑ), vehicle classified as follows

1. If detection time is less than two wheeler length limit, vehicle detect as a two wheeler.
2. If detection time is greater than two wheeler length limit and less than equal to the three wheeler length limit, vehicle detect as a three wheeler.
3. If detection time is greater than three wheeler length limit and less than equal to the Car/jeep length limit, vehicle detect as a car/jeep.
4. If detection time is greater than the car/jeep length limit and less than equal to the LCV length limit, vehicle detect as a LCV.
5. If detection time is greater than the LCV length limit and less than equal to the Truck length limit, vehicle detect as a truck.
6. If detection time is greater than the truck length limit, vehicle detect as a bus.

V. Result

More than 100 samples of 10-15 minutes were collected at the JVLIR at 20 Sec. sampling rate for speed computation. The video recording synchronized with the RTCC data collection was verified manually for accuracy of the data provided by the RTCC. The total system accuracy is found to be between 80 to 92 % on classified vehicle count of six classes of vehicles.

Table 3: Sample Counting / Classifier result

Vehicle type	Manual Counting	CDAC RTCC Vehicle Count	Classification Accuracy %	Counting %
2w	182	163	89.56%	94.58%
3W	195	171	87.69%	
Car/Jeep	275	409	67.24%	
LCV	178	158	88.76%	
Bus & Truck	77	58	75.32%	

Result Comparison with Manual Count

Results obtained for different samples by the RTCC and verification by manual count are plotted in Figure 4 – 9.

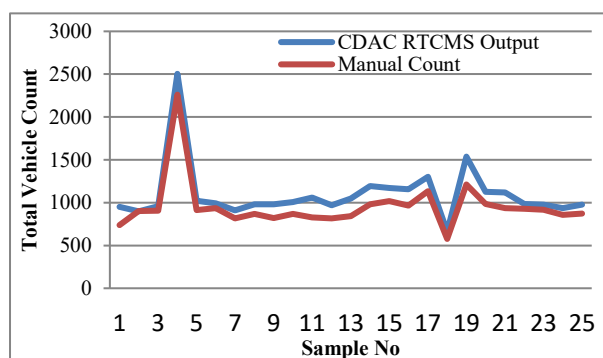


Figure 4: Total Count for different samples

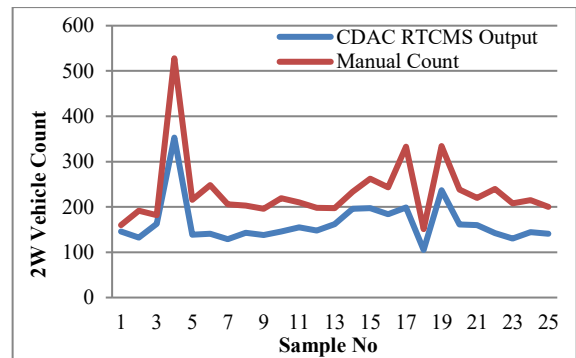


Figure 5: Two-wheeler Count

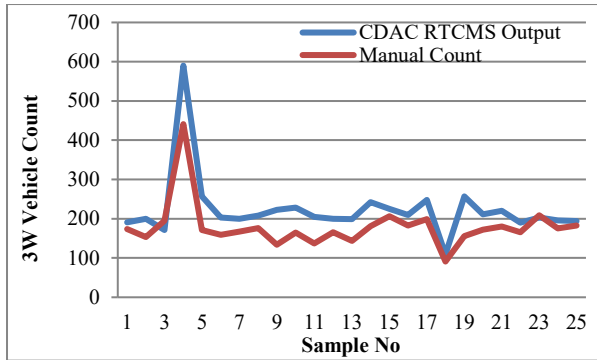


Figure 6: Three-wheeler Count

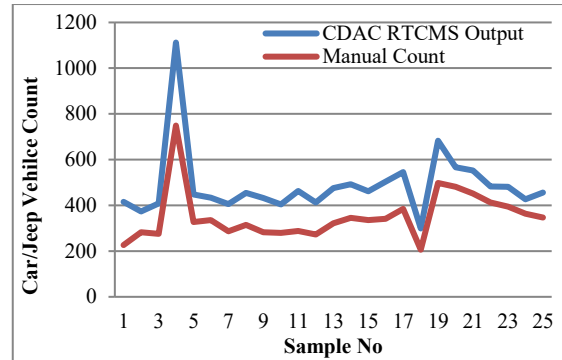


Figure 7: Car and Jeep Count

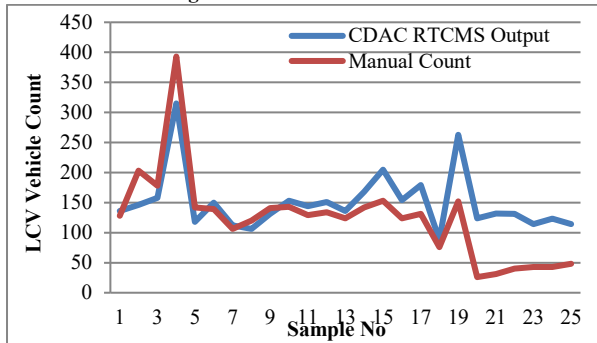


Figure 8: LCV Count

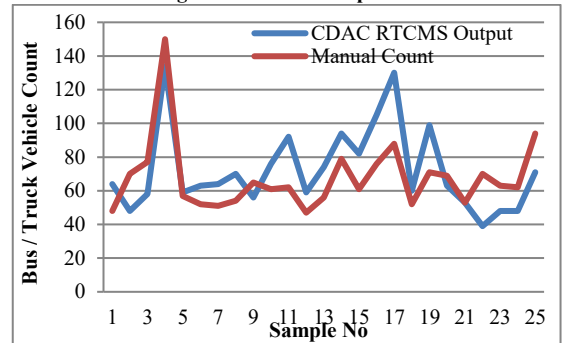


Figure 9: Bus and Truck Count

VI. Conclusion

Development of a method for classified count of vehicles on urban roads where the traffic is highly heterogeneous and the lane discipline is poor was challenging. To the best of our knowledge the available systems address only disciplined traffic and limited heterogeneity. The test results obtained by CDAC at the JV Link Road, Mumbai is highly promising. This needs to be tested in different cities having different vehicle composition to prove universal. However, we believe this invention is a frog-leap in the frontier of traffic data collection.

Acknowledgments

The authors acknowledge the support provided by the Team ITNS, CDAC; Department of Electronics and Information Technology (DeitY), Govt. of India; and the Department of Civil Engineering, Indian Institute of Technology Bombay who helped us in developing the real time vehicle counting and classifier for urban roads.

References

- [1] U.S. Department of transportation, 3rd Edition, Federal highway Administration (Oct 2006). *Traffic Detector Handbook*.
- [2] S. Sheik Mohammed Ali, Niranjan Joshi, Boby George and Lelitha Vanajakshi (2012). *Application of Random Forest Algorithm to Classify Vehicles Detected by a Multiple Inductive Loop System*, IEEE International Conference on Intelligent Transportation Systems, Anchorage, Alaska, USA.
- [3] Sung-Wook Kim, Kwangsoo Kim, Joo-hyung Lee and Dong-il (Dan) Cho (2012). *Application of Fuzzy Logic to vehicle Classification Algorithm in Loop. Piezo-Sensor Fusion Systems*, IEEE International Conference on Intelligent Transportation Systems, Anchorage, Alaska, USA, September 16-19, 2012.
- [4] Soner Meta and Muhammed G. Cinsdikici, Member, IEEE (2010). *Vehicle-Classification Algorithm Based on Component Analysis for Single-Loop Inductive Detector*, in IEEE transaction on vehicular technology, Vol. 59, No. 6, July 2010.
- [5] Glenn Arr, Carlos Sun and Ravi P. Ramachandran (2004). *Fusion of Wavelet Transform and Color information feature for Automatic vehicle reidentification in Intelligent transportation Systems*, in IEEE ICASSP 2004.
- [6] Ryszard Sroka (2004). *Data Fusion Methods Based on Fuzzy Measures in Vehicle Classification Process*, in IMTC 2004 - Instrumentation and Measurement Technology Conference Como, Italy, 18-20 May 2004.
- [7] S. S. M. Ali, B. George, L. Vanajakshi, V. Jayashankar and V. J. Kumar (2011). *A Multiple Loop Vehicle Detection System for Heterogeneous and Lane-less Traffic*, in IEEE International Instrumentation and Measurement Technology Conference (I2MTC-11), Hangzhou, China, May 10-12, 2011, pp. 1413- 1417.
- [8] S. S. M. Ali, B. George, L. Vanajakshi (2011). *A Simple Multiple Loop Sensor Configuration for Vehicle Detection in an Undisciplined Traffic*, in Proc. IEEE 5th Int. Conf. ICST, Palmerston North, New Zealand, Nov28th – Dec.1st 2011, pp. 644 – 649.
- [9] S. S. M. Ali, B. George, L. Vanajakshi (2001). *A Magnetically Coupled Inductive Loop Sensing System for Less –lane Disciplined Traffic*, in IEEE International Instrumentation and Measurement Technology Conference (I2MTC-12), Graz,Austria , May 13-16, 2012, pp.827–832.

Layout techniques & matching strategies for CMOS analog integrated circuits

Chetali Yadav¹, Sunita Prasad², M. Bharath Reddy³ and Manoj Kumar⁴

^{1,2}VLSI Design Department Centre for Development of Advanced Computing Noida, Uttar Pradesh, India

^{3,4}VLSI Design Division Semiconductor Laboratory (ISRO) Mohali, Punjab India

Abstract: In this paper, the state of art layout techniques are presented. The gate folding techniques have been used which decreases the parasitic gate resistance and gate capacitance to improve the performance of CMOS circuits. Most of the CMOS circuits are differential for suppressing common mode noise. The differential circuits shall be perfectly matched so that common mode noise cancellation is perfect. The various techniques of matching the differential pair like interdigital & common centroid is presented. For relieving the stress of STI, one or more dummy transistors at each end of the transistor are inserted. The guard rings are used around sensitive transistors for reducing the effect of substrate noise. The layout of an operational amplifier has been drawn as an illustration.

Keywords: layout; CMOS; Integrated circuit; Common centroid; Interdigital; folded gate, dummy transistors

I. Introduction

The simulation of analog / digital circuits provides the schematics and size of transistors. The first step of physical design is to draw layout from schematics and size of transistors obtained from simulation. The analog circuits are differential and matching of devices are required to obtain same results after fabrication as obtained from simulation using professional softwares. Basic Analog modules for CMOS circuits are Differential transistor pair, Current Mirrors and operational amplifiers. The device matching is required in these circuits to reduce offset voltage. The layout of any analog circuit affects the performance of the circuit due to stress mismatch, thermal and process gradients. The various layout techniques are described in this paper for producing best performing analog circuits.

II. Techniques for reducing parasitic resistance and capacitance

The every transistor have parasitic resistance and capacitance. The polysilicon gate of the transistor has finite resistance. Also there is parasitic capacitance formed by polysilicon and bottom ground of silicon chip. This is distributed resistance and capacitance along the length of the gate of transistor.

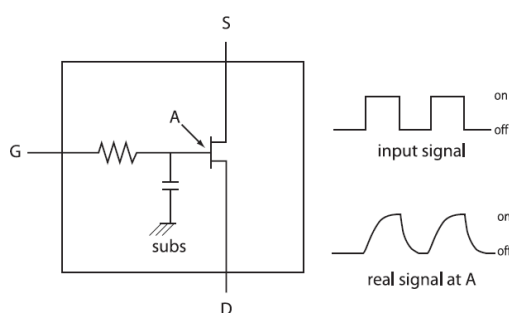


Figure 1 Effect of parasitic resistance and capacitance

There is finite charging and discharging time due to parasitic resistance and capacitance at gate. The charging and discharging time depends on RC time constant of parasitic resistance and capacitance. The parasitic gate resistance increases the charging & discharging time of the parasitic gate capacitance. If we apply high frequency square waveform at the gate of transistor then it is observed that the output is not square. If we want to get square wave at the output then RC time constant of gate parasitic resistance and capacitance shall be reduced.

For reducing the parasitic Gate resistance & parasitic Gate capacitance the single transistor is replaced by multiple transistors of smaller gate width.

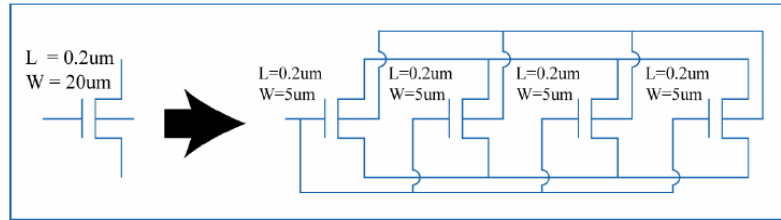


Figure 2 Large size single finger transistor is splitted in smaller multi finger transistor

In drawing layout of four transistors, we use source-drain sharing. The same region is a source for both transistors and a drain for both transistors at the same time. This technique of multiple finger gates shall be used for further reducing parasitic resistance and capacitance.

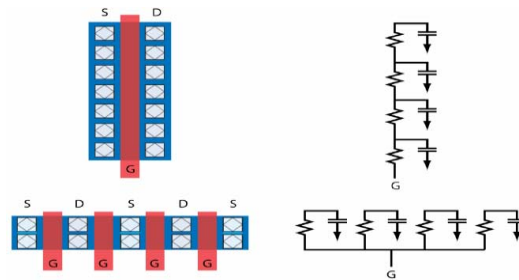


Figure 3 Multi Finger transistors for reducing effect of parasitic resistance and capacitance

All drains are connected together. Similarly all gates & sources are connected together.

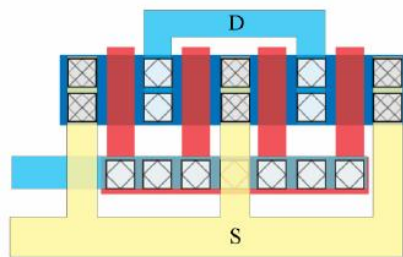


Figure 4 Complete layout of Multi Finger transistor

III. Matching of devices

Matching layout is used to enhance the relative precision of device pair (e.g. a differential transistor pair, a current mirror) (around $\pm 1\%$). Consider a differential pair having two matched transistors with one node in common

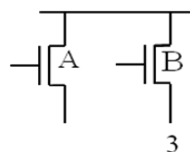


Figure 5 Differential pair transistors

A. Interdigitated Devices

We split them in an equal part of fingers, interdigitate the 8 elements: AABBAABB or ABBAABBA

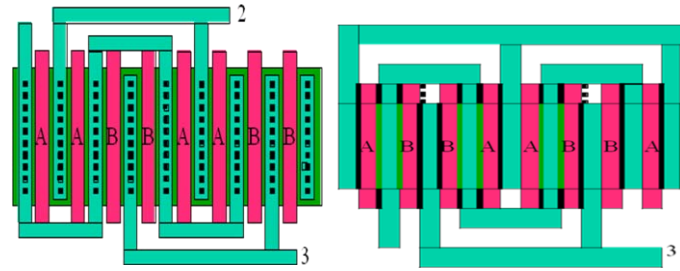


Figure 6 Two ways of interdigitated transistors

B. Common Centroid Devices

The fluctuation of the device characteristics may be canceled using the common centroid. Gradients in features are compensated for (at first approximation). The centroid of the matched devices should be coincident. The common centroid devices will be symmetrical around both the x and y-axis. This topology has immunity from process gradient effects from any direction.

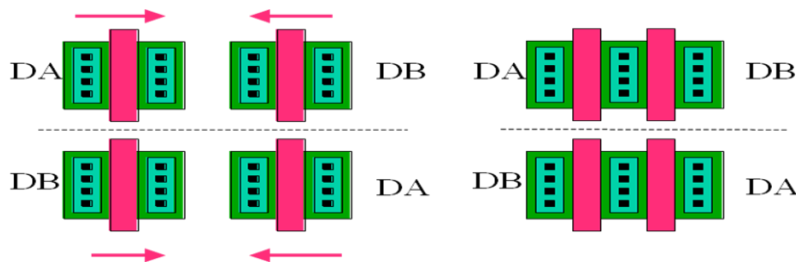


Figure 7 Common centroid transistors

C. Dummy Devices on Ends

Dummies are shorted transistors. Ending elements have different boundary conditions than the inner element. By adding dummy devices, the boundary conditions of all the devices becomes same.

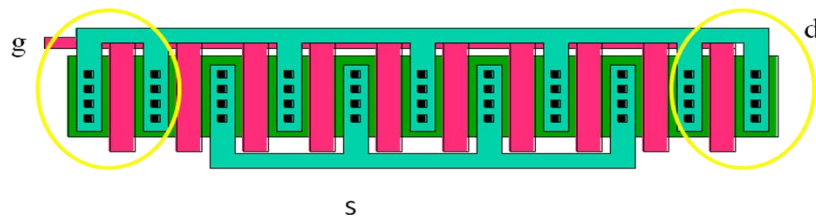


Figure 8 Dummy devices in the ends

D. Matched Metal connections

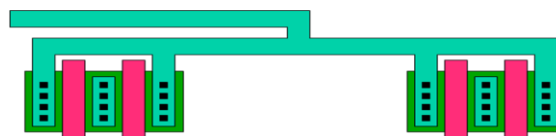


Figure 9 Matched metal interconnects

The interconnect lines widths and path lengths are also taken identical so that parasitic resistance and parasitic capacitance is also equal.

IV. Guard Ring around transistors for reducing noise coupled through substrate

We Surround nMOS transistor by p+ connection to VSS, surrounded by n-well with n+ connection to VDD. Also surround pMOS transistor by p+ ring connected to VDD, surrounded by n-well with n+ connection to VSS.

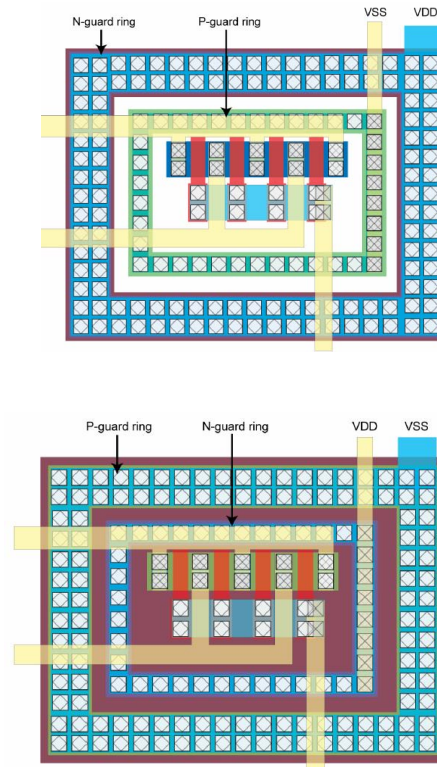


Figure 10 Guard rings around transistors

V. Layout of Operational Amplifier

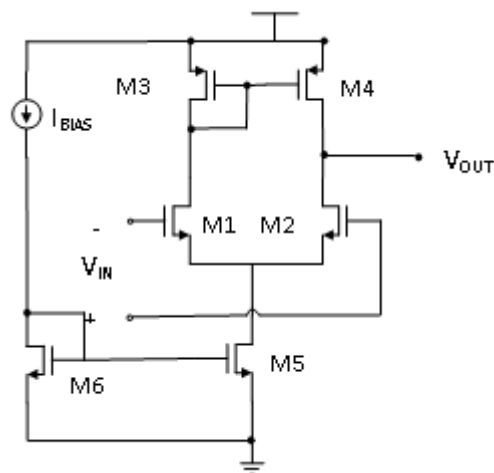


Figure 11 single stage operational amplifiers

As a example the layout of single stage operational amplifier is given below. The transistor M1 is matched to transistor M2, transistor M3 is matched to transistor M4 & transistor M5 is matched to transistor M6.

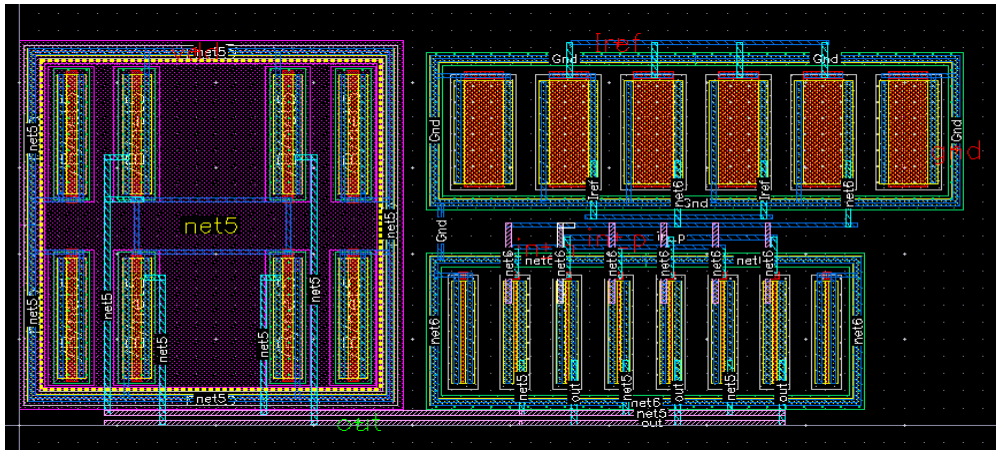


Figure 12 Layout of Single stage operational amplifier

The Schematic diagram of two stage operational amplifier and its layout is given in fig 13 & 14 respectively

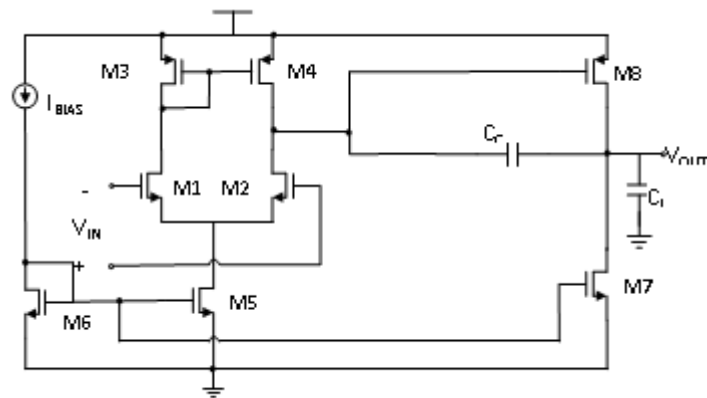


Figure 13 Two stage operational amplifier

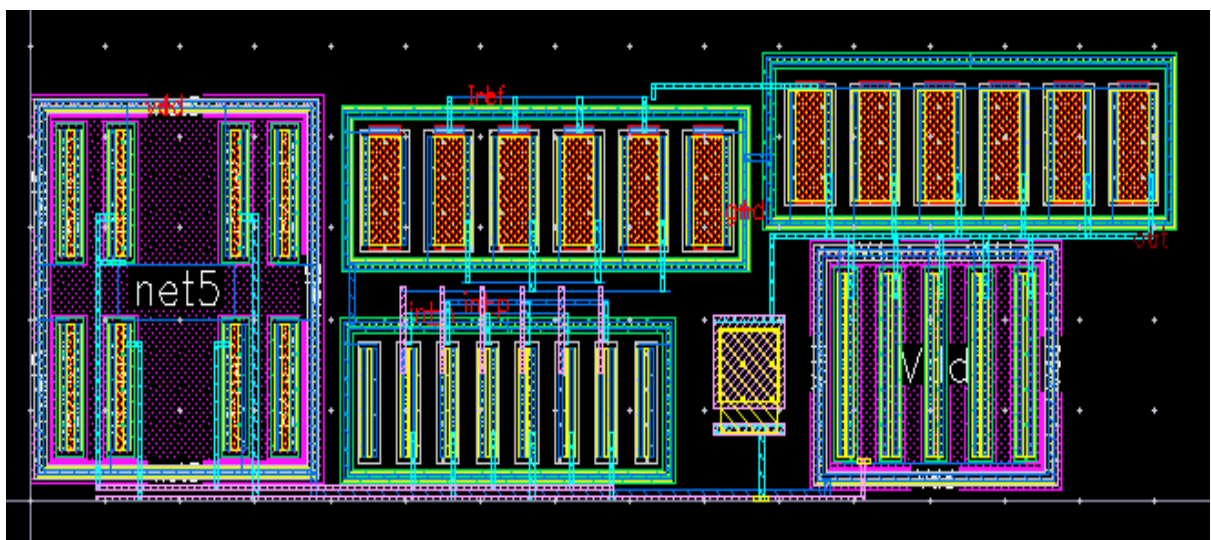


Figure 14 Layout of two stage operational amplifier

The layout given in Fig. 14 makes use of multi-finger transistor for reducing parasitic resistance and capacitance. Interdigital matching has been used for M1-M2 differential pair and M5-M6 current mirror

transistors. Common centroid matching has been implemented in M3-M4 PMOS current source load pair. Dummy devices have been added at the outer ends of each transistor for relieving the stress of STI. Equal matched metal interconnects path lengths have been used for each transistor of matched pair so that parasitic resistance and capacitance of the path lengths are also equal. Guard rings have been added around matched transistors as well as sensitive transistors for reducing noise coupled through the substrate. The total size of the layout is 76um x 25um. The design rule check and layout versus schematic check have been successfully carried out using Calibre software of Mentor Graphics.

VI. Conclusion

The best layout practices like reduction of parasitics by using multi-finger transistors, reduction of substrate coupling by using guard rings etc. are described. The different matching strategies with emphasis on common-centroid and interdigital for differential transistor pair, matching of interconnect metal, and stress effect mitigation is presented.

VII. References

- [1] Behzad Razavi, Design of analog CMOS integrated circuits, McGraw Hill Book Co, 2000.
- [2] T. Massier, H. Graeb and U.Schlichtmann, "The Sizing Rules Method for CMOS and Bipolar Analog Integrated Circuit Synthesis", Transactions on Computer-Aided Design of Integrated Circuits and Systems, pp. 2209 – 2222, Dec. 2008.
- [3] Saravanan Balakrishnan et al, "Analog Layout Design Optimization and Verification for SoC" IEEE RSM2011 Proc., 2011, Kota Kinabalu, Malaysia, pp. 128-131.

VIII. Acknowledgments

Authors are grateful to the director of Semiconductor Laboratory (ISRO) Mohali, India for providing permission for training & facilities for doing this work. The faculty members of CDAC Noida deserves special mention for their technical support and timely reviews.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

FPGA Implementation of Sensor Fusion Technique for Obstacle Detection

Danish¹, Dr. Sunita Prasad²

School of Electronics

Centre for Development of Advanced Computing
B-30, Institutional Area, Sector-62, Noida, India

Abstract: Ultrasonic sensors and Infrared (SHARP) sensors are widely used in mobile robotics applications for distance measurements. This paper presents an obstacle detection system which has been built using these two types of range sensors. But there are problems when working with these sensors for practical applications like uncertainties in results obtained from these sensors, failure of sensors etc. These problems can be solved by using sensor fusion techniques. In today's smart world, the smart device requires reliable sensory data which can be obtained by fusing the results obtained from different types of sensors. In this paper, a less complex sensor fusion technique has been applied for obstacle detection system. The fusion techniques that have been implemented so far are microcontroller based which do not provide very good speed. So, it is not good choice for real-time applications. So, a less complex fusion technique has been implemented on FPGA which is ideal choice for real-time applications. This proposed fusion technique will not only detect the obstacle but also find its distance from that obstacle using two range sensors. The hardware implementation has been done on Spartan3E FPGA using Xilinx ISE Design Suite 14.5. The proposed technique is less complex in terms of computation and as FPGA has been used as hardware so it provides very good speed.

Keywords: Sensor Fusion, Ultrasonic Sensor, Infrared Sensor, Basys2 FPGA Board, PCF8591 ADC.

I. INTRODUCTION

In order to gather the real-time information from the environment to perform some specific tasks sensors are used. The sensor systems are not only used for collecting the data from the environment but also used for translating this data to some meaningful data which can be sent to controller of the system to perform some task based on this information. In today's world, everything is getting smarter and smarter day by day so the electronics devices are. To make any device smarter, the first thing should be that it must give reliable and accurate results. For making any system smarter, sensor fusion plays an important role in which different types of sensory data are combined in such a way the resultant or fused result is more accurate and reliable than result obtained from individual sensor [1]. The sensor fusion can also be applied in obstacle detection and navigation application of mobile robotics. But the problem in working with sensors for real-time applications is the uncertainties present with these sensors. For example, sensor failure may lead to disastrous results and improper working of sensors may give false result which ultimately causes system failure and many more such problems may arise. But the problems can be overcome by sensor fusion. In simple words, sensor fusion means integration of information obtained from different types of sensors into a unified interpretation [2].

Today, mobile robotics plays a vital role in many application areas and therefore it is one of the progressive fields of technology. But main problem with the mobile robot system is obstacle detection and their localization. It is not possible for the robot to take the decision that where it is, where the path is, where it needs to go, which path to take or has it reached the destination yet. It can be made possible using sensor fusion which is a technique that combine the information obtained from the sensors and improve the data rather than the case single sensor used [1]. The obstacle position should be obtained as accurate as possible in order to support robot self-localization procedures. A robot interacts with its surroundings in a flexible manner using external sensors. For example, a robot can operate an obstacle flexibly based on sensor data without intervention by a human operator [2]. This paper focuses on the two different types of range sensors to avoid obstacles and find out the distance from that obstacle with the help of an FPGA. In this research work, we have worked on real-time application of mobile robotics which is obstacle detection application, using simple and less complex sensor fusion technique in order to obtain accurate result and obstacle would be detected based on this accurate result. There are some sensors available which can be used to measure the distance to an obstacle and as a result these sensors may be used in robotics application where a robot can take decisions based on presence or absence of that object. Here, for applying sensor fusion technique we have used two different range finding sensors. Both of these sensors have some advantages and disadvantages and sensor fusion technique has been applied in such a way that the limitation of one sensor can be compensated by advantages of other sensor [3]. In this way of applying sensor fusion a more accurate result may be obtained which is not possible to obtain if these sensors

are employed individually. The two range finding sensors that have been used in this research work are Ultrasonic range sensor and Infra-red Sharp range sensor. This paper would not only focus how to detect the obstacle if that obstacle come within a particular range but also able to find its accurate distance from that obstacle. This range finding ability of this project could be very useful in mobile robotics in taking some decision based on distance.

Nowadays, the use of FPGA has become very popular for solving complex computational problem and therefore design engineers prefer FPGAs for solving complex tasks. Due to numerous advantages of FPGAs, traditional microcontroller based hardware systems can be replaced by FPGAs for field computation. The main feature of FPGA is that it is reconfigurable which makes the system flexible [1]. FPGAs also provide other benefits over microcontrollers such as performance, prototyping capabilities, and reliability. The design based on FPGAs provides more accuracy and it is much faster than the traditional microcontroller based system. Because of its high speed advantage, FPGAs are the ideal choices for real-time applications. This paper explains the implementation of a simple fusion technique for infrared and ultrasonic sensors using FPGA.

II. SENSOR CHARACTERISTICS

Range sensors can be used to find out the distance to an obstacle and in this project two different range sensors have been used. These are ultrasonic sensor and infrared sensor. Both of these are used for distance measurement and each sensor has its some advantages and disadvantages which will be discussed in this section. Due to different characteristics, these two sensors give different results in same situations. Infrared sensor provides a sharp focus therefore it gives more accurate result than other sensor. The stability of this sensor is also very good. Another main advantage this sensor provides is its faster response time which is much desirable features for improving the real-time response of a mobile robot [3]. Apart from these benefits this sensor also exhibits some limitations like it cannot detect transparent or glass-based objects because the IR beams emit by this sensor refract from the glass-based object and does not reflect back to sensor or only a small proportions of IR beams reflected back which results into no result or false distance calculation [13]. Another problem with IR sensor is that it produces inaccurate results for the objects having dark surfaces like black because it depends on reflectance of the object's surface and reflectivity of IR beam with dark surfaces is less. The major problem with IR sensor is that it shows non-linear characteristics with distance. According to the datasheet of IR sensor provided by the SHARP that it shows highly non linearity for the distance less than up to 15 cm and the curve becomes linear as the distance increases therefore the minimum distance measured by IR sensor has been kept above 10 cm [13]. Below 10 cm it gives inaccurate results. The IR sensor is based on triangulation method for finding the distance to the obstacle [3]. Figure 1 depicts these two range sensors. On the other hand ultrasonic sensor is based on sound wave propagation which is not affected in detecting the transparent obstacles because



Fig 1: (a) Ultrasonic sensor (HC-SR04) (b) Infrared sensor (SHARP GP2Y0A21)

sound waves easily reflect back from the obstacle's surface to the sensor module [1]. Another benefit of ultrasonic sensor is that it gives accurate result even conditions of poor lighting and any kind of obstacle's surface. As this sensor has limitation of wider beam width therefore it does not gives sharp focus. One more problem with US sensor is that it also depends on the temperature and shape of the obstacle's surface [4] because the velocity of sound wave travel in air is affected by environmental parameters like temperature and humidity. The shape of surface should be wide enough so that sound waves can strike the surface completely and then reflected back to sensor module. In the project, ultrasonic and infrared sensors have been used in complementary fashion, where the benefits of one sensor can compensate the limitation of the other sensor. The range that these sensors can measure is 10 cm to 80 cm for IR sensor while for US sensor this range is around 400 cm [14]. These sensors can be used for various applications apart from distance calculation like navigation systems in vehicles, home etc. as obstacle detection or avoidance, counting devices, edge detections etc. When these two contactless sensors are used for distance measurement then time of flight method is preferred for the distance measurement in which the time taken by a pulse of energy travelling from its transmitter to an obstacle and then back to the receiver of the sensor unit is measured [3]. The distance can be calculated by multiplying the velocity of the received energy pulse by the time of flight.

III. PROPOSED APPROACH

This project work is basically the hardware implementation of low level sensor fusion technique that can be applied to both Ultrasonic and Infrared range sensors for measuring the distance to an obstacle using FPGA [1]. The hardware which has been taken for this project work is Basys2 FPGA board which consist Spartan3E FPGA. In this proposed approach, a simple sensor fusion technique has been proposed and implemented on FPGA itself which is less costly both in terms of complexity and computationally, that will allow an autonomous robot to detect an obstacle accurately, and also find the accurate distance to that obstacle. The block diagram for the proposed approach of the project is shown in Figure 2.

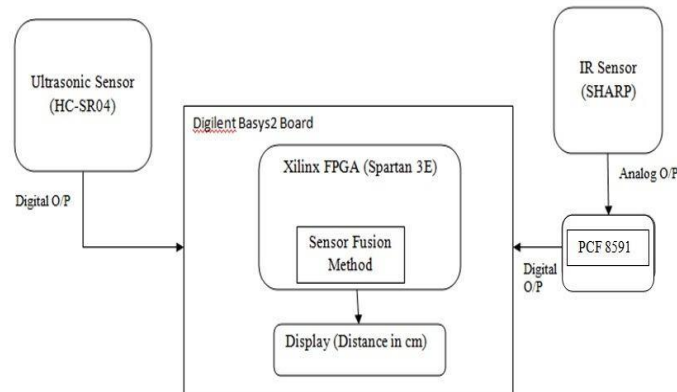


Fig 2: Block diagram of the proposed system

In this proposed work, we have used two different sensors, an IR sensor (SHARP GP2Y0A2I) and an US sensor (HC-SR04) for measuring the distance to an obstacle. These sensors have been utilized in such a way that they are used in a complementary manner to give reliable distance measurement. The results coming from these sensors have been fused in such a way that the advantages of one compensate for the disadvantages of the other. The proposed system has been classified into three modules. The first module is ultrasonic range sensor module in which US sensor has been interfaced with the given FPGA board and the measured distance to an obstacle has been displayed on seven segment display mounted on the board. Similarly, the second module is infrared range sensor module in which IR sensor has been interfaced with FPGA and displayed the result on displaying unit of that board. The last module is sensor fusion module which combines these two modules to obtain the fused data.

A. Ultrasonic Range Sensor Module

The US sensor gives the output in pulse form which is digital in nature therefore the data coming from this sensor can be directly sent to the FPGA. Due to this, we do not need any additional hardware unit for interfacing the US sensor with the FPGA. The US sensor which has been used in this work is HC-SR04 module which consists of a transmitter, a receiver and a control unit [14]. The distance can be calculated using sound wave propagation method. The block diagram of ultrasonic range sensor module is shown in Figure 3.

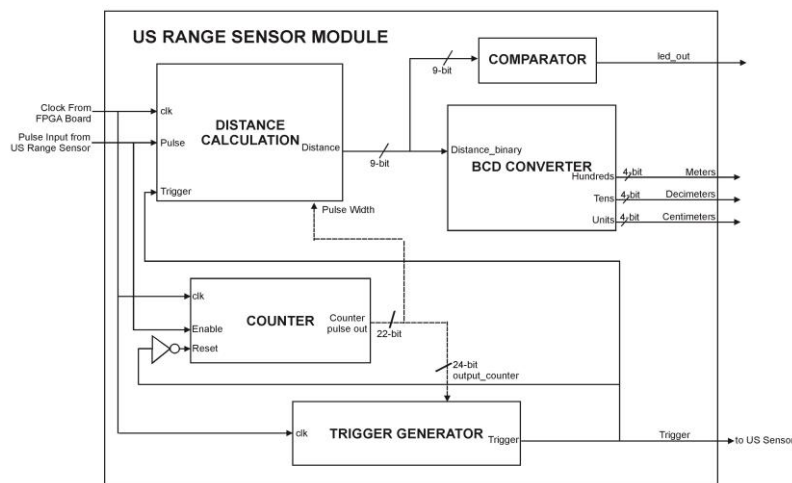


Fig 3: Block diagram of ultrasonic range sensor module

The US sensor module can be activated by sending a trigger pulse (which is a high Transistor-to-transistor logic (TTL) pulse of minimum 10 us duration) to trigger pin of sensor module. This trigger pulse is generated by trigger generator module. After sending the trigger pulse the transmitter unit of this sensor radiates the sound

waves in the form of 8 cycle bursts at a frequency of 40 kHz. At this time, the sensor generates an echo pulse which is a high TTL pulse and this pulse remains high till the 8 cycle bursts of sound waves reflect back to sensor's receiver after striking to the obstacle [4]. The width of this echo pulse is proportional to the distance to that obstacle. Now this echo pulse is sent to the FPGA from the echo pin of sensor module. The distance calculation module will convert this pulse width into the distance in cm. The distance calculation module is designed to translate the response of US sensor into a numeric value. This module waits for the echo pulse (coming from US sensor) to rise from low to high. Once this pulse goes high, this module counts the number of clock cycles until the echo pulse goes low again. Now this number of clock cycles is multiplied by the clock frequency of FPGA in order to get the high level time of echo pulse [4]. Now, the distance to the obstacle can be calculated by using the following equation [1].

$$\text{Distance} = \text{High level time of echo pulse} * \text{velocity of sound waves (340 m/s)} / 2$$

This distance would be in binary format therefore using BCD converter this distance can be converted into BCD format in order to display the result on seven segment display unit. For binary to BCD conversion, the Double Dabble algorithm has been used. According to the US sensor datasheet, to calculate the distance in centimeter, the echo pulse width is measured in microseconds and then divided by 58 because every 5800 clock cycles at a frequency of 100 MHz corresponds to 1 cm measured [14].

B. Infrared Range Sensor Module

The output of IR sensor is analog in nature because this sensor gives the output in voltage form which is inversely proportional to the distance to an obstacle [1]. So the data coming from this sensor cannot direct send to the FPGA because FPGA cannot work with analog signals. Therefore we need to convert this analog voltage in digital form before sending to the FPGA. For that purpose we have used PCF 8591 module which is an I2C compatible device. This device can be used either as 8-bit ADC or 8-bit DAC. The PCF 8591 module is an 8-bit CMOS data acquisition device which consists of 4 analog inputs for ADC, one analog output for DAC and a serial I2C interface for transferring the converted data to the controller [11]. In this implementation, the analog output (voltage) of IR sensor has been applied to the first analog input pin of ADC (PCF 8591) and the remaining inputs pins have been grounded. The ADC has been connected to given Basys2 FPGA board using I2C bus interface because PCF 8591 is an I2C compatible device. In this case, FPGA board is the master and the PCF 8591 is a slave device. I2C stands for Inter-Integrated Circuit. It is very popular bus protocol which uses only two wires (SCL line for giving clock signal and SDA line for carrying data bits) for bidirectional serial data communication between master and slave devices [12]. The 8-bit ADC output (which is inversely proportional to distance) is given to the FPGA via bus I2C interface. In order to read the ADC data coming from PCF8591 device

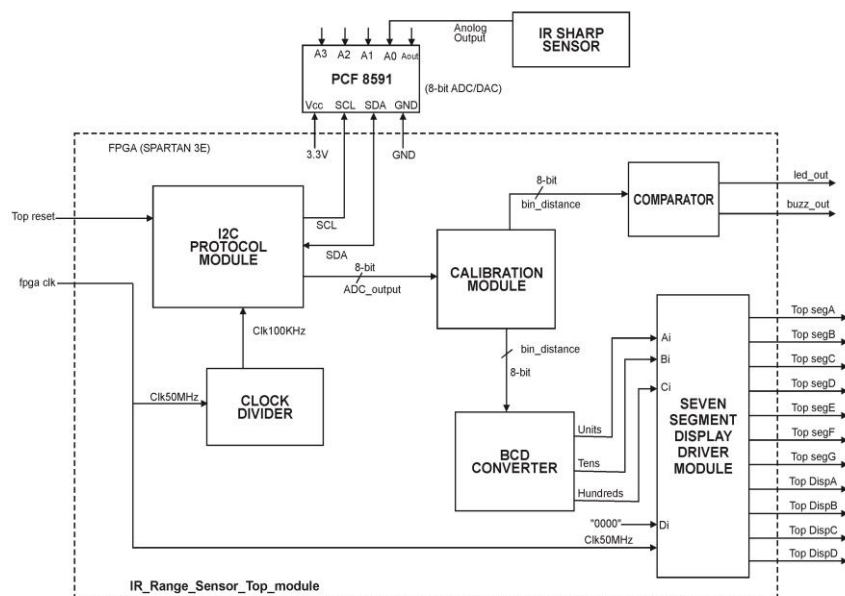


Fig 4: Block diagram for the IR range sensor module

through I2C bus, we have implemented the I2C bus protocol for master device (FPGA) using VHDL. The I2C protocol module can be seen in block diagram of infrared range sensor module (see Figure 4). This module reads the 8-bit ADC data from the PCF 8591 device using I2C protocol. This 8-bit ADC output of PCF 8591 module is not actual distance but an 8-bit binary number which is corresponding to analog voltage of IR range sensor. To get the actual distance in cm, this ADC output needs to be calibrated. A simple technique has been used for calibration. As the ADC output is inversely proportional to the distance to an obstacle, the distance can be

obtained by using the formula: $\text{Distance} = K / V_o$, where K is the calibration constant and V_o is the output of ADC which is 8-bit digital equivalent of analog voltage of IR sensor. To obtain the value of K , a calibration experiment has been performed in which the 8-bit output values of ADC are measured by placing a white board in front of IR sensor and moving that board farther to sensor. The value of ADC output has been recorded in decimal for different positions of white board and then K value has been calculated for each measurement and to get the more accurate value of K , the mean of K values has been taken.

C. Sensor Fusion Range Sensor Module

In this module, actual fusion of sensory data has been performed on the data coming from two different range sensor modules. Sensor fusion technique which is employed in this proposed work is less complex in terms of computation. Here, the fusion technique is applied in such a way the limitation of one sensor is compensated by other sensor. In this way it gives reliable and accurate distance which is used to detect the obstacle for various applications like mobile robotics for object localization etc [2]. As per the IR sensor characteristics which gives unexpected outputs for distance less than 6 cm and also range of this sensor limited to 40 cm for 3.3 V operating voltage. Here fusion technique has been applied in such a way that if distance to an obstacle is less than 6 cm and greater than 40 cm then it passes the result obtained from ultrasonic sensor as fused data. But if the distance ranges from 10 cm to 40 cm then the mean value distances obtained from both sensors is considered as fused data. This fused result more reliable and accurate distance to an obstacle. In some cases like detecting objects having sharp edges or shape in not uniform then the result obtained from US sensor is not accurate so this technique gives more accurate result as well.

III. RESULTS

The objective of this work is to detect the obstacle if that obstacle comes within the specified range and also find its distance in cm. Firstly, distance to an obstacle has been measured using individual sensors and then this measured distance has been plotted with the actual distance of that obstacle. Finally, the distance measured from fusion of both sensors has been plotted with the actual distance. For observing the results an experimental set up has been designed where both range sensors are mounted on a piece of hardware along with given FPGA. In this set-up, a measuring tap has been used for comparing the actual distance with the distance obtained from the design. First of all we measured the distance obtained from ultrasonic sensor alone and compared that result with actual distance and a graph has been used for showing the comparison result between measured distance and actual distance. The graph showing the observation result for ultrasonic sensor is shown in Figure 5 (a). The graph shows relation between actual reading and distance measured by ultrasonic sensor. It shows a linear relationship. In second observation, only IR sensor has been used for getting results and the observation result has been shown graph as shown in Figure 5 (b). The result shows that IR sensor gives variation below 10 cm distance and afterward it gives the linearity.

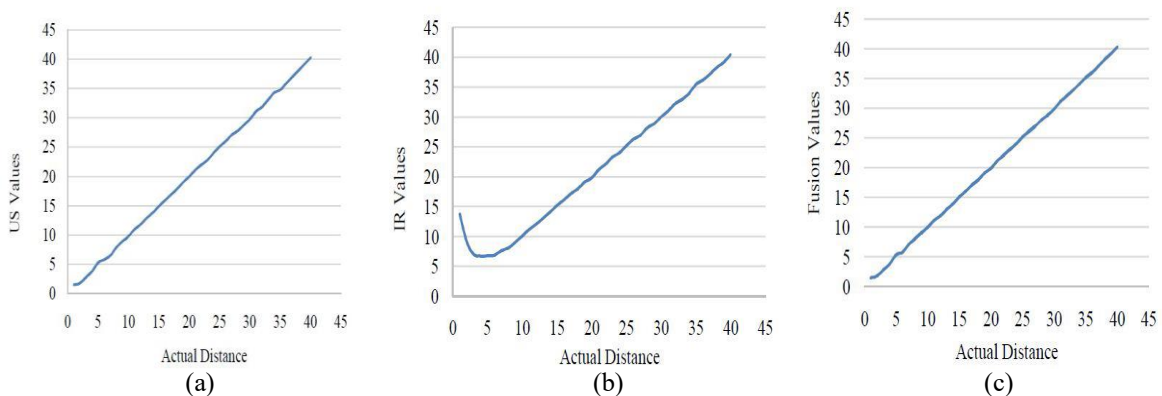
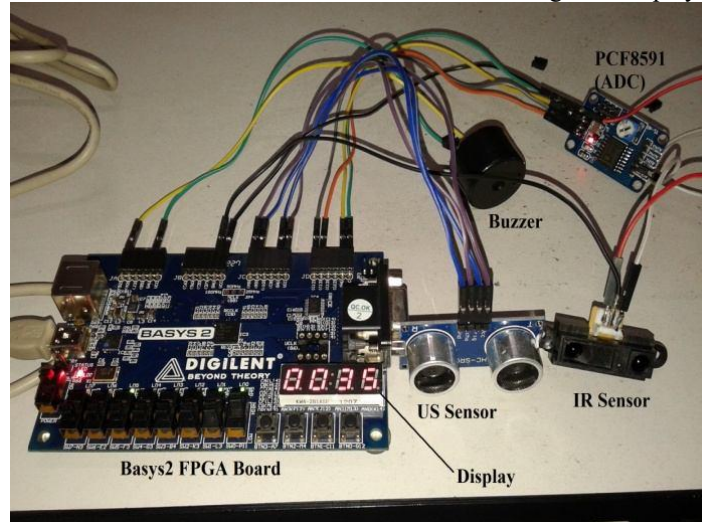


Fig 5: Observation result for (a) Ultrasonic sensor (b) Infrared sensor (c) Fusion data

Finally, both the modules connected to given FPGA, and observe fusion result. Figure 5 (c) shows the relation of actual distance and fusion data and it shows the linear relationship. This characteristic shows that instead of the single sensor if two or more sensors are used then it will give better results than single sensor. For distance (< 6 cm) estimated in such a way that fusion data take the reading of ultrasonic sensor only because there is large variations in the IR sensor reading. After the 6 cm distance, the mean of both the sensors will be considered as estimated distance called fusion data. The plotted graphs have similarity to the shapes that given in data sheets provided by the manufactures. The snapshot of hardware implementation of sensor fusion system arrangement is shown in Figure 6. A piezoelectric buzzer is used in this system for detecting the obstacle. When that obstacle

comes within the specified range of these sensors then this buzzer gives a beep sound which indicates that obstacle has been detected. The measured distance can be seen on seven-segment display of FPGA.



III. CONCLUSION

This paper presented a simple sensor fusion technique for obstacle detection has been implemented successfully on the Spartan 3E FPGA. In this proposed work, two different range finding sensors that are ultrasonic and infrared sensors have been used for distance measurement in the development of an obstacle detection system. It can be clearly concluded from the observation results that ultrasonic sensor gives a linear output characteristic whereas infrared sensor shows a nonlinear output characteristic for distance measurements. The approach for sensor fusion which has been used in this proposed work is less complex in terms of computation. This advantage of low computational complexity in hardware leads to less hardware resource requirements without compromising the speed of operations. In this work, for interfacing the IR sensor with FPGA an I2C compatible ADC (PCF8591) has been used which helps in reducing the number of interconnection pins because here all the data transmission occur using only two wires.

IV. REFERENCES

- [1] Salina. B., Dr P. Malathi, "FPGA Implementation of Data Fusion Algorithm for Object Localization", published in *IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 2014, ISBN 978-1-4799-4040-0.
- [2] M E. Conde, Sergio Cruz, D M. Munoz, Carlos H.Llanos, Eugenio L F Fortaleza, "An efficient data fusion architecture for infrared and ultrasonic sensors, using FPGA", published in *IEEE Transaction Circuits and Systems*, 2013, ISBN 978-1-4673-4900-0.
- [3] Baharuddin Mustapha, Aladin Zayegh, Rezaul K. Begg, "Ultrasonic and Infrared sensors Performance in Wireless Obstacle Detection System", published in *IEEE International Conference on Artificial Intelligence, Modeling and Simulation*, 2013, ISBN 978-1-4799-3251-1.
- [4] Mike Trent, Kevin Laubhan, Ahmed Abdelgawad, Kumar Yelamarthi, "An FPGA-based Portable Real-time Obstacle detection and Notification System", published in *IEEE International Conference on Electrical, Electronics, and Optimization Techniques*, 2016, ISBN 978-1-4673-9939-5.
- [5] Kavita Palaskar, Prof. S.A Shaikh, "Distance Estimation Using Multi-sensor Data Fusion Technique with FPGA", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 4, Issue 8, August 2016.
- [6] Bollam Eswari, N.Ponmagal, K.Preethi, S.G.Sreejeesh, "Implementation of I2C Master Bus Controller on FPGA", published in *IEEE International Conference on Communication and Signal Processing (ICCCSP)*, 2013, ISBN 978-1-4673-4866-9.
- [7] Vidya Venkatesh, Deepthi Dayanand, Shri Kanhu Charan Padhy, "VHDL implementation for design of an I2C interface for temperature sensor and EEPROM Memory", published in *International Journal of Advanced Research in ComputerEngineering & Technology (IJARCET)*, Volume 4 Issue 4, April 2015.
- [8] Jonathan Valdez, Jared Becker, Understanding the I2C Bus, Texas Instruments Application Report, June 2015.
- [9] Xilinx Spartan-3E FPGA Family Data Sheet, DS 312 July 19, 2013.
- [10] Digilent Basys2 FPGA Board Reference Manual.
- [11] PCF8591 8-bit A/D and D/A Converter Module Datasheet, Philips Semiconductor, 2003.
- [12] Philips Semiconductor, UM10204, I2C-Bus Specification and User Manual, Rev. 6 - 4 April 2014.
- [13] SHARP GP2Y0A21 YK Infrared Sensor Datasheet.
- [14] Ultrasonic Range Sensor Module, HC-SR04, Datasheet.



Interfacing of ADC 0809 with FPGA Development Board

Tushar Puri¹ Hemant Kaushal²

School of Electronics

Centre for Development of Advanced Computing

B-30, Institutional Area, Sector-62, Noida, India

Abstract: In nature every sense a human perceives is an analog signal. But the advancements in the computing power of manmade machines are mostly related to digital values. However to bridge this gap Analog to digital converters are made. Now, to interface ADC's with a digital device is another challenging task. Solving such a problem, this paper explains the interfacing of a primitive but famous ADC, ADC 0809 with a field programmable gate array board. The FPGA board used is BASYS 2 with a SPARTAN-3E FPGA installed over it.

Keywords: ADC 0809, FPGA interfacing, finite state machine, Analog to digital conversion for FPGA, hardware integration.

I. INTRODUCTION

Every signal in nature exists in analog form. Analog signal must be transformed into equivalent digital form so that it can be processed by any machine. As, machines understand digital signals only. Also, transmission and processing of analog signals have major drawbacks. Some major ones are requirement of large bandwidths for signal transmission, complexity of circuits required to process these signals, signal attenuation through transmission media and ease of processing digital signals rather than analog signals. To overcome these issues, the analog signals are transformed into digital signals and then transmitted. At the receiver end, the original analog signal can be extracted from the received digital signal. Thus, the conversion from analog to digital form is of prime importance in the field of communication and microelectronics.

Analog to Digital converters are used to transform analog signal into equivalent digital form. There are many types of ADCs, namely flash, sigma delta, dual slope and successive approximation, which are used according to specific application. This paper shows interfacing of an ADC with FPGA. The ADC used for this purpose is 0809 which uses successive approximation for its conversion. The FPGA board used is Basys 2 Spartan 3E FPGA board. The ADC0809 offers high speed, high accuracy, minimal temperature dependence, excellent long-term accuracy and repeatability, and consumes minimal power.

The paper is organised as follows: Section II explains some pre requisites for implementing ADC like its pin diagram and signal description. Section III describes method used for implementing ADC which is done using finite state machine. Section IV shows Simulation Results which include simulated waveforms and RTL schematic of ADC.

II. LITERATURE SURVEY

The pin diagram shows all the signals of the ADC.

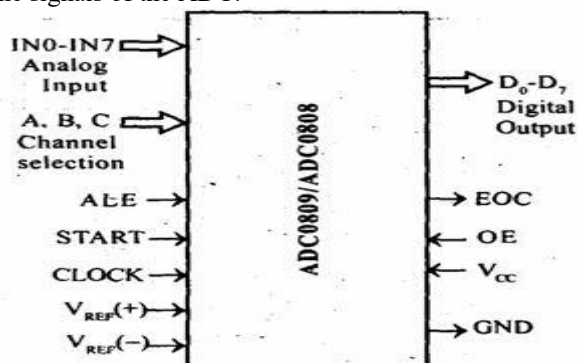


Fig 1: Pin Diagram of ADC 0809[1]

Regrettably, one cannot just connect the ADC with the field programmable gate array and expect analog to digital conversion with digital data as output [2]. There are always some control signals which are required by ADC like clock, start pulse, etc. The ADC 0809 chip has an 8 channel multiplexer, therefore there are three address select lines: A, B, and C. The channel used is 0th one so A, B and C are all set to 0. ALE is required to load the selected address lines into the ADC. The ALE should be pulsed for at least 100ns in order for the addresses to get loaded properly. The clock signal is required to cycle through the comparator stages to do the conversion. There are 8, 8 clock cycle periods required in order to complete an entire conversion. This means that an entire conversion takes at least 64 clock cycles. The maximum frequency of the clock is 1.2MHz. The purpose of the start signal is twofold. On the rising edge of the pulse the internal registers are cleared and on the falling edge of the pulse the conversion is initiated. Like the ALE pulse the minimum pulse width is 100ns. The signal can be tie to the ALE signal when the clock frequency is below 500kHz. So for simplicity we take clock frequency 400kHz and tie these two signals together. The Output Enable signal causes the ADC to actually output the digital values on the output lines. The ADC stores the data in a tristate output latch until the next conversion is started, but the data is only output when enabled. The End of Conversion signal is sent to the FPGA from the ADC. The signal goes low once a conversion is initiated by the start signal and remains low until a conversion is complete. The timing specifications of the ADC indicate that Start and ALE pulse width must be between 100ns to 200ns.[1]

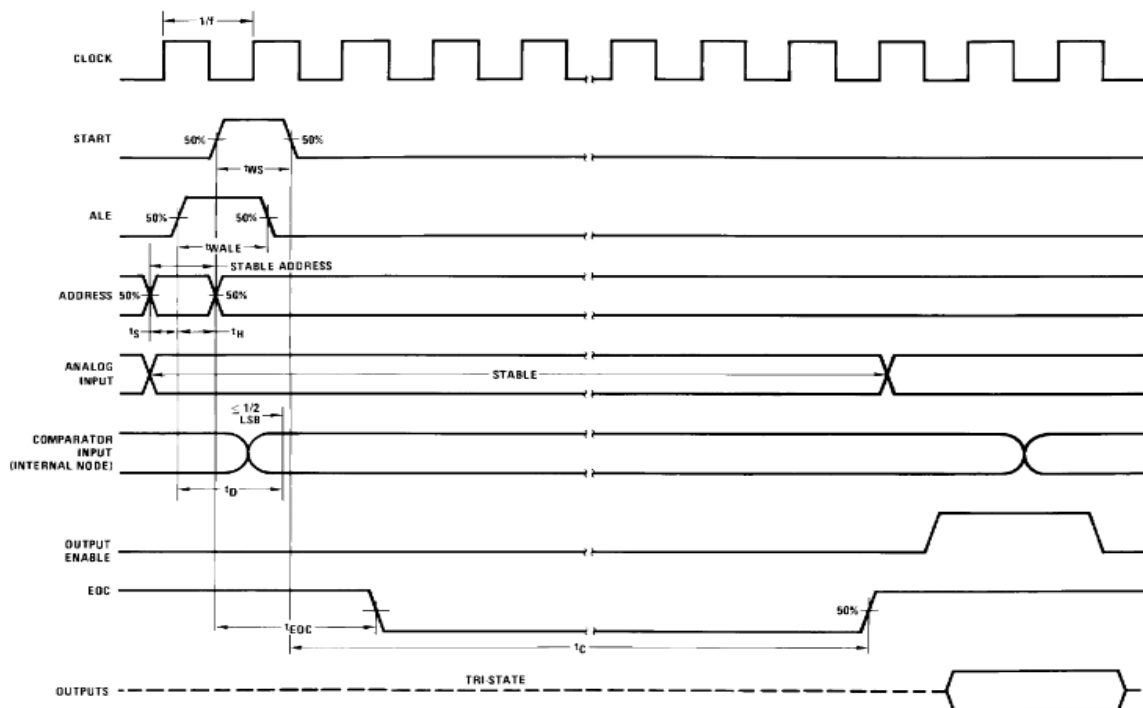


Figure 2: Timing Diagram of ADC 0809[1]

Pin Number	Label	Input/Output	Description	Required
Note: All control signals should have a high voltage from $V_{cc} - 1.5$ to 15V and a low voltage from 1.5V to -0.3V.				
1-5, 26-28	IN0 to IN7	Input	Analog data in for 8 different channels.	No, can tie to ground if no input.
6	Start	Input	It is a control signal from the FPGA, which tells the converter when to start a conversion. It is a pulse of at least 100ns in width.	Yes
7	EOC	Output	Signal from the ADC. It goes low when a conversion is started and high at the end of a conversion. Users can look for a rising edge transition.	Yes

9	Output Enable	Input	Control signal for FPGA that turns the output of the ADC on while high. Useful for handshaking.	No, can tie to Vcc.
10	Clock	Input	Clock signal from FPGA. Max 1.2MHz.	Yes
11	Vcc	Input	Power to the chip. Range 4.5V to 6.0V DC.	Yes
12	V _{REF} (+)	Input	Top rail of Reference voltage. The voltage level that, when received as an input, will output "11111111" to the FPGA. Max Value Vcc + 0.1V	Yes
13	GND	Input	Ground. 0V	Yes
16	V _{REF} (-)	Input	Bottom rail of Reference voltage. The voltage level that, when received as an input, will output "00000000" to the FPGA. Min Value -0.1V	Yes
8,14,15, 17-21	MSB to LSB ALE	Output	This is a bit of the digital converted output. 2 ³ is the LSB.	No
22		Input	Control signal from FPGA. This should be a pulse from the FPGA sent when the address is ready to be loaded into the ADC. The minimum pulse width is 100ns. It can be tied to the Start line if the clock is operated under 500kHz.	Yes
23	ADD C	Input	Control signal from FPGA. This is an address select line for the multiplexer. It is the MSB of the select lines.	No, can tie to ground
24	ADD B	Input	Control signal from FPGA. This is an address select line for the multiplexer. It is the Second bit of the select lines.	No, can tie to ground
25	ADD A	Input	Control signal from FPGA. This is an address select line for the multiplexer. It is the LSB of the select lines.	No, can tie to ground

Figure 3: Signal Description of ADC 0809

- The source resistance must be below 10kohms for operation below 640kHz and below 5kohms for operation around 1.2MHz.[2]
- The analog source must remain stable for 72 clock cycles for correct conversion. [2]
- The resolution is quite low to perform high precision tasks.

III. IMPLEMENTATION

In order to get it to work, there is a total of seven control signals that must be sent from the FPGA. These are the address lines, A, B, and C, Address Latch Enable (ALE), Clock, Start, and Output Enable (OE). There is also one control signal which is sent to the FPGA, it is the End of Conversion (EOC) signal. Also the converted analog-to-digital data from the ADC also acts as input to the FPGA. The following diagram shows simple block diagram of interfacing between ADC 0809 and FPGA.

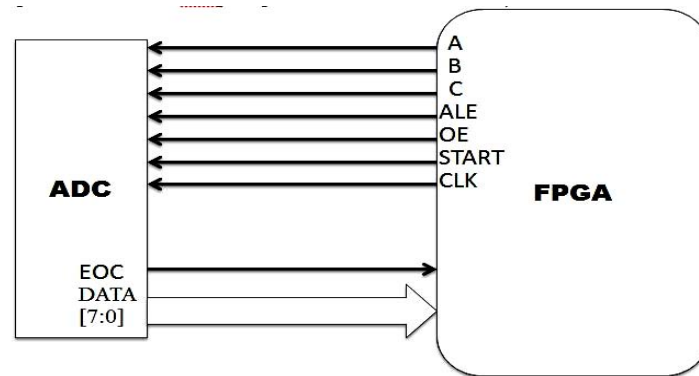


Figure 2: Interfacing between ADC 0809 and FPGA

Any sensor can be connected to the ADC to give it the analog inputs. Some are pressure sensor, ultrasonic sensor, infrared sensor and so on. The inputs to ADC from FPGA are A, B, C, ALE, OE, START, CLK. A, B and C are given permanent ground as we have attached our sensor to 0th channel of the ADC. According to the ADC datasheet ALE and START can be 100-200ns width pulses. We have chosen them to be of 140ns. Also, the signal can be tie to the ALE signal when the clock frequency is below 500 kHz. So for simplicity we take clock frequency (CLK) 400 kHz and tie these two signals together. Inputs to FPGA from ADC are EOC and digitally converted 8-bit data. The analog sensor readings are converted to digital form by the ADC and are given to FPGA for further processing.

ADC is implemented using a finite state machine. There are total 8 states in our FSM. Each state is explained as follows. Figure 3 diagrammatically shows these discussed states.

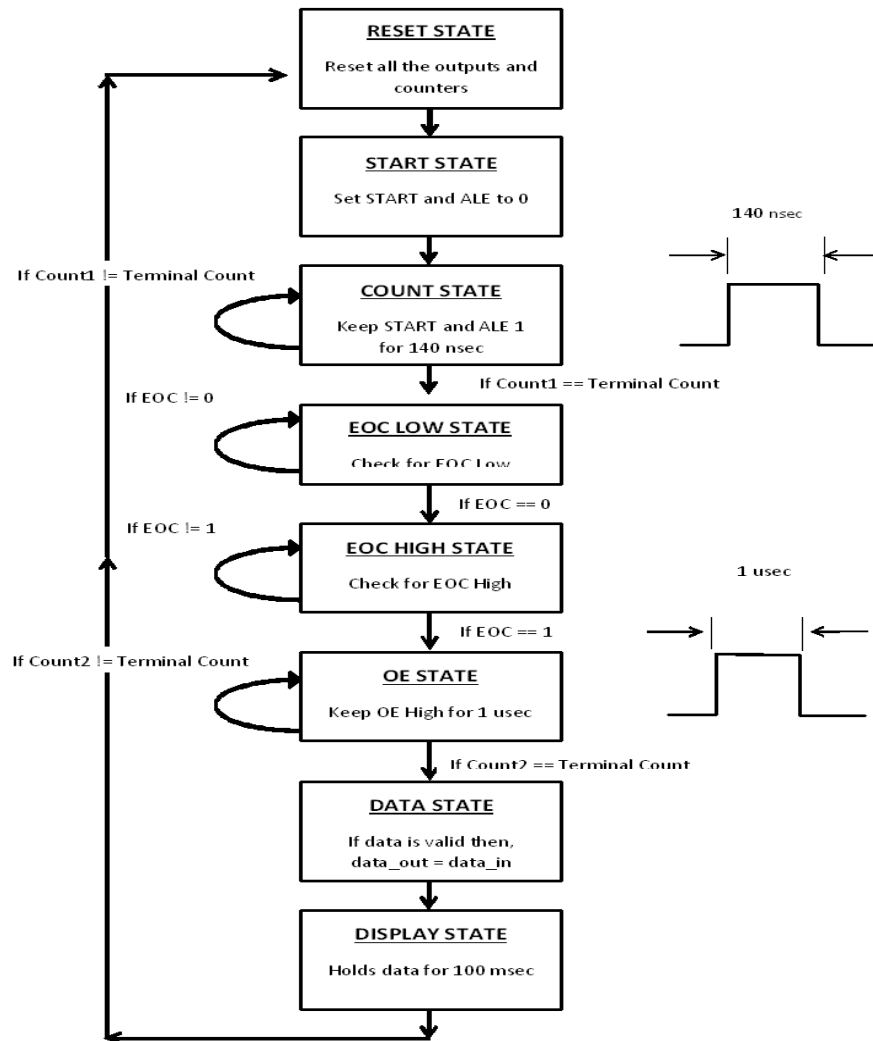


Figure 3: FSM for implementing ADC 0809

- **Reset State:** In this state all the outputs (from FPGA), namely, A, B, C, ALE, START and OE are initialized. Also the internal registers/reg's (counters) used for implementing FSM are initialised in this state.
- **Start State:** Here START and ALE are set to 0 and given to ADC. So that they can be set high in subsequent state and data conversion can start.
- **Count State:** This state generates START and ALE as a pulse of 140ns. ALE latches the addresses and START marks start of the conversion.
- **EOC Low State:** As EOC is low by default, so in this state EOC is checked whether it is low or not. This is done so that ADC works properly and a proper flow in the FSM is established.
- **EOC High State:** When FPGA receives EOC high signal from the ADC, it means that conversion has been done. So, this state checks for EOC high which marks the end of conversion of 1 byte (8 bits) of data.
- **OE State:** After the conversion has happened, OE is set to high and given to the ADC. This signal tells the ADC to store the data in the output latches until the next conversion starts. We have given a 1us pulse to the OE signal so that data is latched properly on the ADC output latches and is stable.
- **Data Acquisition State:** In this state, the data is latched onto FPGA. i.e the data which is saved onto output latches of ADC is saved into a reg in our FPGA.
- **Display State:** In this state, we hold the data onto the same reg for 100msec so that it can be shown on LEDs of the FPGA. More speed than this will make the data invisible or flickering on the LEDs.

IV. SIMULATION RESULT

The simulation and implementation is done using Xilinx ISE 14.2 Design Suite using Verilog HDL. The synthesized top level schematic of the ADC implemented using Finite State Machine (FSM) discussed in previous section is shown in figure 4.

The inputs to ADC are data_in[7:0], clk, eoc, rst. Data_in[7:0] is the input digitized signal coming from the ADC onto FPGA. The ADC FSM works on the on-board clock of 50MHz. A separate clock is given from FPGA to the ADC of 400kHz so that ADC works on the same. The outputs are abc[2:0], data_out[7:0], ale, oe and start. Data_out[7:0] is digitized data of 8 bit which is latched onto FPGA. This is shown on FPGA LEDs. The figure 5 shows the simulated waveform of the ADC module.

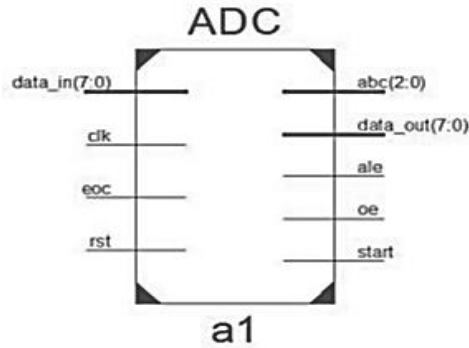


Figure 4: Top Level Schematic of synthesized ADC

In reset state, all the outputs and internal registers are initialized. The system is in reset state when rst signal is low. After getting rst high, ale and start signals are given a pulse of 140ns. Now, if eoc high is detected then oe is set to high and stays high for 1usec. Figure 6 shows further simulation results. After 1usec, oe is again set to low and data_in[7:0] is latched onto data_out[7:0]. Count_display signal is a counter which holds this data for 100msec.



Figure 5: Simulation Result of ADC (Part 1)



Figure 6: Simulation Result of ADC (Part 2)

VII. REFERENCES

- [1] Texas Instrument ADC 0809 Datasheet.
- [2] https://sites.ualberta.ca/~delliott/ee552/studentAppNotes/1999f/ad_converter/
- [3] https://reference.digilentinc.com/media/basys2:basys2_rm.pdf



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Design and Implementation of Population Based Ant Colony Optimisation Algorithm

Namrata Prakash¹ and Dr. Sunita Prasad²

VLSI Design

Centre for Development of Advanced Computing, NOIDA

B-30, Institutional Area, Sector 62, Noida, India

Abstract: In this project, hardware design of population based ant colony optimisation algorithm for travelling salesman problem is developed using Verilog codes. The problem size of the TSP is taken as 4. The PACO hardware is designed such that the population matrix stores the solutions for previous 3 iterations along with the elite solution. The design is first, simulated on the ISE Design Suite 14.5 and the simulation waveforms are checked for the correctness of the PACO algorithm logic. Then this HDL code is synthesised and implemented on Basys2 Xilinx Spartan-3E-100 FPGA Board. The research and development of hardware design for ant colony optimisation algorithm is guided by the thought that dedicated hardware present for the purpose of running ant colony optimisation algorithms for specific problems like travelling salesman problem would speed up the process. This will provide faster results and improve the throughput of the system. The design is implemented takes in rst from the toggle switch and 50 MHz clock frequency from the board and displays the output of the PACO algorithm on the seven segment display section. The distance between the cities in the TSP is hardcoded in the design, so this design is problem specific and not general.

Keywords: PACO; FPGA implementation; ant colony optimisation; TSP; Spartan-3E;

I. Introduction

It is observed that when a trail of ants is moving in straight line along a path and an obstacle is found in their way they branch out and take separate paths around the obstacle to converge on its other side and continue towards their destination. After some time is passed, all the ants still on the former side of the obstacle are observed to be taking the shorter path around the obstacle. This gives the impression that ants are very smart creatures. But the reality is different from this deduction. In the words of Deborah M. Gordon, a biologist at Stanford University, "Ants aren't smart, Ant colonies are." [1] The truth is that it is the entire colony of ants which figures out the shorter path around the obstacle and all the individual ants contribute to get this result but there is nothing intelligent required about their contribution.

This type of behaviour where a group can accomplish tasks unimaginable by an individual is found not only in ants but in other insects, birds and animals like honey bees, termites, fishes etc. This group behaviour is termed as 'Swarm Intelligence'. Swarm intelligence can be defined as the collective ability of a group of agents working in a distributive manner to accomplish a task, where the individual agents follow simple rules and have no clue about the task they are assigned to complete. These agents communicate with the environment in order to improve their performance instead of communicating with each other. [1] This method of communication in which individuals communicate with one another by modifying their local environment. [2] Swarm intelligence works on the principle of stigmergy. Simple creatures follow simple rules, each one acts on local information and has no awareness about the actual task the group is supposed to accomplish. [1] Stigmergy enables complex, coordinated activity without any need for planning, control, communication, simultaneous presence, or even mutual awareness. The resulting self-organization is driven by a combination of positive and negative feedbacks, amplifying beneficial developments while suppressing errors. [3]

It has been observed that ants communicate indirectly by disposing traces of pheromone as they walk along a closed path. The following ants prefer the paths having strongest pheromone information, thereby further increasing the pheromone traces on these paths. Since ants on short paths are quicker, pheromone traces on these paths are increased very frequently. On the other side, pheromone information is permanently reduced by evaporation. This diminishes the influence of formerly chosen unfavourable paths. [4]

Observing the foraging behaviour of ants, a mathematical model of this behaviour was constructed in order to simulate swarm intelligence in artificial agents. In ACO algorithm, a set of software agents called *artificial ants* search for good solutions to a given optimization problem. Ant system algorithm for TSP consists of two phases: Tour construction and Pheromone update. In the tour construction phase, each ant constructs its own

tour by visiting all the cities once. In the pheromone update phase, pheromone evaporation and intensification is done.[5] This may sometimes lead to a situation in which all the ants follow the same tour, because of the excessive growth of pheromone trails on arcs of a good, although suboptimal, tour. This is known as stagnation. So, the algorithm was modified by applying a more aggressive action choice rule. The improved algorithm, known as Ant colony system algorithm, exploits the experience gathered by the ants during their tour construction phases more efficiently. It uses local and global update of pheromone values.[6] In the global update, pheromone evaporation and pheromone deposit take place only on the arcs belonging to the best-tour. In the local update, each time an ant uses an arc, it removes some pheromone from the arc to increase the exploration of alternative paths. This helps in removing stagnation.

II. Population based Ant Colony Optimisation (PACO) Algorithm

The need for the development of PACO algorithm came into being because FPGA implementation or hardware realisation of ACO algorithms is very difficult due several restrictions. The multidimensional array type of data structure is not permitted in Verilog. Hence the storage of 2D population matrix is not that easy in hardware description. So, the storage of population matrix requires a long 1D array to store the 2D array elements either row wise or column wise. Evaporation and integration of heuristic information requires multiplication operations but with the computational resources available on fine grained FPGAs it is very difficult to achieve the required design for large problem size. In order to apply the pseudo random probability rule for making the selection choice for the next city to be chosen in each tour construction, floating point division is required. This is not efficient and very complex to implement on the hardware.

Hence, population based ant colony (PACO) algorithm was developed to create hardware design for ACO in a simplified manner. In PACO, a population matrix is used to store the solutions of previous iterations to keep an account of the cities visited by the ants and use high frequented cities in order to achieve the selection of strong pheromone paths.

Pseudocode for PACO algorithm for solving TSP

Until maximum iteration limit is exceeded [4]

Step 1: Place each ant on any random city i and place this city in the neighbourhood list for that ant.

Step 2: Until all cities are visited by each ant

Select next city j for each ant based on the probability distribution for the corresponding neighbourhood list

The probability distribution is obtained by counting the number of previous occurrences of j in the population matrix

Remove the selected city for each ant from its corresponding neighbourhood list.

Step 3: Select best tour from the tours generated by m ants and check for elite solution

Step 4: Update population matrix (insert the current best tour and update elite solution if required)

Step 5: Increment iteration

End

III. Proposed Design for implementation of PACO algorithm

This is the topmost module in PACO architecture. It describes the interconnections between the sub-modules used in the hardware design.

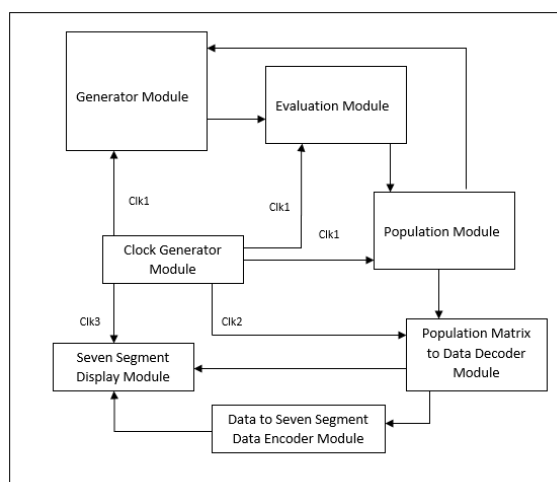


Figure 1: Block Diagram of PACO Module

The above block diagram shows the interconnections of the various modules in the top module. Three different clock frequencies are used in this design. The block diagram shows the different clock signals used in the design

for different submodules. Here $\text{clk1}=5\text{Hz}$, $\text{clk2}=1\text{Hz}$ and $\text{clk3}=1\text{kHz}$. The algorithm starts working from the generator module. As the reset input rst is initially at zero logic, all the signals and registers are initialised. When rst is set to one. The generator module starts the tour construction procedure and generates the three tours t_1 , t_2 and t_3 corresponding to each ant and sends these to the evaluation module along with the control signal e for controlling the working of the evaluation module. The evaluation module takes the three tours and calculates the total distance travelled in each tour and compares them to find the best solution for the iteration and the this pbest is checked to be the elite solution. In case the pbest is found to be the elite solution, the is_el signal is set to high. This pbest and is_el are sent to the population module along with the control signal upd . The population module updates the population matrix stored in a 48 bit register in FIFO manner. It also updates the pheromone matrix stored as 5 bit array of size sixteen. Pheromone is incremented corresponding to the solution entering the matrix Q and decremented corresponding to the solution leaving the matrix Q . Then it selects the query q_{ih} and sends it to the generator module along with the control signal qr . It also counts the iterations and sends the stop signal stp to terminate the ACO algorithm if the itr signal is $5'b1000$. The population matrix to data decoder takes the 48 bit Q as input and sends four 3 bit signals data1 , data2 , data3 and data4 as output to the data to seven segment data encoder module for the proper display of the population matrix Q on the seven segment display. It works on the clock signal clk2 . The data to seven segment data encoder module converts these four 3 bit data into four 8 bit data ssd1 , ssd2 , ssd3 and ssd4 to be sent to the cathode input of the seven segment units in the seven segment display module. The display module takes in ssd1 , ssd2 , ssd3 and ssd4 and sends each one of these to separate anodes of the seven segment units for refreshing at clock signal clk3 . The generator module evaluation module and the population module work on the clock frequency clk1 .

The functioning and design of each of these submodules is explained below.

The hardware design for PACO needs a single clock frequency for proper working of the PACO algorithm but it requires three clock signals for the correct display of the population matrix Q on the seven segment display. Thus, clock generator module is designed to generate three different clock signals of frequency 5Hz, 1Hz and 1kHz from the input clock signal of frequency 50MHz taken from the FPGA board.

The clock generator module has three different counters, each for generating a different clock signal. The counter counts the number of clock cycles arrived of the input clock signal. The output clock signals are generated by keeping the input clock signal high and then low for the required number of input clock cycles. The required number of clock cycles of the input signal clk for which the output clock signals clk1 , clk2 and clk3 are to be set high or low is calculated as described below with the help of three different counters. The clock outputs clk1 , clk2 and clk3 initialised to $1'b1$ at rest and they are toggled whenever the counter resets to 0 after counting the required number of cycles. This is how the three clock signals are generated.

The generator module is one of the three main modules for the PACO algorithm design. This module consists of three solution generators having their corresponding random generators and an and gate for generating the control mechanism for the evaluation module. Here three solution generators are taken as 3 ants are assumed in this design. Each solution generator constructs a tour in one iteration and then sets output signal t to logic 1. The and gate takes as inputs the output signals t_1 , t_2 and t_3 generated by each solution generator and generates the output e which controls the state transitions in the evaluation module at the arrival of positive edge of clock. As the three solution generators in this module construct separate tours for each ant and send it to the evaluation module. The need for 3 different random generators is there because the design three solution generators work concurrently. So, if only one random generator is used then the solution one generated for each solution generator would be same (sl1 , sl2 and sl3). This is undesirable in the algorithm. Hence, three different random generators are included in the design.

Each random generator consists of 3 t flip flops and 3 xor gates connected in different arrangements to generate different random numbers at each clock pulse. In the PACO design, only four output values which are $3'b001$, $3'b010$, $3'b011$ and $3'b100$ are desired from the random generator but the logic implemented by the 3 toggle flip flops and 3 xor gates can generate any 3 bit value. So, another module named as rnd is instantiated in the random generator module to convert the output generated by the combination of flip flops and xor gates to the desired four values. These are basically the codes for city1 , city2 , city3 and city4 in the TSP. The rnd module decrements the values generated by the flip flop and xor gate logic values by 3 if they are greater than $3'b011$ and otherwise increments these values. The output given by rnd is sent as the output of random generator module to be stored as the first selected city for each tour.

The solution generator module is made up of a synchronous finite state machines working at clock signal clk1 . In this design, the fsm stays at state s0 if $\text{rst}=0$ or $\text{stp}=1$. When $\text{stp}=0$ and $\text{qr}=0$ the fsm is at state s0 , then it waits for the completion of the current iteration and the setting of the qr signal to high logic. The qr signal generated by the population module controls the working of the generator module by controlling the transition from state s0 to state s1 in the solgen module. For first iteration the fsm goes to s1 without waiting for $\text{qr}=1$

The functionality of different states is summarised as follows:

S0 : Set get_sl1 high to trigger the random generator to receive sl1 and $t=0$.

S1 : Assign neighbourhood cities n11 , n12 , n13 according to the sl1 .

S2-S6 : Calculate pij for neighbourhood cities.

S7 : Assign sl2 according to pij11, pij12, pij 13

S8 :Assign neighbourhood cities n21, n22

S9 -S13: Calculate pij21,pij22

S14: Assign sl3 and sl4 according to pij21, pij22

S15: Assign sl1, sl2 , sl3, sl4 to tour and t=1

The evaluation module takes the three tours and calculates the total distance travelled in each tour and keeps them as d1, d2 and d3. Then it compares d1, d2 and d3 to find out the shortest tour. This shortest tour is stored in pbest and compared with the least tour distance so far stored in the register best_tour. In case the pbest is smaller than the is_el signal is set to high. It sends this best solution pbest and the is_el signal as output to the population module.

This module is made up of a synchronous finite state machines working at clock signal clk1.

The following table shows the state transitions in the fsm

In this design, the fsm stays at state s0 if rst=0 or stp =1. When stp =0 and e=0 the fsm is at state s0, then it waits for the completion of the current iteration and the setting of the e signal to high logic. The e signal generated by the generator module controls the working of the evaluation module by controlling the transition from state s0 to state s1 in this module.

The following function are done by each state:

S0 : Set upd=0, is_el =0 Initialize pbest, dbest, best_tour.

S1-S2 : Total distance of each tour is calculated

S3 : Calculate pij for neighbourhood cities.

S3 : find pbest and dbest

S4 : update best tour and assignis_el

S5 : set upd=1

The population module stores the stores the population matrix Q in the form of a 48 bit register. It consist of 12 bit elite solution and three 12 bit solutions for past three iterations. This module updates the Q according to the pbest and is_el signals received and updated tauij according to the solution entering or leaving the Q and sends qih and qr to the generator module for the starting of the next iteration. It also counts the number of iterations and sets the stop stp signal high so that finite state machines in all the other modules get stuck at their reset states or the state specified in the design.

In this design, the fsm stays at state s0 if rst=0 or stp =1. When stp =0 and upd=0 the fsm is at state s0, then it waits for the completion of the current iteration and the setting of the upd signal to high logic. The upd signal generated by the evaluation module controls the working of the population module by controlling the transition from stse s0 to state s1 in this module. The following function are done by each state:

S0 : Set qr=0

S1: counter and count are incremented (count counts from 1-3 counter counts from 1-4)

S2-S3 :Incremettauij for solution entering Q

S4-S5 :Decremettauij for solution leaving Q.

S6 : assign pbest to Q (The bit location is calculated by counter) and update elite solution if is_el is high

S7 : assign qih and increment itr

S8 : set qr=1 set stp=1 if itr=5'b10000

The population matrix to data decoder module decodes the data to be sent to the seven segment display. Here the 48 bit data has to be sent to four seven segment display nodes. As the population matrix Q is 4x4, so each column of q can be seen on the display at once.

So an fsm is designed to send each row to the display panel at once at the clock signal clk2

In this design, the fsm stays at state s0 if rst=0. When stp =1 the fsm enters into an infinite loop of state s4 and s5. When rst sets to 1 the fsm starts with state s0, then it moves to state s1, then s2 then s3 and back to s0 until stp is set to high. When stp=1 is encountered, the fsm goes to state s4 through s3 and gets stuck between s4 and s5. The task of this module is to send the proper data for display on the seven segment display units from the 48 bit Q. State s0 sends the first row of the population matrix to the seven segment display, s1 sends the second row, s2 sends the third row and fourth row. When stp signal is high, s4 sends the elite solution to the seven segment display. Then s5 sends all ones to the seven segmentdisplay to set it off. So, the display flashes the elite solution on and off when the stp signal gets high. This module is made up of a synchronous finite state machine working at clock signal clk2.

The data to seven segment data encoder module continuously encodes the four 3 bit data, sent by the population matrix to data decoder module to four 8 bit data and sends it to the seven segment display displaymodule. This module is very necessary to design as the seven segment display takes 8 bit value to set the corresponding 8 leds on or off. The basic job of this module is to send the correct data at the cathode data input of seven segment display. So that the correct numbers appear on the display. Seven segment unit consists of 8 leds named as a, b,

c, d, e, f, g, dp. The correct sequence of zeroes and ones is needed to be sent to the seven segment input so that the correct number corresponding to the output generated gets displayed on the FPGA board.

The seven segment display module contains a synchronous fsm working at clock signal clk3 for seven segment display. This is required as the only one unit of the seven segment display can stay on at a particular time. So, each of the four seven segment units has to be refreshed within 1ms to 16 ms, so that it appears to the viewer as if the display is continuously on. This is why clk3=1kHz is used.

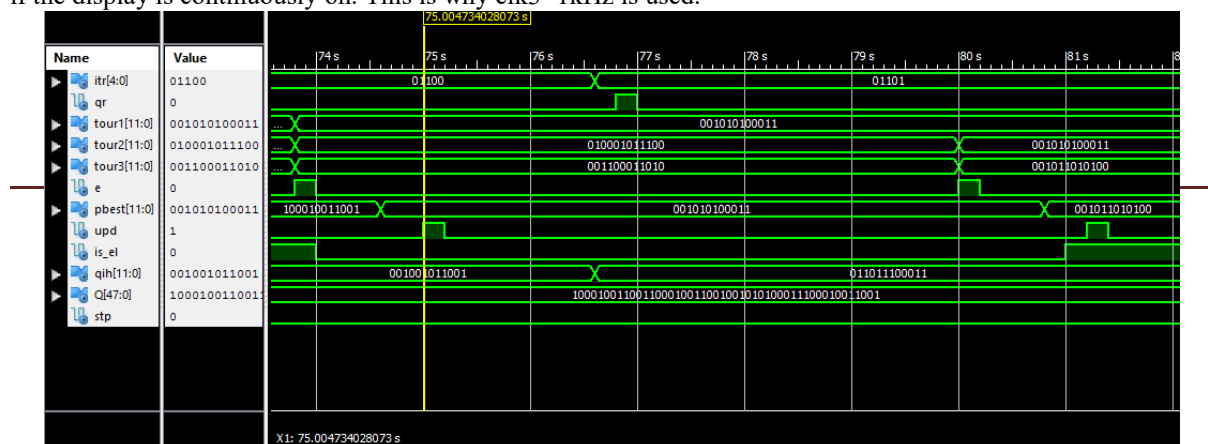


Figure 2: Simulation waveforms of PACO module from 74-81 sec

IV. Conclusion and Future Work

In this project hardware design for the implementation of Population based Ant Colony Optimisation Algorithm is generated. The correctness of the design is verified from the simulation results run on the ISE Design Suite 14.5 and then the design is synthesised and implemented on the Basys2 Xilinx Spartan-3E-100 FPGA Board. The hardware design is found to be working correctly on the FPGA board. The required output is obtained on the seven segment display present on the board.

Table 1: Design Utilization Summary

Logic Utilization	Used	Available	Utilization
Number of slices	667	960	69%
Number of slice flip flops	700	1920	36%
Number of 4 input LUTs	1258	1920	65%
Number of bonded IOBs	20	83	24%
Number of GCLKs	2	24	8%

The utilisation of resources can be reduced by removing the tauij register and tauij increment and decrement logic in the population module design. This would reduce some states in the population module fsm and also there will be no need to store a 5 bit 1D array with 16 entries. This would reduce the device utilisation.

Apart from the removal of tauij and its updation logic, efforts can be applied to make the design faster by getting the generator module, evaluation module and the population module work simultaneously. The design can be made to work such that when population module is updating the Q in the i^{th} iteration, the evaluation module is evaluating tours generated in the $i+1^{\text{th}}$ iteration and the generator module is constructing tours for the $i+2^{\text{th}}$ iteration. This would speed up the design and improve the performance of the PACO hardware.

References

- [1] <http://ngm.nationalgeographic.com/2007/07/swarms/miller-text>
- [2] <http://journal.media-culture.org.au/0605/03-elliott.php>
- [3] Francis Heylighen, "Stigmergy as a Universal Coordination Mechanism I: Definition and Components", Cognitive Systems Research, Elsevier, Volume 38, June 2016, Pages 4–13
- [4] B. Scheuermann, K. Sob, M. Guntch, M. Middendorf, O. Diessel, H. ElGindy, H. Schmecka, "FPGA implementation of population-based ant colony optimization" Elsevier, Applied Soft Computing Volume 4, Issue 3, August 2004, Pages 303–322, Hardware Implementations of Soft Computing Techniques
- [5] http://www.scholarpedia.org/article/Ant_colony_optimization
- [6] Marco Dorigo and Thomas Stützle, "Ant Colony Optimization", 2004, A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

VHDL Implementation of AES with Random Delay to Resist the Power Attack and To Confuse With DES

Arvind Kumar Singh¹, BM Suri² and SP Mishra³

Scientists, SAG, DRDO, Metcalfe House, Civil Lines, Delhi-110054

Abstract: VHDL implementation of 128 bits AES with random delay is described in this paper. This does not only resist the Correlation Power Analysis (CPA) Attack but also confuses with the processing of DES algorithm in terms of their power patterns. This is achieved by inserting six random clocks as delay. These extra random clocks are distributed in beginning and end of encryption using two bit random number. The value of the random number decides the number of extra clocks before and after the AES execution processing. During these extra clocks, processing of dummy functions of AES is being carried out. This implementation considerably resists CPA attack as well as confuses with the DES processing as power patterns of both appear very similar. None of the key bytes is being extracted by mounting CPA attack on 20000 power traces acquired from FPGA (Spartan 6) board processing the VHDL module of AES with random delay at 4MHz.

Keywords: AES, random delay, correlation power analysis attack, DES, power pattern, confuse, traces

I. Introduction

The secrecy of the crypto systems is becoming vulnerable with advent of Side Channel Attack (SCA) techniques like Power Attack, EM Attack, Timing Attack, etc. Adversary measures power consumptions [1], EM radiations [2], execution time [3], etc. of crypto systems and analyses them to extract the secret key. SCA has been proved very efficient in breaking the crypto algorithms like AES [4], [5] which are undefeatable so far by the classical cryptanalysis technique. It is reported [6] that all 16 key bytes of AES are being determined by mounting Correlation Power /EM Analysis (CPA/CEMA) Attacks on appropriate number of power consumption / EM radiation patterns. The time and computations required in breaking the crypto systems with the help of SCA techniques is surprisingly very less. CPA attack requires less than 30 minutes time to extracts all key bytes of AES using 600 power traces [6].

The extraction of secret key of crypto systems by just measuring and analysing these leakages has seriously troubled the crypto systems designers. This motivated them to come out with newer ideas and techniques to counter these kinds of attacks. As a result of continuous efforts made, many techniques (countermeasures) [7], [8] have been developed and reported in the literatures. Different types of countermeasures offer different level of resistance against attacks. An ideal countermeasure would protect a system and prevent any kind of attack, which is impossible [9]. In real world scenario, specific types of countermeasures are used for specific types of crypto algorithms. Their aim is to offer resistance to a great level against success of attacks, instead of preventing an attack completely. Countermeasures used against SCA are classified into three categories [9] as per the level of resistance offered by them. To stop the leakage of information in the first place, thus to close the side-channel, first type of countermeasures are used. Working principle of second category of countermeasures is to decrease the information leakage by typically reducing the signal to noise ratio of the side-channel. Different kind of approach is adopted by the third category of countermeasures. Their goal is not to prevent or minimize side-channel leakage, but to live with it. The basic idea they follow is to allow side channel leakage but to prevent the exploitation of leaked information by any attack.

Selection of countermeasures is based on strength of security required as per the need of application, performance, availability of resources, computational overhead, cost, etc. Introducing countermeasures increase security strength of crypto systems at the cost of increased computational overhead and resources requirement, which ultimately increase the price of the crypto systems and also may degrade the performance. Therefore a countermeasure is selected /designed keeping in mind the performance, strength of security required and cost. This can be achieved by developing optimized countermeasures depending on specific requirement/use to provide sufficient strength without degrading much system performance and with minimal extra overhead.

To achieve the sufficient security strength with respect to SCA without increasing much extra overheads, the use of random clocks and random delay, changing the sequence of operations, adding of noise [1], [2] through extra computations, etc have already been reported. Introducing random delay increases execution time and hence makes the system sluggish. This problem can be resolved by using random clocks for different encryptions. But

in this case, it is possible to identify the different random clocks used for encryptions with the help of signal processing and spectrum analysis. This enables adversary to separate power traces and plaintexts/cipher texts of different encryptions performed at different clocks. After this, adversary can mount the attack to extract the key. Hence the obvious choice is random delay though it makes system sluggish.

In this work random delay as a countermeasure has been used during the AES encryption process. Particular numbers (six) of delay clocks have been introduced in specific manner. These six delay clocks are distributed in beginning and end of AES algorithm processing using two bits random number. Six delay clocks have been inserted purposely to extent AES execution time from 10 clocks to 16 clocks. Thus AES with six delay clocks produces power patterns almost similar to the power patterns of DES performing 16 rounds processing in 16 clocks. Therefore, AES with six delay clocks does not only resist the CPA attack but also confuses with the DES algorithm execution in terms of power patterns. This misguides adversary in wasting time to mount SCA assuming DES processing if the plaintexts/cipher texts are not clearly available corresponding to each encryption/decryption.

The rest of the paper is organized as follows: Section 2 describes VHDL implementation of AES with random delay. CPA attack is explained in section 3. Section 4 contains results and conclusions.

II. VHDL implementation of AES with random delay

AES (128 bits) algorithm with random delay is implemented on FPGA (Spartan 6) board (Sakura G) using the lab setup as shown in figure 1. The lab set up contains FPGA board, oscilloscope (DPO 7254), workstation equipped with Xilinx tool and Oscopce utility. Bit file of VHDL logic module of AES with random delay, shown in figure 2, performing multiple encryptions is ported to the FPGA board through JTAG port using Xilinx tool. At start of each encryption, a signal is generated and sent to the one input of oscilloscope as trigger signal. Power signal from FPGA board is fed to another input of oscilloscope. Oscilloscope captures the power signal at the arrival of trigger signal [6]. Oscopce utility running at workstation copies data from oscilloscope to its memory with the help of LAN/USB cable. Sufficient delay (1.5s) has been inserted between consecutive encryptions. This enables the safely (without overwriting issue) transfer of data from oscilloscope to the workstation memory.

First, VHDL logic module of AES generates 10 rounds keys using the main key in the 10 clocks. Then control switches to the start of encryption process. Before starting the encryption, two bit random number (nc) and trigger signals are set. The nc and 16-nc clocks are used for dummy AES processing before the start and after the end of actual AES processing respectively.

First nine rounds of AES execute identical functions namely substitute bytes, shift rows, mix column, and add round keys. Last round (tenth) excludes mix-column operation. Dummy AES executes all operations of AES using separate plaintext and key. Insertion of six extra delay clocks extends the module execution time from 10 to 16 clocks. This results in producing power patterns similar to that of DES algorithm completing 16 rounds of processing in 16 clocks.



Figure 1 Lab setup

Simulation window of VHDL module is shown in the figure 3. It shows clock, trigger, inputs (plaintexts), keys signals etc. It is clearly visible that trigger signal remains high for 16 clocks duration in which the processing of AES with random delay completes. After delay of five clocks, the trigger signal again gets activated for next encryption. Here the delay of 5 clocks between adjacent encryptions has been inserted for simulation purpose only. But actual number of clocks needed for 1.5s delay has been used during implementation phase. At the end of the delay, the plaintexts, random number (nc) used for delay and key for dummy AES are being updated. The next encryption starts with the updated signals.

Figure 4 depicts plot of the power traces for single encryption of AES without random delay and with random delay. The power trace of DES algorithms completing execution in 16 clocks is also illustrated in figure 4. Power

trace of AES without random delay exhibits 10 patterns corresponding to ten rounds of processing performed in 10 clocks. Power traces of AES with random delay shows sixteen patterns and power trace of DES also shows sixteen patterns corresponding to sixteen rounds of processing performed in 16 clocks. Very similar patterns available in power signals of AES with random delay and DES misguides adversaries. Adversary can assume processing of DES algorithm in place of AES.

Figure 2 VHDL Module of AES with random delay

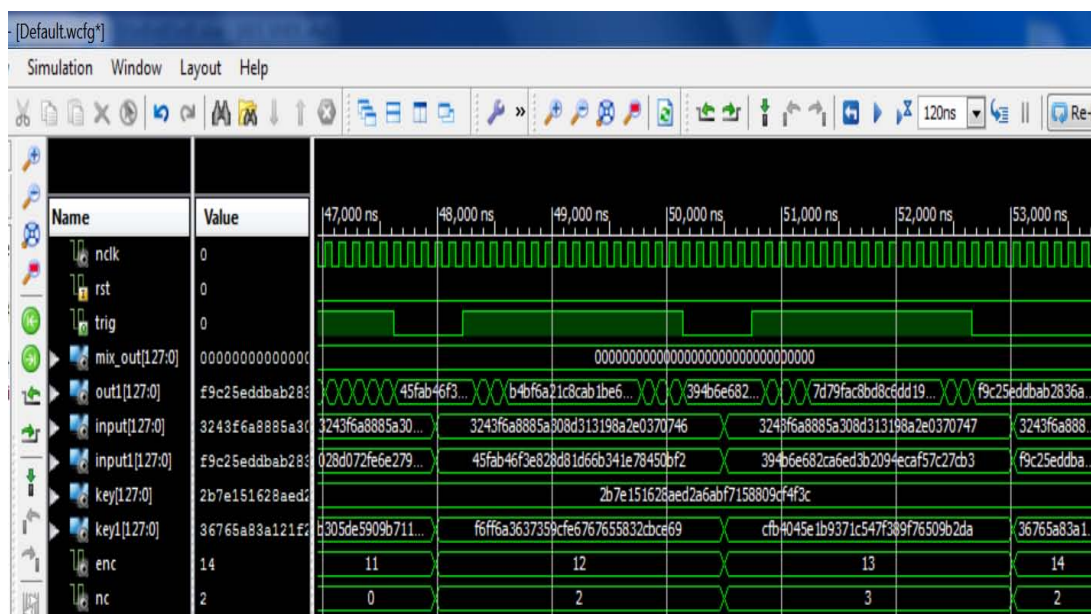
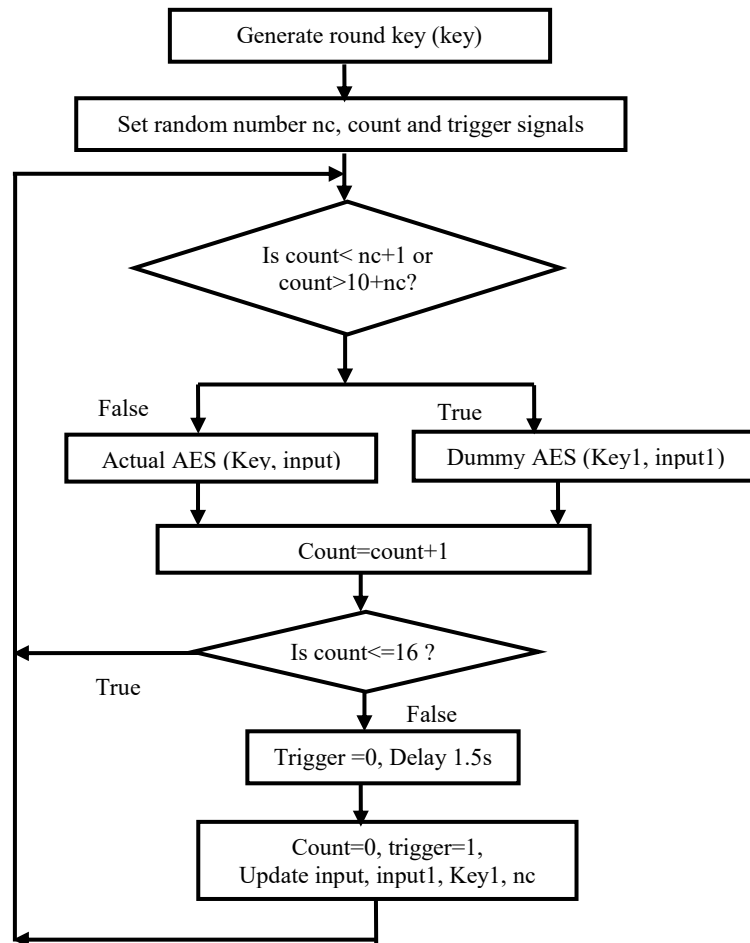


Figure 3 Simulation window of VHDL module

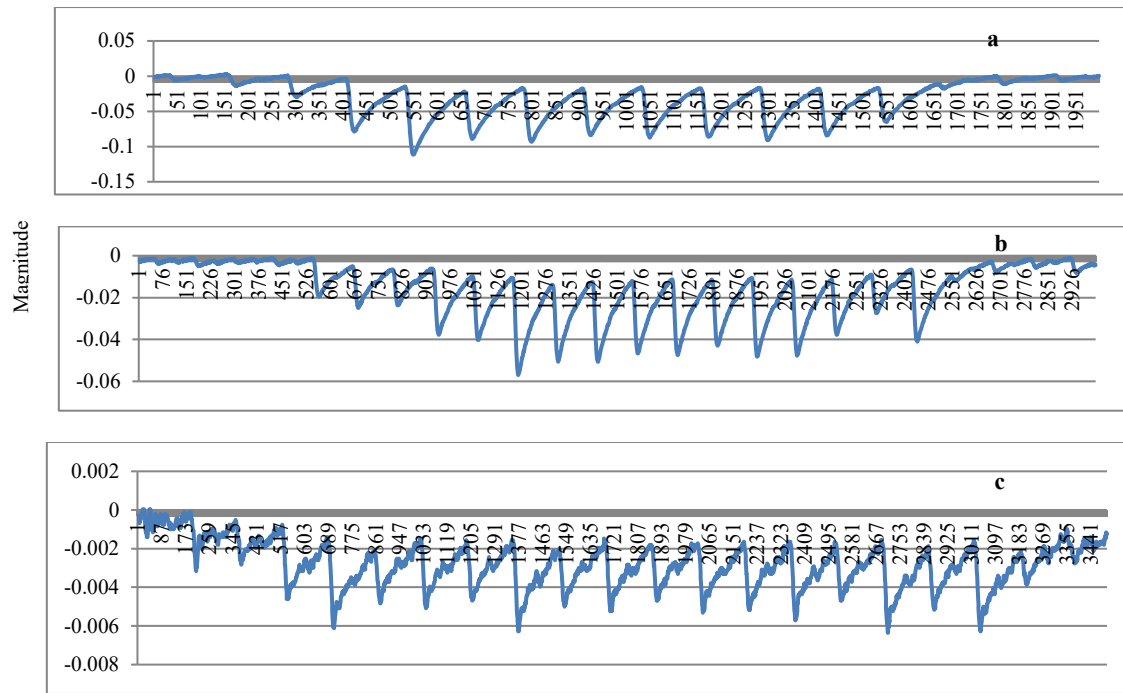


Figure 4 Power signal plot of a) AES algorithm without random delay b) AES with random delay c) DES algorithm

III. CPA Attack

CPA attack [10] is the efficient version of Differential Power Analysis (DPA) [1] attack. DPA attack segregates power traces of several encryptions into two groups based on the intermediate values and sees the peak in differential trace for correct guessed key. But CPA attack compares actual power signals with estimated power leakages computed from intermediate values involving possible key bytes. It requires an appropriate model to estimate the leakages. Hamming Weight (HW) and Hamming Distance (HD) models are being widely used. According to HW model, number of 1 present in the intermediate value is proportional to the power consumption. HW model is used for smart card and microcontroller based systems. HD model is usually used for CMOS based devices. According to it, leakages are proportional to the, switching activity in CMOS devices, number of 0 to 1 and 1 to 0 transitions [6], [11]. It assumes equal amount of power consumption in both types of transitions. Leakage in HD model is equal to the exclusive-or of two consecutive values of intermediate variables. The example of power leakage approximation methodologies of HW and HD models are given in table 1.

Table 1 Example of leakage approximation by HW and HD models

Model	Intermediate values	Approximation of Power leakage
HW	01011001	01011001 \Rightarrow 4
HD	00001111	00001111 \oplus 11000000 = 11001111 \Rightarrow 6
	11000000	

After the estimation of power leakages and measurement of actual power consumptions for sufficient number of plaintexts, CPA attack is mounted. CPA attack is generally mounted on first round or last round as it is easy to compute the intermediate values due to knowledge of plaintext or cipher text. CPA attack performs comparison between estimated and measured leakages by computing correlation. Highest correlation coefficient, Pearson's correlation coefficient [12], between measured and estimated power leakages is expected for correctly guessed key bytes. Basically, it computes correlation for all 256 possible values of a key byte and then it searches for the maximum correlation value. The guessed key byte which results in maximum correlation value is considered as the actual key byte being used by crypto algorithm. This process is repeated to find all 16 key bytes of AES. The values of Pearson's correlation coefficient for first 4 rank key bytes are given below. Rank 1 key byte (correct guessed key byte, 0xD0) has highest correlation value.

- rank: 1, candidate: 208 (0xD0), confidence: 0.1567
- rank: 2, candidate: 225 (0xE1), confidence: 0.1528
- rank: 3, candidate: 143 (0x8F), confidence: 0.1420
- rank: 4, candidate: 155 (0x9B), confidence: 0.1416

IV. Results and conclusions

Sufficient number of power traces corresponding to number of encryptions performed using random inputs (plaintexts) is required to extract the key by mounting CPA attack. Parts of power traces associated to the last round of AES processing are used to mount the CPA attack. The results of mounting the CPA attack on power traces of AES (running at Sakura G board, Spartan 6) without random delay and with random delay are given in table 2. Power traces are acquired at 500MS sampling rate when processing was performed at clock frequency of 4 Mhz. Table 2 shows that all sixteen key bytes of AES have been extracted by mounting CPA attack on 2100 power traces of AES running without any random delay. None of the key bytes has been determined by mounting CPA attack on 20000 power traces (acquired in one working day) of AES running with random delay (six clocks) inserted in beginning and end of encryption based on two bit random number. This shows the level of resistance offered by random delay countermeasure against CPA attack on the cost of 60% increase in execution of encryption time. The insertion of six clocks as delay does not only increase the resistance against CPA attack but also provides an important feature. The power pattern of the AES with random delay of 6 clocks becomes very similar to the power pattern of the DES. This misguides adversary. Adversary waste his time in mounting the attack assuming DES algorithm if the plaintexts/cipher texts are not available clearly corresponding to each encryption/decryption.

Table 2 Result of CPA attack

S. No	AES Module	No of Power traces	Key bytes retrieved
1	Without delay	2100	All key bytes
2	With random delay	20000	No key byte

Acknowledgement

Authors would like to thank Smt Sunita Maithani, Sc 'G', Divisional Head and Smt Anu Khosla, Director, SAG for their support, motivation and providing opportunity and infrastructure to complete this work. Thanks also go Sh Akash Gupta, Sc 'C' and Ms. Purnima Hansda, Sc 'D' for providing power traces of DES.

References

- [1] Kocher, P.C. & Jaffe, J., Jun, B. 1999. Differential Power Analysis. Springer, Heidelberg In: Wiener, M. (ed.) CRYPTO 99. LNCS, vol. 1666: pp. 388-397,
- [2] Gandolfi, K., Mourtel, C., Oliver F. 2001. Electromagnetic Analysis: Concrete Results. In the proceedings of the workshop on Cryptographic Hardware and Embedded Systems 2001, LNCS 2162 Paris, France: pp. 251-261
- [3] Kocher, P. 1996. Timing attacks on implementations of Diffie-Hellman, RSA, DSS and other systems. in Advances in Cryptology: Proceedings of CRYPTO'96, N. Koblitz, Ed., vol. 1109 of LNCS, pp. 104-113.
- [4] Advanced Encryption Standard, http://en.wikipedia.org/wiki/Advanced_Encryption_Standard
- [5] Stallings, W. 2003. Cryptography and Network Security: Principles and Practice. 3rd edition: Prentice Hall.
- [6] Singh, A.K., Mishra, S.P., Suri, B.M., & Khosla, Anu. 2015. "Investigations of Power and EM Attacks on AES Implemented in FPGA", 2nd international conference on Soft computing in Problem Solving, SocProS-2015, held at IIT Roorkee.
- [7] J. Borst. Block Ciphers: Design, Analysis and Side Channel Analysis. PhD thesis, K.U. Leuven, September, 2001.
- [8] S. Mangard. Securing Implementations of Block Ciphers Against Side Channel Attacks. PhD thesis, Institute for Applied Information Processing and Communications (IAIK), TU Graz, 2004.
- [9] Benedikt Gierlichs. Statistical and Information-Theoretic Methods for Power Analysis on Embedded Cryptography, PhD thesis, K.U. Leuven, 2011.
- [10] Brier, E., Clavier, C. & Olivier, F. 2004. Correlation power analysis with a leakage model. In the proceeding of CHES 2004. Springer, Hiedelberg, LNCS, vol. 3156: pp. 16-29.
- [11] Mestiri, H., Benhadjyoussef, N., Machhout, M., Tourki, R.: A Comparative Study of Power Consumption Models for CPA Attack. In I.J. Computer Network and Information Security. Pp.25-31, 2013
- [12] Clarke, G. M. & Cooke, D. 1998. A basic course in statistics, Arnold London, 4th edition.



DESIGN AND IMPLEMENTATION OF FPGA BASED USB 2.0 CONTROLLER

Bhavya¹, Niharika²

School of Electronics

Centre for Development of Advanced Computing

B-30, Institutional Area, Sector-62, Noida, India

Abstract: USB stands for Universal Serial Bus. Our project proposes a novel design for 8 bit binary data transmission and reception using USB2.0 protocol. We implement physical layer and protocol layer of both the transmitter and receiver using FSM. The physical layer of transmitter incorporates NRZI Encoder and Bit Stuffer to improve the performance of USB 2.0. The correctness of the design is verified by implementing it on FPGA BASYS2 Board. We also use the Double Dabble Algorithm to convert the 8-bit binary value into BCD for seven segment display.

Keywords: Universal serial bus, FPGA, Verilog HDL, Implementation, Design, Application specific integrated circuit.

I. INTRODUCTION

USB is an input/output interface standard. It is connection between personal computer (PC) and peripheral devices. USB2.0 replaces different communication protocols with high data transfer rate. Basically, USB2.0 provides three types of data transfer speeds such as low speed(1.5 mbps), full speed(12 mbps) and high speed(480 mbps). It has many advantages includes low cost, easiness of use, hot plug capability, simple construction, auto-configuration, low cost, expandability and outstanding performance. The developments in the electronics industry are aimed to make devices as small as possible and to get them to market quickly. So the designers focus on FPGAs rather than the traditional PCBs. Other connections like RS232 ports can only be connected to one device at a time. USB allows multiple devices to be attached to a single port. This paper aims to develop a controller FPGA based USB device core. The USB controller is responsible for the correct transmission and reception of data through USB interface. It consists of a transmitter and a receiver. USB controller has error checking features built into it. The USB controller is capable of carrying out low speed (1.5 Mbps) USB transactions.

II. SYSTEM ARCHITECTURE

The basic idea of this proposed project is to implement hardware of a USB2.0 controller based on FPGA. A typical USB 2.0 controller contains a USB transmitter protocol layer, transmitter physical layer, receiver physical layer and receiver protocol layer.

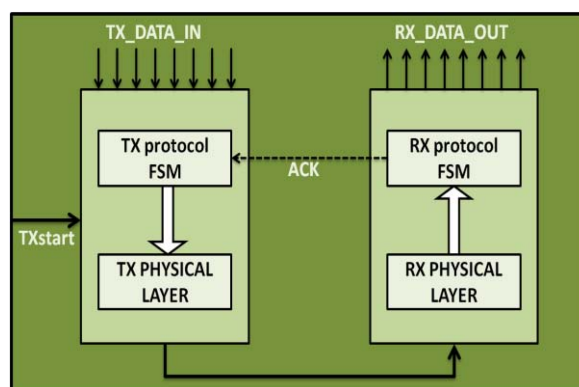


Fig. 1 General Architecture

TRANSMITTER PROTOCOL LAYER

Transmitter protocol layer is implemented by using FSM (finite state machine). It is basically used by physical layer in order to generate data packets.

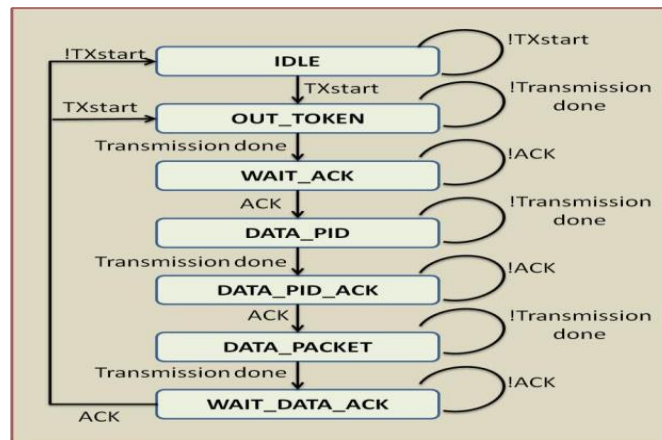


Fig. 2 Transmitter Protocol layer

The default state of fsm is ideal state and continuously waits for the TXstart to be asserted by the external input. Whenever it gets TXstart signal the state of FSM changes to OUT_TOKEN and in this state protocol layer transmits the OUT_TOKEN packet to the transmit physical layer for the further processing and remain in this state until the transmission of OUT_TOKEN is not completed by the transmitter physical layer. After the transmission of the OUT_TOKEN packet by the transmitter physical layer generates transmission done acknowledgement to the transmitter protocol layer after getting transmission done acknowledgement from the transmitter physical layer, the state of FSM changes to WAIT_ACK and remain in this state until it will get acknowledgement from the receiver protocol layer. Now, the state of FSM changes to DATA_PID. It will remain in the DATA_PID until it gets transmission done from transmitter physical layer otherwise FSM changes to DATA_PID_ACK and if gets acknowledgement from receiver protocol layer then it will go to DATA_PACKET state, now valid data is ready to transfer. The WAIT_DATA_ACK signal, after getting acknowledgement it checks for TXstart from the input. If TXstart pin is asserted then next state will be OUT_TOKEN state else next state will be IDLE.

The signal transmission done is getting from transmitter physical layer and we will get this signal when 8bit packet is completely transmitted.

The signal ACK is getting from receiver protocol layer and we will get this one when receiver ensures that it is the valid data.

TRANSMITTER PHYSICAL LAYER

Transmitter physical layer includes shift register, bit stuffer and NRZI encoder.

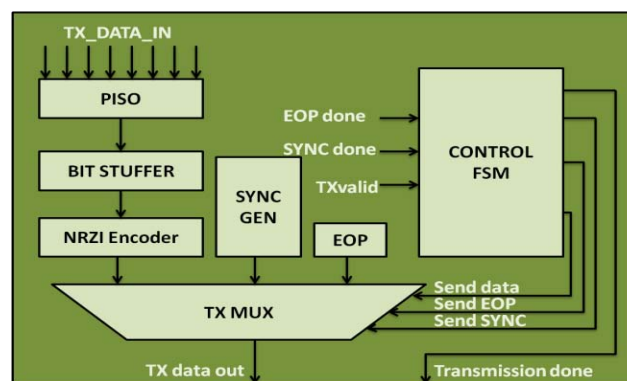


Fig. 3 Transmitter Physical Layer

NRZI Encoder

The USB devices employ NRZI data encoding for transmission of data packet. In NRZI encoding, logic 1 is represented by change in level and logic 0 is represented by a no change in level.

Bit Stuffer

It is employed by the transmitting device when sending a packet on USB. A zero is inserted after every six consecutive ones in the data stream before the data is NRZI encoded, to force a transition in the NRZI data stream. This bit stuffer adds the stuff bit if six consecutive ones are found.

PISO

It is a serial shift register. It converts the parallel data to serial data stream.

SYNC generator module

Sync pattern which is transmitted before the transmission of any data packet.

EOP generator module

Generates EOP (End of Packet) pattern which is transmitted after the transmission of any data packet.

TXMUX module

This block ensures which data pattern is to transmit by the transmitter.

Controller FSM module

This module generates the entire necessary signal required by the physical layer module to transmit valid data.

RECEIVER PHYSICAL LAYER

Transmitter physical layer includes shift register, bit unstuffers and NRZI decoder.

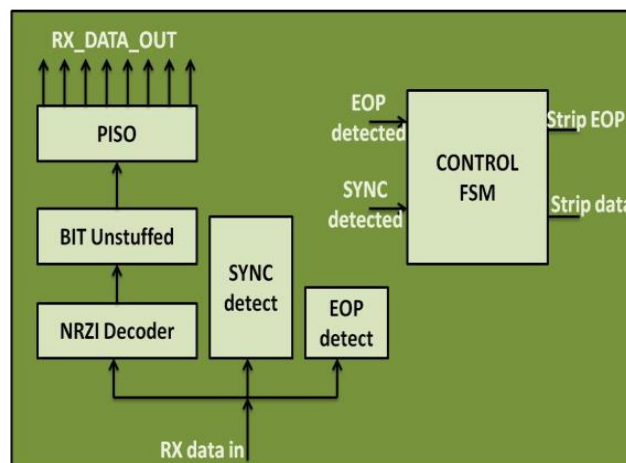


Fig. 4 Receiver Physical Layer

NRZI Decoder

This module decodes the NRZI encoded data. To do so xor-ing of current input and previous input is done.

Bit Unstuffer

After NRZI decoding, the bits that are stuffed in the transmitter side has been unstuffed to get the valid data at the receiver end.

SIPO

It is a serial shift register. It converts the serial data to parallel data stream.

SYNC detector module

Detects the Sync pattern transmitted by the transmitter.

EOP detector module

Detects the EOP pattern transmitted by the transmitter at the end of data packet.

Controller FSM module

This module generates the entire necessary signal required by the physical layer module.

RECEIVER PROTOCOL LAYER

Default state of receiver protocol FSM is wait for OUT_TOKEN. In this state the receiver protocol FSM is continuously waiting for token packet transmitted by the transmitter. After getting valid OUT_TOKEN, the state of FSM changes to wait for EOP, in this state FSM continuously checking for the EOP acknowledgment from

receiver physical layer .When we get EOP,it will wait for DATA_PID and wait for the EOP. When getting EOP, now receiver will get valid data and send ACK signal to transmitter protocol layer.

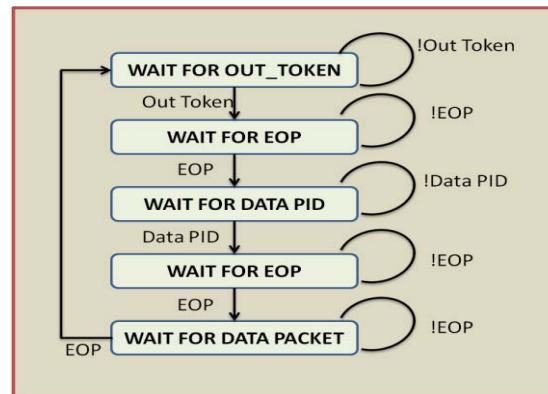


Fig. 5 Receiver Protocol Layer

III. TOP MODULE

This controller is designed to transmit the 8 bit binary data serially using 2.0 protocol. In this both protocol layer and physical layer of both transmitter and receiver are designed using Verilog HDL.

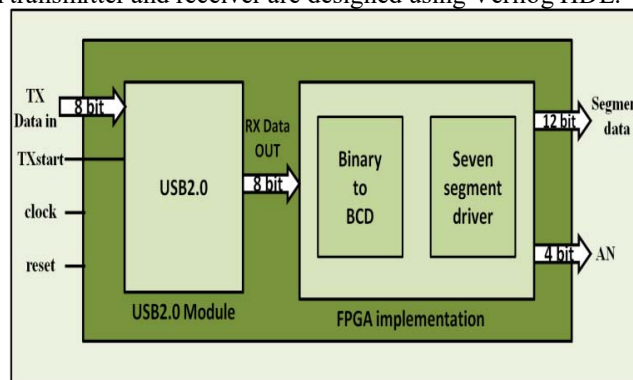


Fig. 6 Implementation of top module

Protocol layer of both receiver and transmitter is implemented by using FSM the physical of transmitter includes NRZI Encoder and Bit Stuffer. All these functionality is written in Verilog and implemented on FPGA Basys 2 Board. The 8-bit value to be transmitted is given from the 8 up down switches of FPGA kit and the output is displayed on the seven segment display. To display 8-bit value on seven segment,8-bit binary value is first converted into BCD value for which an algorithm known as Double Dabble algorithm is implemented .The BCD value so obtained is then decoded into seven segment form.

On the single FPGA we have implemented both transmitter and receiver, here input is given by 8 up down switches of FPGA and output is displayed on seven segment display. As we know, FPGA basys2 board is only having 8 up down switches, so it only display data up to 255.

IV. SIMULATION RESULTS

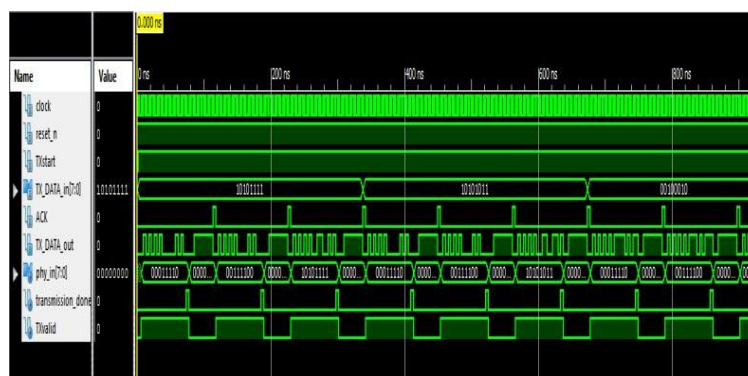


Fig. 7 Transmitter Section

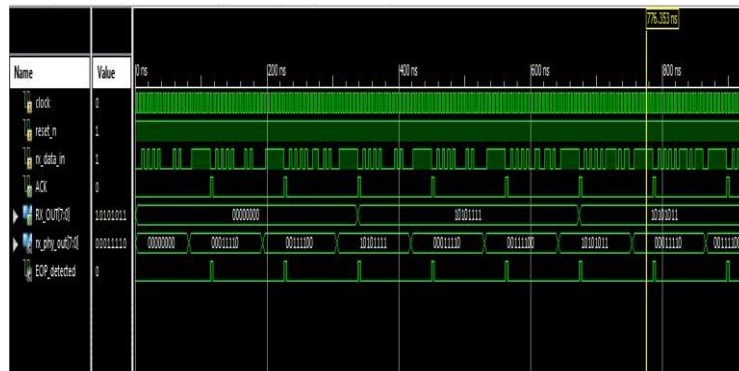


Fig. 8 Receiver Section



Fig. 9 Top Module

IMPLEMENTATION OF USB2.0 ON FPGA

STATUS:

TX_START=1 (HIGH), that means data is ready to transmit. At the input pins i.e. switches we have entered 11011111 (binary bits) same will be displayed on the seven segments in the decimal form (223) that we have taken as the output.



Fig.10 Implementation on FPGA basys2 board

V. CONCLUSIONS

This paper is an approach for the implementation of the USB 2.0 controller on FPGA basys2 board. To conclude the proposed USB 2.0 controller has a number of important advantages one of them is small space as it

requires only one FPGA board ,it has both transmitter and receiver section. The 8 up down switches used as a transmitter section and seven segment display as a receiver section. We use the Double Dabble Algorithm to convert the 8-bit binary value into BCD for seven segment display.

VI. REFERENCES

- [1] .Przemyslaw M. Szczółka, Kamil J. Pyrzynski, "USB receiver/transmitter for FPGA implementation",International Conference on Signals and Electronic Systems (ICSES), pp.1-6, Sept. 2012 [2].A.abba,F.Caponio, A.Cusimano ,A.Geraci, "Controller IP for a Low Cost FPGA Based USB Device Core", IEEE Nuclear Science Symposium and Medical Imaging Conference , pp.1-4,Oct. 2013
- [3]. Elio A. A. De Maria, Edgardo Gho, Carlos E. Maidana, Fernando I Szklanny, Hugo R. Tantignone, "A LOW COST FPGA BASED USB DEVICE CORE",4thSouthern Conference on Programmable Logic, pp.149-154, march 2008
- [4]. Fatemeh Arbab Jolfaei, Neda Mohammadizadeh, Mohammad Sadegh Sadri, Fatemeh FaniSani, "High Speed USB 2.0 Interface for FPGA Based Embedded Systems",Fourth International Conference on Embedded and Multimedia Computing, pp.1-6, Dec. 2012
- [5] .Guangling Guo, Zhiqiang Li, Fan Yang, "TRANSMISSION AND RECEPTION OF DATA THROUGH USB USING FPGA" International Conference on Multimedia Technology, pp.5374-5376, july 2011 [6].Universal Serial Bus Specifications Revision 2.0.
- [7] "Xilinx Basys2 FPGA kit, Complete Data Sheet", Xilinx Corp., Aug 2005.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

DIMENSIONALITY REDUCTION IN BIG DATA: A SURVEY

Suhani¹ and Nidhi Jain²

Computer Science

Centre for Development of Advanced Computing,
Sec-62, Noida-201309,
India

Abstract: The term big data is generating big buzz all over the world in recent years. Many techniques and software are proposed to handle big data. The features of big data (i.e., variety, volume, velocity, value, veracity) add complexity to the data which makes traditional data mining techniques invalid. Dimensionality reduction is recommended to handle high dimensional data prior to other tasks. So, it becomes very important to perform dimension reduction efficiently. The curse of dimensionality is the main reason for considering high-dimensional problem.

In this paper, various techniques are presented to lessen the dimensions of data.

Keywords: Big Data, Dimension reduction, Hadoop, dimensions

I. INTRODUCTION

A **Dimension** is a construction that classifies facts and measures in order to allow users to answer the business questions. Dimensions provide organized classification information to else unordered numeric events in a data warehouse. Also, a dimension is a data set which is composed of individual data elements which are non-overlapping.

The main function of dimensions is threefold: to provide filtering, labelling and grouping.

Dimension Reduction refers to the process of translating a dataset with massive dimensions into a data with reduced dimensions ensuring that alike information is tersely conveyed. While resolving **machine learning problems** to attain improved features for a classification or regression task, these techniques are typically used.

There are several **benefits** of applying Dimension Reduction process, some are as follows:

- It benefits in data compression and storing space reduction.
- It decreases the time obligatory for executing the same calculations. Fewer dimensions leads to less computation, also fewer dimensions can permit usage of algorithms that are unfit for higher dimensions.
- It improves the model performance by taking care of multi-co linearity. It removes the redundant features. For example: a value kept in two different units (meters & inches) is redundant information. Decreasing the dimensions of data to 2D or 3D makes it stress-free to scheme and picturize. Then the patterns can be observed clearly.
- It is also helpful in removing the noise and as a result, the performance of models can be improved.

Why Dimensionality Reduction?

- It is very convenient and easy to collect data.
- Data accrues in an unknown rapidity.
- Data pre-processing is a significant part for operative data mining and machine learning.
- Data is not composed for only data mining.
- Dimensionality reduction is an effective approach for reducing the size of data.
- Most data mining and machine learning methods may not be effective for high-dimensional data because of:
 - Curse of Dimensionality
 - As the dimension rises, query accurateness and effectiveness degrades quickly.
- The intrinsic dimension may be small.

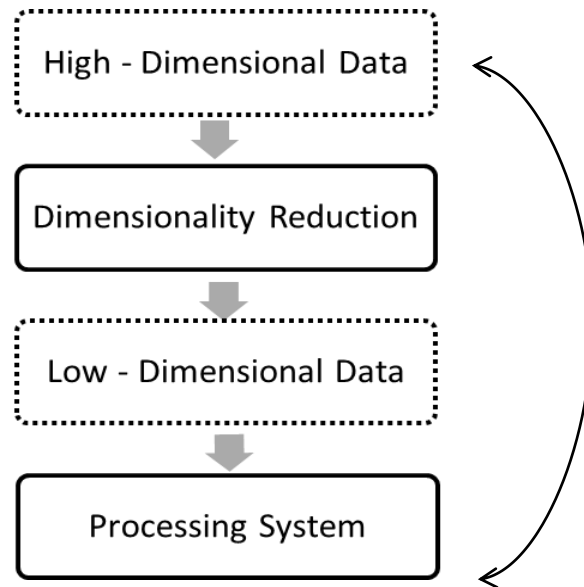


Fig.1 shows the basic architecture of Dimension Reduction

About Big Data

Big Data refers to huge amount of data which includes complex structured, semi-structured, and unstructured data that has been too complex in the past to be accurately and efficiently interpreted. Though it is very difficult to put a number on what quantifies the big data, big data does not refer to not only Exabyte's or petabytes of data. The data is said to be big data when the amount of data that is needed to be processed is greater than what a system could actually process. There are no strict rules about the size of database that requests to be in sequence for the data in it to be deliberated as "big." Typically big data is defined as the necessity for more and more new tools & methods in order to be able to process it.

Characteristics of Big Data (The 'V's)

Initially there were 3 vs,

1) Volume – It is the measure of the amount of produced and stored data. The magnitude of the data firmly decides the value and latent insight and also whether it can actually be deliberated as big data or not. About approximately 2.5 exabytes of data is generated every day. Data Volume can be measured by quality of events, dealings and the quantity of history. Big data is not just the spoken or written account of raw volume. The real challenge is to identify or develop the most reliable cost-operative approaches for mining value from all the terabytes and petabytes of data available now. This is where Big Data analytics become essential.

2) Variety - It is the medley of data. Data can be: structured, semi-structured and unstructured. Usually, data (exclusively operational data) is structured because it is placed into a database based on its kind i.e., numeric, character, floating point, etc.

Wide variety of data: Internet data (Social Network- Instagram, Twitter, Facebook, Social media), Primary Research (Surveys, Experiences, Observations), Location data (Mobile device data, Geospatial data), Secondary Research (Competitive and Market place data, Consumer & business data, Industry reports), Supply Chain data (vendor Catalogues, Pricing, etc.), Image data (Video, Satellite image, Surveillance), Device data (Sensor data, RF device, Telemetry) etc.

Types of data:

Structured Data: They have predetermined data models and can be fitted into relational database. Especially, the operative data is organized because it is stored based on the kind into a database (i.e., numeric, character, floating point, etc.).

Semi-structured data: These are the type of data that do not have predefined formal structure of data models. Semi-structured data is often a mixture of various types of data that has nearly structure or pattern that is not as firmly distinct as structured data. To separate the semantic rudiments (that contains the ability to impose hierarchies in the data), semi-structured data contains tags.

Unstructured data: is not appropriate as a relational database and /or do not have a pre-defined data model. Many a times, image, text, audio, video, Internet (including log files and click streams), and geospatial data are well-thought-out as unstructured data.

3) Velocity- Data velocity is about the swiftness at which data is generated, accrued, consumed, and treated. As the speed of the world is increasing, the loads on businesses to process information in real-time or with near

real-time responses is also increasing. This may mean that data is treated on the go, to make fast, real-time decisions or for more timely decisions, monthly group processes are run inter-day.

Other 2 were added (in Hadoop 2.0) to emphasize need for data authenticity and value.

4) Veracity -The worth of seized data can highly vary, disturbing the precise analysis. It is related to privacy, security & accountability, which creates a need to verify secure data.

5) Value - Analysis of this data can yield counter intuitive insights and actionable intelligence through predictive models that handle the what-if queries.

Challenges with Big Data

- **Data redundancy** is the beingness of additional data to the definite data which does not make any difference to the meaning of data. It permits the error correction in stored or communicated data. This supplementary data can simply be a whole copy of the real data, or only the designated pieces of data that permit error detection and reconstruction of injured or misplaced data up to a certain level.
- **Data integration** involves joining data from different sources and also provides users with a integrated view of these data.
- **Data creation**, 2.5 Exabyte's every day. Sources: Transactions, log data, social media, events, emails etc.
- **Data curation** includes all the processes required for controlled and principled data creation, management, and maintenance. These processes also have capacity of adding value to data.

II. MAJOR TECHNIQUES OF DIMENSIONALITY REDUCTION

1) Feature Selection is a process that according to an objective function chooses an optimum features subset.

Goals of feature selection-

- To remove noise from data and to reduce dimensions of data.
- To improve performance of mining
 - Learning speed
 - Accuracy of prediction.
 - Comprehensibility and simplicity of excavated results.

2) Feature Extraction

- Feature reduction is defined as the plotting of the high-dimensional data (original data) against a space with lower-dimensional.
- On the basis of different problem settings, the criterion for feature reduction can be different that are as follows:
 - Supervised setting: maximizes the discrimination of class.
 - Unsupervised setting: minimizes the loss of information.

III. MAJOR CHALLENGES WITH HIGH DIMENSIONALITY

- **Computational issues** - High dimensional data leads to computational issues as many algorithms do not scale linearly but exponentially.
- **Space complexity** – this is another major issue as the dimensions of the data increases, it requires more storage and thus leading to space complexities.
- **High computational cost** – with high dimensionality of data arises unavoidable computational issues which brings along the increasing cost of computation.
- **Curse of dimensionality** – This is indeed the major challenge with high dimensional data. It states to the fact that, the size of trial data required to determine a purpose of numerous variables to an accuracy degree assumed (i.e., to get an estimate of rationally low-variance) raises exponentially with the quantity of variables. All this is in the absence of simplifying assumptions.

Dimensionality reduction addresses these problems, while (hopefully) preserving most of the information that is relevant in the data needed to learn predictive and accurate models.

IV. APPROACHES

1) Principal Component Analysis (PCA)

PCA is a linear dimension reduction method. It reduces the dimensions by transforming the original dataset into new variable set known as Principal components.[1][2].

Tonglin Zhang *et al.* (2016)[3] proposes a new approach as because of the barriers of memory and storage, the traditional PCA cannot be applied to big data. The basic idea of this article was to scan data by rows in order to develop an array of tolerable measurements. It was shown that the approach proposed here can deliver precise solutions if in the follow up analysis, the linear regression approach is used. Bouzalmat Anissa *et. al.* (2015) [4] shows the comparisons of linear and non-linear methods on the basis of their efficiencies. Also, author has shown the effect of great dimensionality of features using PCA, ICA, LDA, and Sparse Random Projections. Young Kyung Lee *et. al.* (2012)[5] proposes a modification of the standard PCA that works for such high-dimensional data when the loadings of principal components are sparse. Their method starts with an initial subset selection, and then performs a penalized PCA based on the selected subset. They have shown that their procedure identifies correctly the sparsity of the loading vectors and enjoys the oracle property, meaning that the resulting estimators of the loading vectors have the same first-order asymptotic properties as the oracle estimators that use knowledge of the indices of the nonzero loadings. Their theory covers a variety of penalty schemes. They also provide some numerical evidence of the proposed method, and illustrate it through gene expression data.

2) Locally Linear Embedding (LLE)

LLE is a nonlinear dimensionality reduction technique that calculates neighbourhood preservative embedding of high-dimensional inputs into low-dimensional.[6][12]

S. T. Roweis *et. al.* [6] introduces LLE as an unsupervised learning algorithm that unlike clustering algorithms, plots its inputs into a solitary universal coordinate system of low dimensions. Lawrence K. Saul *et. al.*[7] describes locally linear embedding (LLE), an unsupervised learning algorithm that calculates neighbourhood preserving embedding of high dimensional data into low dimensional data. LLE exploits local symmetries of linear reconstruction to determine nonlinear structure in data of high dimensions. They illustrated the technique on images of lips used in audio-visual speech synthesis. Jing Chen *et. al.* (2011) [8] comprehensively reviewed and discussed the existing extensions of LLE along with their advantages and disadvantages. Also they suggested several directions for future research to generalize different tactics in various extensions that are related to LLE stages and evaluating their performances.

3) ISOMAP

Joshua B. Tenenbaum *et. al.* (2000) [9] describes a framework i.e. ISOMAP and discusses how this framework overcomes the disadvantages of the linear methods: PCA & MDS.

Conclusions:

- Isomap handles non-linear manifold.
- Isomap keeps the advantages of PCA and MDS & efficiently (non-iterative, polynomial time) computes a globally optimal solution.
- Within a single coordinate system, it denotes the universal structure of a data set.

Minkook Cho *et. al.*(2009)[10] proposed a method to measure data pair distance that utilizes class-membership so as to discover a low-dimensional manifold. The distance between data points and the distance between classes is also preserved. It was confirmed that the anticipated method provides improved performance than the orthodox ISOMAP by carrying out computational experimentations on real facial data sets and on artificial data sets. M. Balasubramanian *et. al.*(2002) claims that in the context of levelling cortical surfaces, “the basic idea” of their Isomap technique for non-linear dimensions reduction has “long been known”. However, their problem of finding low dimensional structure differs the problem of cortical levelling in critical ways, in a cloud of high dimensional data points. State-of-the-art ways for cortical levelling, takes input as a triangulated mesh of fixed dimensionality & topology, which signifies an supplementary structure which is unavailable in the broad-spectrum problem they tried to solve. They took as input only a collection of unorganized data points and dimensionality of the underlying manifold are unknown which was estimated in the construction of a faithful low-dimensional embedding process. Their algorithm estimates the inherent geometry of a data diversity which was based on an irregular approximation of each data point’s neighbours on the manifold. The straightforwardness of the algorithm, in contrast to the special methods used in cortical levelling, makes it very effective and usually appropriate to a wide range of data sources and dimensions. Based on their failure (that was corrupted by additive Gaussian noise) to create a topology-stabilising 2D implanting of a Swiss roll data set, they also assert that Isomap is “topologically unstable”.

4) Random Projections

Random Projections takes a high-dimensional data-set and then maps it into a lower-dimensional space, also it provides some guarantees on the estimated maintenance of the distance. This was done as follows, suppose input data is $n \times d$ matrix - A . Then, for doing projection, they have chosen a “suitable” $d \times k$ matrix R , and then defined the projection of A to be - $E = AR$, which stores k -dimensional approximations for the “ n ” points. [13]

Jing Wang *et al.* (2015) [14] realized Random Projection in dimension reduction & improves the limitations of data with high dimensions. Effectiveness of RP was compared with PCA in terms of Recognition rate of recognition of face. The **results** of the experiment performed uncovers that in the great dimension space, the enactment of RP is better than PCA (For high dimensional data, dimensions > 100, the recognition rate of RP > recognition rate of PCA). Bouzalmat Anissa *et al.* (2015) [4] shows the comparisons of linear and non-linear methods on the basis of their efficiencies. Also, author has shown the effect of features with high dimensionality using PCA, LDA, ICA and Sparse Random Projections.

Conclusions:

- RP is an optimal method, and is capable of functioning very fast even if the amount of dimensions is very high.
- Random projections have farsuperior runtime than PCA, LDA & ICA.

Ella Bingham *et al.* (2001) [15] have shown that when data is projected onto a random low-dimensional subspace produces outcomes which are as good as to orthodox dimensionality reduction methods such as PCA: the similarities of data vectors is well conserved under random projections. However, using random projections is significantly computationally less costly than using other methods. They have also shown experimentally, that for additional savings on computations, using a sparse random matrix is a good idea in random projection.

5) Holistic Approach [16]

In this approach, 3 fundamental problems of big data are addressed that are nearly linked to the scattered dimensionality reduction of big data:

- Big data fusion
- Dimensionality reduction algorithm
- Construction of distributed computing platform.

A chunk tensor method was offered to integrate the structured, semi-structured and unstructured data as a fused model in which all the features of the dissimilar data are arranged applicably along the tensor orders. A Lanczos based High Order Singular Value Decomposition (HOSVD) algorithm was suggested to lessen dimensions of the unified model. In terms of convergence property, storage scheme, computation cost, & theoretical analyses of the algorithm were also given. To perform the dimension reduction task, this paper employed a Transparent Computing paradigm for constructing a scattered computing platform as well as it exploits the linear predictive model for partitioning the blocks of data. The anticipated holistic approach is effectual for distributed dimensionality reduction of big data as demonstrated by the experimental results

V. CONCLUSION

In this paper, various techniques are presented to lessen the dimensions of novel data. From the survey, it comes to know that non-linear dimensionality reduction approaches are advantageous over linear methods as the data now days are highly non-linear. But, PCA can be used as a pre-processing phase for any method of dimensionality reduction. Also, if the dimensions are more than 100 then Random Projections is very efficient to use. Lastly, the Holistic Approach described is very effectual for scattered dimensionality reduction of Big Data.

REFERENCES

- [1] J.E. Jackson. "A User's Guide to Principal Components", New York: John Wiley and Sons, 1991.
- [2] I.T. Jolliffe. "Principal Component Analysis". Springer-Verlag, 1986.
- [3] Tonglin Zhang and Baijian Yang, "Big Data Dimension Reduction using PCA", 2016 IEEE International Conference on Smart Cloud.
- [4] Bouzalmat Anissa, Belghini Naouar, Zarghili Aarsalane, Kharroubi Jamal, Laboratory of Intelligent System and Applications, Faculty of Science and Technology, Fes, Morocco, "Face Recognition: Comparative study between linear and non linear dimensionality reduction methods", 978-1-4799-7479-5/15/\$31.00 ©2015 IEEE.
- [5] Y.K. Lee, E.R. Lee, and B.U. Park. "Principal component analysis in very high-dimensional spaces," *Statistica Sinica*, 22, 933-956, 2012.
- [6] S. T. Roweis and L. K. Saul. "Nonlinear dimensionality reduction by locally linear embedding" *Science* **290**, 2323-2326 (2000).
- [7] Lawrence K. Saul, Sam T. Roweis, "An Introduction to Locally Linear Embedding", *SCIENCE*, 2000.
- [8] JING CHEN, ZHENGMING MA, "LOCALLY LINEAR EMBEDDING: A REVIEW", *International Journal of Pattern Recognition and Artificial Intelligence* Vol. 25, No. 7 (2011) 985_1008.
- [9] Joshua B. Tenenbaum (Stanford), Vin de Silva (Stanford), John C. Langford (CMU), "A Global Geometric Framework for Nonlinear Dimensionality Reduction", *SCIENCE*, 22 Dec 2000, Vol. 290, Issue 5500, pp. 2319-2323.

- [10] Minkook Cho, Hyeyoung Park, "Nonlinear dimension reduction using ISOMap based on class information", International Joint Conference on Neural Networks. 2009, IEEE.
- [11] M. Balasubramanian, E. L. Schwartz, J. B. Tenenbaum, V. de Silva, and J. C. Langford, "The Isomap Algorithm and Topological Stability" *Science*, 295(5552), 2002.
- [12] L. Donoho and C. Grimes. "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data". *Proc. Natl. Acad. Sci. U.S.A.*, 100(10):5591-5596, 2004.
- [13] Aditya Krishna Menon, "Random projections and applications to dimensionality reduction", University of Sydney, 2007.
- [14] Jing Wang, Dept. of Information Management, The Central Institute for Correctional Police, "Efficient Face Recognition Research Based on Random Projection Dimension Reduction", 978-1-4673-7143-8/15 \$31.00 © 2015.
- [15] Ella Bingham, Heikki Mannila, "Random projection in dimensionality reduction: applications to image and text data", in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245-250, 2001.
- [16] Liwei Kuang, Yaoyue Zhang, Laurence T. Yang, Jinjun Chen, Fei Hao, and Changqing Luo, "A Holistic Approach to Distributed Dimensionality Reduction of Big Data", *IEEE Transactions on Cloud Computing*, 2015.
- [17] Nebu Varghese, Vinay Verghese, Prof. Gayathri .P and Dr. N. Jaisankar, "A Survey of Dimensionality Reduction And Classification Methods", *International Journal of Computer Science & Engineering Survey (IJCSES)* Vol.3, No.3, June 2012 .

ACKNOWLEDGMENTS

I take this opportunity to acknowledge all those who have guided me in this project work. I express my earnest gratitude towards Ms. Nidhi Jain, my research guide for her valuable encouragement and guidance.

I would like to thank faculty members of C-DAC, NOIDA for their constant support and guidance during my project work. Their motivation and suggestions were invaluable in the project work. I am grateful to them for their most cooperative attitude and suggestions, without which I could not have been able to do this work.

Last I wish to thank my mother, father and my sister who were always there for me by giving everything they have, their love and support.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Water Quality Monitoring using Data Mining Techniques

Mitali Kathpal¹, Kriti Saroha²

School of Information Technology,

Centre for Development of Advanced Computing, Noida, India

Abstract: This paper presents a case study that applies statistical techniques to river quality monitoring database to discover the patterns of water quality deterioration in river water. The declining quality of water can be handled properly if the quality of water is predicted beforehand. In this some techniques have been introduced to examine the future quality of water. A dataset that includes a total of 8451 samples of water quality collected with 1 hour time interval from different sites was used to examine and analyze the water quality. This paper includes the values of 15 parameters which adversely affect the quality of water.

Keywords: ANN with NAR, MSE, RMSE, regression, WQI

I. Introduction

Water quality is becoming a major issue now-a-days as the quality of water is degraded due to various factors which, includes industrial, commercial activities, and poor sanitation infrastructure. The water contamination is adversely affecting health, environment and infrastructure. Therefore, it is very important to introduce approaches for predicting future water quality trends. For carrying out effective analysis and prediction of water quality pattern trends, it is important to involve time dimension as one of the attributes to carry out analysis, so that the quality of water according to temporal variation is addressed. The algorithm for analysis and prediction which have been used before may include techniques from the area of artificial intelligence and others like Bayesian networks (BN), artificial neural network (ANN), support vector regression (SVR), decision support system (DSS), neuro-fuzzy inference and Auto-Regressive Moving Average (ARMA). However, the non-linear data of water quality makes it difficult to predict quality of water.

The main aim of the study is to introduce a method that would analyze and predict the water quality beforehand using water quality parameters. The 15 parameters used are biological, physical and chemical, which affect the quality of water.

This paper proposes a solution by introducing a model based on Machine Learning Techniques for predicting the future water quality trends by analyzing historic water quality dataset. ANN with NAR is applied to make effective prediction of water quality trends.

The paper is organized as follows; section 2 presents the description of related work in the area. Section 3 describes the proposed approach. Section 4 summarizes the conclusion and results.

II. Literature Survey

Various studies have been done to extract knowledge that can examine the water contamination.

This section briefly describes and compares the techniques used by different researchers with their conclusions. Some of them are described in this section.

Yafra khan; Chai Soo See collected data from U.S Geological Survey's(USGS) National Water Information System(NWIS) in U.S. Data collected with 6 minute time interval for efficient prediction of water quality. Authors presented a novel method for predicting quality of water using ANN with Non-Autoregressive (NAR) time series model. The performance of the model was depicted through Regression, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). This model predicted more accurate results with lowest MSE for Ph 3.7×10^{-4} and the best regression value 0.98 for salinity.

Ting nien wu, et.al. collected data from Eighty-four monitoring wells in Taiwan. Author presented a method for knowledge discovery which included various steps as shown in fig 1:-

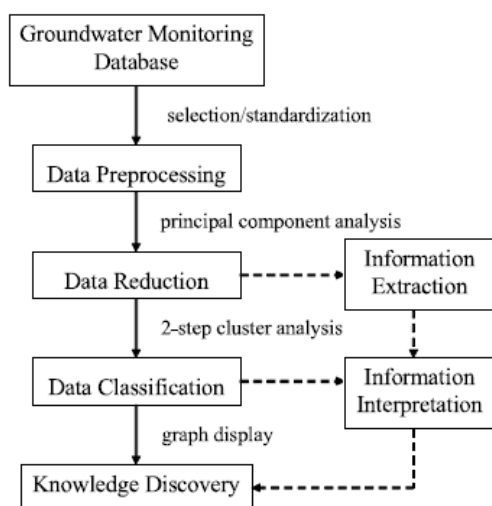


Fig 1

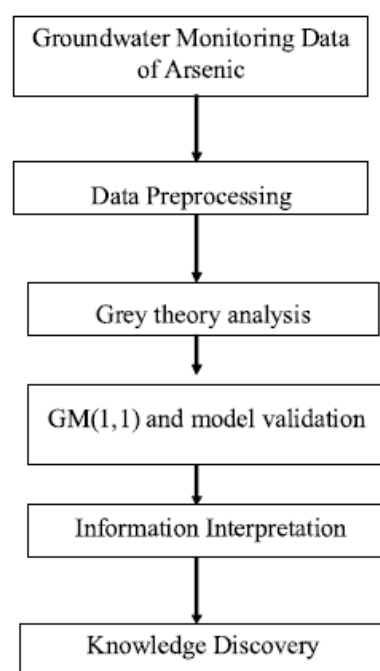


Fig 2

The research identified four pattern trends of groundwater deterioration as arsenic dissolution, organic pollution, salinization and mineralisation.

Jan-Yee-Lee collected dataset from 84 monitoring wells in chinanian Blackfoot disease region from 2009 to 2012. The collected dataset included concentration of arsenic in groundwater which was taken as input for model. Author proposed a model known as GREY MODEL for prediction based on single variable. The method proposed by author is shown in fig 2.

Among the 84 monitoring wells in chianian plain, 16 wells have high arsenic level. The prediction model based on grey theory have improved the accuracy of limited data prediction and forecasting. This model helped in improvement of groundwater monitoring.

Ivana.D.Radojevic, et.al. Collected dataset for Gruza reservoir from five sampling sites which includes 968 samples for analysis and dataset for Grosnica reservoir was collected from three sampling sites which include 172 samples that were considered for analysis. The author used clustering using k-means and classification with decision trees for prediction of state of coliforms and water quality monitoring. This approach proposed by author proves to be reliable one in predicting the total coliforms in the dataset.

Maqbool Ali, et.al. Collected water samples from twelve distinct locations of Rawal watershed. A dataset consists of 663 samples from different locations was collected for years 2009 to 2012. For analyzing the water quality, it includes physical and chemical parameters and also bacteriological parameters. The author proposed methodology which is divided into five parts. In the first part, the quality parameters trends were presented according to month. In second part, parametric satisfactory analysis is given. In third part, the data is pre-processes and outliers are removed. In fourth part best quality index is found by means of various clustering methods. In the final phase, the months with high contamination of fecal coliforms were found. The results analysed in this paper were as follows:- Using hierarchical clustering with ALM applying Euclidean distance to determine water quality index have given better results as compared to any other technique. Also, MLP has given better results for classification.

Neetu Arora, et.al. collected dataset for Satluj river for duration of 7 years. The author presented a approach which is useful in understanding the main pollutants in water quality deterioration. In this case study proposed by author, cluster analysis was used to study the temporal and spatial variation in surface water quality of Satluj river. It helps in reducing number of monitoring stations and understanding the main factors which influence water quality.

Comparison Table

	Yafra khan; Chai Soo See	Ting nien wu, et.al.	Jan-Yee-Lee	Ivana.D.Radojevi c, et.al.	Maqbool Ali, et.al.	Neetu Arora, et.al.
Algorithm	Artificial neural networks	PCA Cluster Analysis Using K-MEANS	Grey model	Clustering and classification using decision trees	Classification techniques MLP,RBF,KNN,SVM Clustering techniques k-means Hierarchial	Hierarchical clustering
Attributes	4 attributes	14 attributes	Only arsenic	Only coliforms	12 Attributes	8 monitoring stations
Conclusion	This model proves reliable with high prediction accuracy. High regression value for specific conductance being 0.98	The clustering divides the dataset as follows:- 35% for salinization, 11 % for arsenic dissolution, 19% for organic pollution, and 35% for mineralization	Accuracy level is good with error<0.01	Better prediction of coliforms in water quality to maintain ecological balance.	High accuracy with prediction fecal coliforms were high in july, march, june and October.	Better understanding of parameters that are most important for water quality.
Future Work	User centric approach				Time Series forecasting model	Principal component analysis

III. Proposed Approach

This section presents the proposed approach that would be used to mine water quality dataset. The approach includes applying ANN with NAR model on collected dataset containing 15 water quality parameters.

As a first step the class labels are defined as good, average, excellent and bad using water quality index calculation.

WQI is computed that provides a single numerical value to express the overall water quality based on different water quality parameters. It also defines the rating of quality whether it is good or bad based on water quality parameters.

Water Quality Index is calculated from the following given equation. The Unit weight (W_i) has been computed by using this formula:

$W_i = k/S_i$ Where k = Proportionality constant

$K = 1/(\sum 1/S_i)$

S_i = Standard permissible value of i th parameter given by WHO standard.

W_i = unit weight of i th parameter

Water quality index (Weighted Arithmetic Mean method)

$WQI = \sum q_i w_i / \sum w_i$

Where q_i = Sub Index of the quality rating for i th parameter

$Q_i = 100 * (V_i / S_i)$

V_i = observed value in laboratory; n = number of parameters taken S_i = standard value of i th parameter.

In the next step, the water quality trends of 15 water quality parameters are predicted using ANN with NAR model. The dataset used for analysis is with 1 hour time interval.

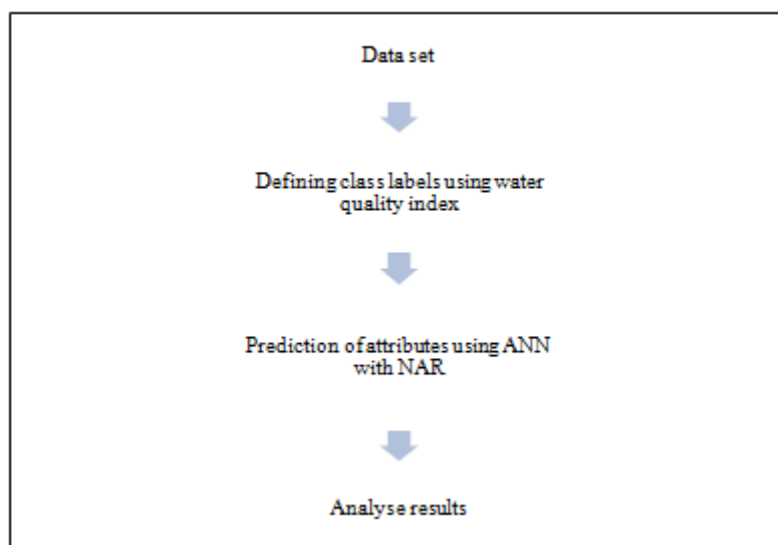


Fig 3

IV. Results and conclusion

Implementation results are shown using ANN with NAR model. The performance of this model is measured through Regression value and mean squared error. The proposed ANN with NAR time series model proves to be the efficient one with the accuracy of prediction with lowest MSE being 0.0006 for phosphate and the best regression value for pH 0.87492.

The plot for regression analysis displays that how strongly the data fitted the function for training, validation and testing. The more the value of regression is close to 1, the better is the accuracy of prediction. Here, the regression plot is shown for PH and Phosphate in fig b and fig f.

The MSE graph shows the number of epochs taken to converge testing, validation and training. Here, the MSE graph is shown for PH and Phosphate in fig c and fig e. In fig a and fig d, The blue dots in the graph shows the target values and the red dots shows the predicted values by this proposed model.

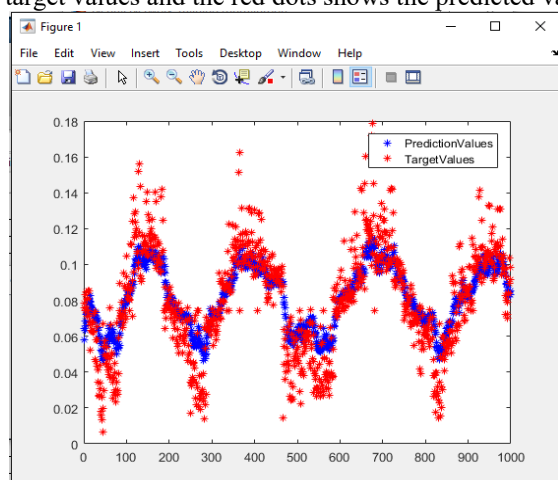


Fig a: Prediction of Ph using ANN with NAR

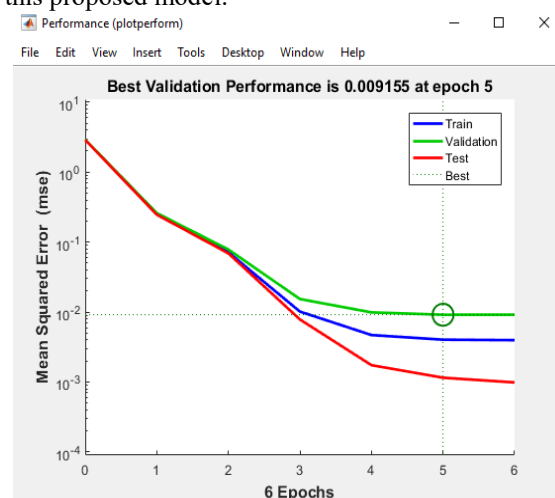


Fig b: Mean Squared Error graph

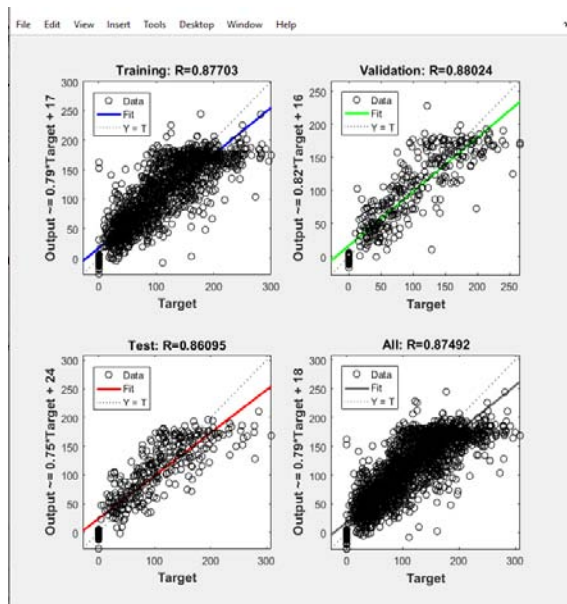


Fig c: Regression Analysis

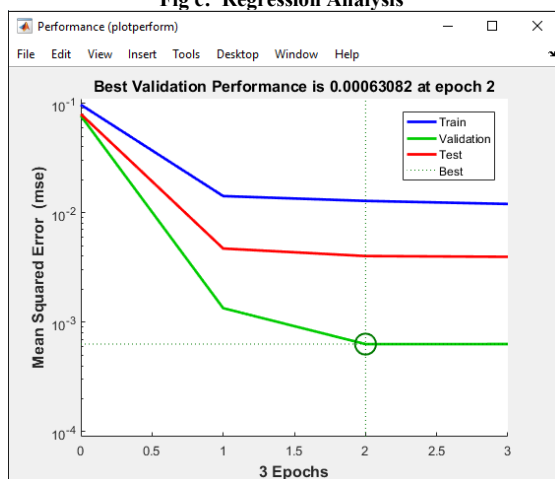


Fig e: Mean Squared Error

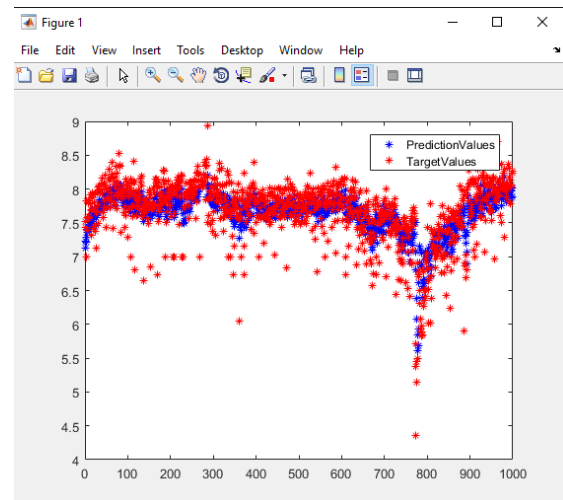


Fig d: Prediction of phosphate using ANN with NAR

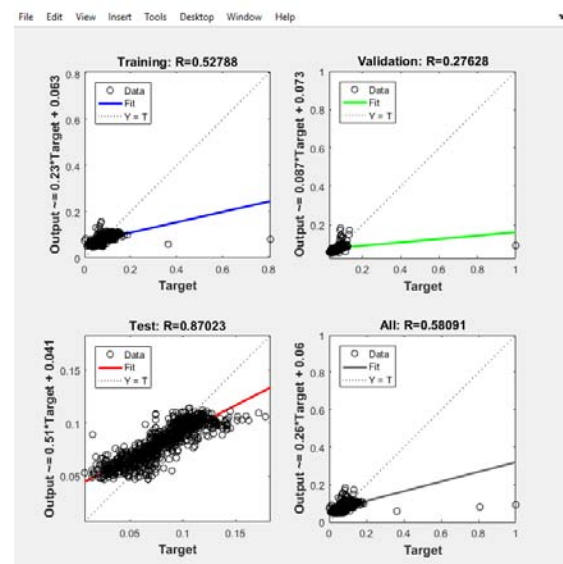


Fig f: Regression Analysis

Comparison of Results:

Table 1 lists the results obtained by applying our approach. It describes the values of performance measures for all the attributes in dataset. This approach has successfully predicted the future water quality trends by predicting the values for attributes which affect the quality of water. Such analysis of water quality would help to take some necessary actions to improve the quality of water. The work can be extended further to identify the relevant attributes and reduce the non-relevant attributes from the data set. The impact of attribute reduction can be studied by analyzing the results thus obtained.

ATTRIBUTE NAME	REGRESSION VALUE	MSE VALUE
PH	0.87492	0.009155
CONDUCTIVITY	0.58	0.007658
DISSOLVED OXYGEN	0.58091	2.6906e-08
TURBIDITY	0.6263	0.021035
NITRITE	0.6345	1.5695e-05
NITRATE	0.30305	0.008596
TOTAL COLIFORMS	0.31438	1.3511e-05
E-COLI	0.6189	3.1997e-05
FAECAL STREPTOCOCCI	0.31438	3.1997e-05
B.O.D	0.021035	0.021035
C.O.D	0.68581	3.1997e-05
PHOSPHATE	0.5809	0.00063082
SALINITY	0.7415	0.0079619
AMMONIUM	0.6513	0.005879

Table 1: Comparison of performance measures for parameters

References

- [1] Jiawei Han and Micheline Kamber, “*Data Mining : Concepts and Techniques*”, 2nd edition.
- [2] Kamakshaiah.Kolli and R. Seshadri, “*Ground Water Quality Assessment using Data Mining Techniques*”, International Journal of Computer Applications (0975 – 8887) Volume 76– No.15, August 2013
- [3] Ivana D. Radojević, Dušan M. Stefanović, Ljiljana R. Čomić, Aleksandar M. Ostojić, Marina D. Topuzović and Nenad D. Stefanović, “*Total coliforms and data mining as a tool in water quality monitoring*”, African Journal of Microbiology Research Vol. 6(10), pp. 2346-2356, 16 March, 2012
- [4] Yafra khan and Chai Soo see, “*Predicting and analyzing water quality using Machine Learning: A comprehensive model*”, IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2016
- [5] J an-Yee-Lee, “*Applying theory in predicting the arsenic contamination of groundwater in historical blackfoot disease territory*”, IEEE ninth international conference on natural computation, 2013
- [6] Ting-Nien Wu and Chiu Sheng Su, “*Application of Principal Component Analysis and Clustering to spatial location of Groundwater Contamination*”, IEEE Fifth international conference on fuzzy systems and knowledge discovery, 2008



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Secure Virtualization Environment with The Aid of SELinux

Soumya Bhowmik

ITSS, CDAC Bangalore Electronics city Phase 1

Bangalore-560100, India

Abstract: In a virtualized environment, it's possible to build number of independent virtualized systems on a server and host different services in isolated environment. Though virtualization helps in better resource allocation, system management, savings on power & cooling, and allocation of resources to virtualized systems depending on demand, it may create greater threat in certain scenarios if not dealt properly.

Imagine a situation where a cracker breaks into a virtual machine and takes full control of the system. Or in case there is vulnerability in the hypervisor. The situation is intensified with the fact that almost all virtualization instances and containers require root privilege.

This paper is searching the answer for this problem with help of another powerful and emerging concept called SELinux. SELinux sub-system can help implementing a policy based on Mandatory Access Control (MAC) mechanism and provides flexible security policy configuration, which in turn allows greater security in virtualized environment.

Keywords: virtualization; hypervisor security; kvm; libvirt; svirt; selinux; mls; mcs; mac

I. Introduction

Virtualization is a generic technology used for running services, usually more than one operating system simultaneously and in isolation from other programs on a single system.

The heart of virtualization is hypervisor. This is in general term, a subsystem or software layer to control hardware and for running multiple operating systems, called virtual machines (VMs) or guests, on a single physical server or system. This system which lets create guests on top of its operating system -- is called a host. Before virtualization came into existence, usually there were isolated physical servers. In case a cracker compromises the integrity of the server, he/she gets control of just that server and network attached systems with respective permissions. The next step is that the cracker has to launch network attacks targeting other servers in the same network. System administrators has numerous features and tools for defending the server against network attacks: firewalls, network traffic analysis tools, Intrusion Detection System (IDS), Intrusion Prevention System(IPS), Unified Threat Management (UTM) etc.

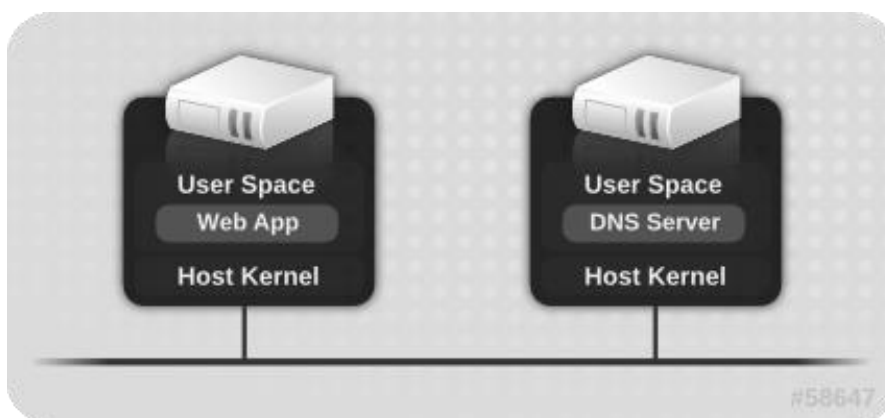


Fig 1. Non-Virtualized Environment [2]

After virtualization comes into picture, more than one services/systems are hosted in same server. If one VM is compromised, the intruder just needs to tunnel through the hypervisor into host system. If a hypervisor has known vulnerabilities, the intruder can gain control over all of the virtual machines in that respective host. The cracker can even access and write into any virtual host images which are reachable from that compromised host machine.

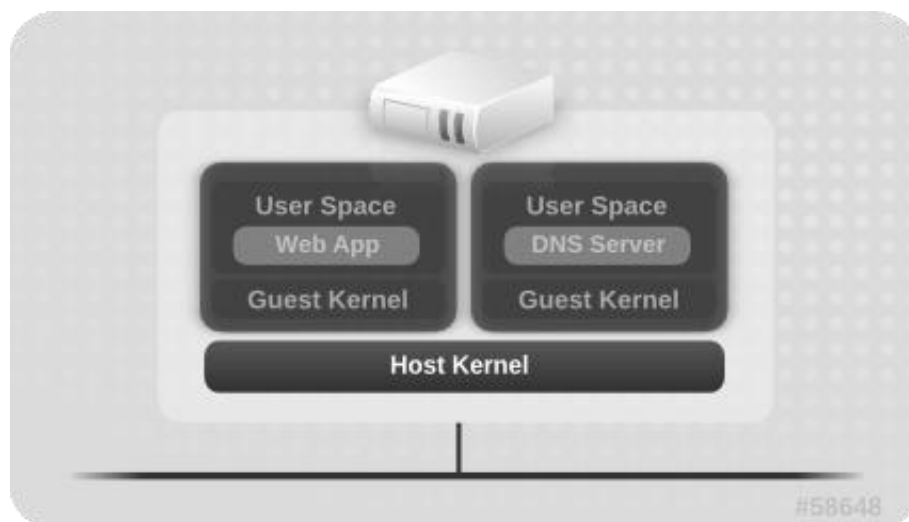


Fig 2. Virtualized Environment [2]

The threat under discussion is not merely theoretical; vulnerabilities already exist on hypervisors which are exploited in number of instances. If one guest gets compromised, then the attack can reach other guest instances within the same network.

News of breaking into the XEN hypervisor is all over the community, as documented here [1]

Necessity of Visualization Security

Among all the security concerns related to visualization, following are worth mentioning [2]

- In case the host/hypervisor becomes prime target - it may turn into a single point of failure for guests & data.
- It's possible for VMs to interact with each other in non-desired manner. Vulnerable hypervisor can let the compromised guest bypass and reach out to other resources residing on the same host.
- Tracking and maintaining services & resources can be difficult in case of large data center environments ; which may raise the need for greater management of resources, including but not limited to proper patching, monitoring and troubleshooting.
- Lack of proper understanding and knowledge of technical staff about virtualization technologies, in addition to inadequate skill sets, sub minimal experience in virtual environments often leads to unidentified vulnerabilities.
- Distributive nature of storage resources which in turn rely upon several physical systems, leads to super complex environments, mismanaged & poorly configured systems.
- Traditional sources of security risks within the computing environment aren't minimized or taken care of fully with Virtualization technology; which alleviates the necessity of securing the entire solution stack, in lieu of securing only the virtualization layer.

II. SELinux

SELinux is the primary Mandatory Access Control (MAC) mechanism integrated into a number of GNU / Linux distributions. Initially incubated as the Flux Advanced Security Kernel (FLASK) development by University of Utah, United States, Flux team and the US Department of Defense. Later supported by the NSA and released as open source software.

The usefulness of SELinux

In brief words [5][7]

- When SELinux is enabled, granting access to resources and relevant operations on them (e.g. read, write) are defined by the policy (i.e. prohibiting all access unless granted by policy). This feature results in 'mandatory access control' (MAC) system with the help of SELinux.
- Particular application can be restricted into its own 'domain', and granted minimal privileges to other applications necessary for its functioning, with the help of SELinux.
- In case a rouge application tries to do something outside its normal privileges granted by policy (intentional or otherwise), SELinux eliminates the possibility.
- It may happen sometimes, an application does something allowed by policy which results in some form of harm to the system, SELinux has capability to sustain certain degree of such damages.
- User login sessions are confined to their own domains, resulting in restricted accessibility.

- Applications such as X-Windows are difficult to restrict as by principle they are designed to have total access to all resources. Sandboxing services offered by SELinux can generally overcome these issues.
- Though SELinux doesn't stop all viruses/malwares to affect the system, it limits the damages or leaks to great extent.
- SELinux limits the effects of Kernel vulnerabilities, though it can't fix it .
- Finally, selecting a good design for SELinux security policy is important, as it doesn't stop anything granted by the policy.

Three forms of access control in SELinux

- Type Enforcement (TE) : Access to objects controlled by types, attached to objects.
- Role-Based Access Control (RBAC) : Used to define a set of roles that can be assigned to users.
- Multi-Level Security (MLS) : SELinux optionally provides MLS abilities, which makes it possible to define hierarchical sensitivity levels and corresponding categories to objects & subjects.

III. Mode of threat transmission in operating system virtualization

- Case 1: Propagate to host from malicious guest
 - Vulnerable virtualized systems can let an attacker get privileges as almost all Guest OS runs with root user access privileges in Linux virtualization system.
- Case 2: Malicious guest attacks other guests
 - In the form of DoS attack, information stealing and more

IV. libvirtd with svirt

In case of most linux OSs, virtualization is achieved with KVM (Kernel Virtual Machine).

Daemon that manages all virtual machines, in case of KVM, is called libvirtd. All VMs run as distinct processes. Virtual images are stored in the form of files or devices like logical volumes and iscsi targets.

libvirt is a stable and well-tested VM management framework which includes in general form– virsh :

sophisticated command line tool for libvirt, used extensively by administrators.

The sVirt project builds upon SELinux to further facilitate virtual machine isolation and control sharing. For example, fine-grained permissions can be applied to group virtual machines together for sharing resources.

sVirt provides a MAC framework for virtual machines in collaboration with libvirt. This architecture in turn makes it possible for all virtualization platforms supported by libvirt & all MAC implementations supported by sVirt to interoperate.

sVirt uses process based operations, labels and restrictions to facilitate additional security and control over guest instances, just like all other services under the shield of SELinux. Generally application of labels to resources on the system, done automatically, depending on the presently running virtual machines (dynamic). Administrator can also manually specify the resource label (static), to address special requirements if exists.

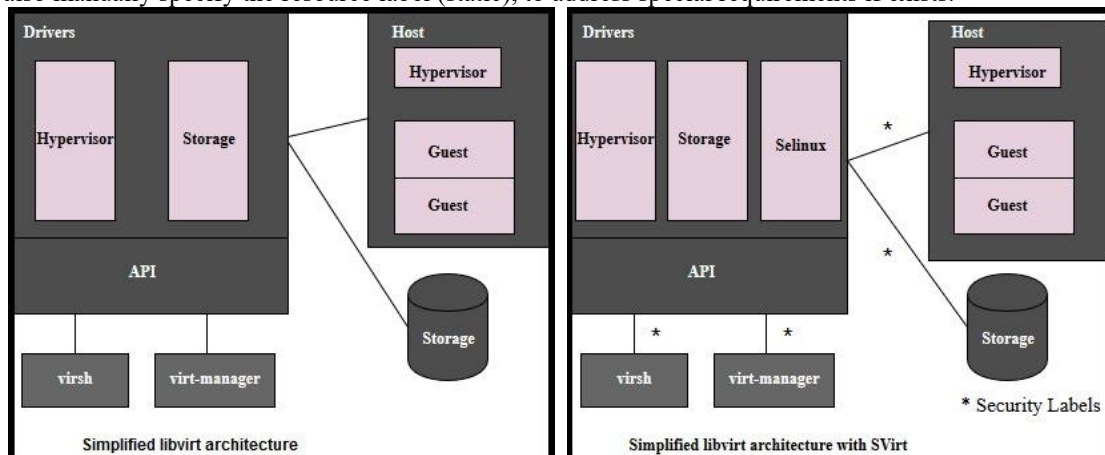


Fig 3 : libvirt architecture with & without selinux

In Fedora 11 sVirt was first introduced as a feature, and later integrated as a security mechanism for Linux-based virtualization applications like (KVM)/Qemu, and lguest. The primary developers are Dan Walsh and James Morris, of Red Hat Inc.

Some of the goals/use-cases of sVirt are listed as follows: [4][6]

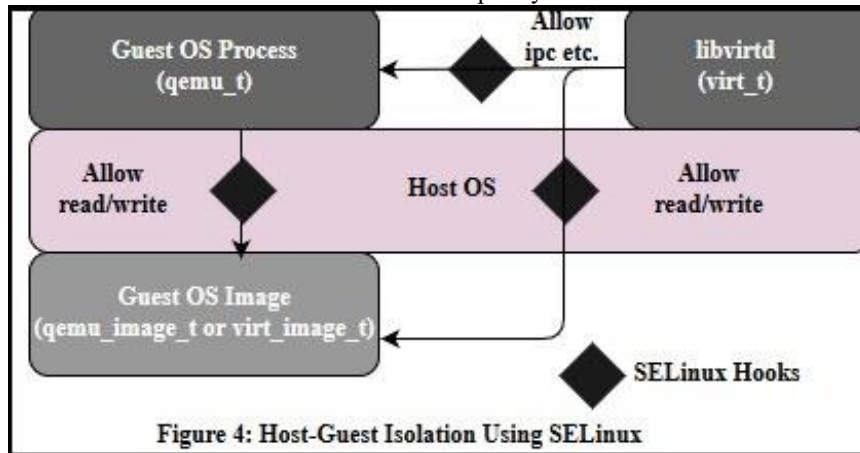
- Facilitate virtualized services with a level of isolation at par with what earlier offered by physical isolation.
- More protection of the host from potentially compromised VM guests.
- Heightened protection of VM guests from each other in case the host is vulnerable or malconfigured. A level of protection also maintained in case of a compromised guest.
- Desktop applications are executed in strongly isolated environment in separately labeled VMs.

V. Proposed Solution

Two broad approaches to handle the situation of malicious guest OS accessing host or other guests in the same host can be envisioned.

1. Isolation between host OS and guest OS with Type Enforcement

In this strategy, distinct security labels are applied to individual VM instances and their resources as well, in order to keep each VM isolated from one another via MAC policy.



During configuration/creation of VMs with the help of libvirt toolchain, there exist an option to make the VM "isolated" from all other VMs, in terms of resources. The libvirt conducts all the labeling & policy configuration behind the curtains.

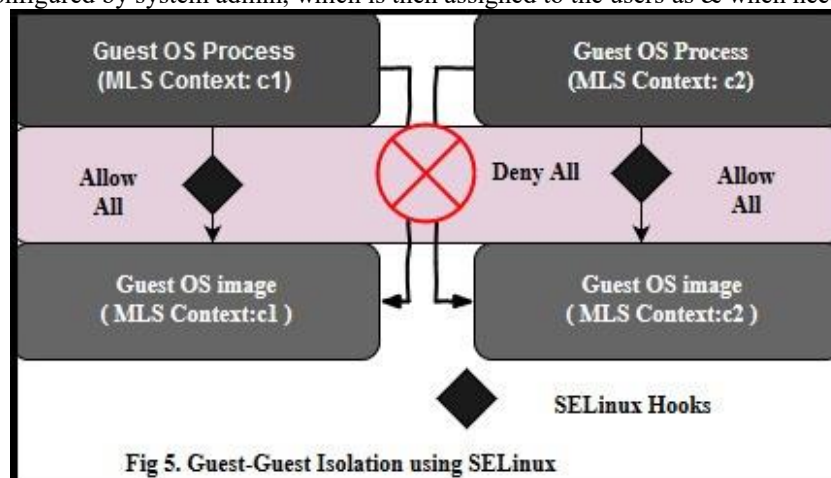
The SELinux policy has rules to grant the *svirt_t* labelled processes read/write access on *svirt_image_t* labelled files and devices.

In this way, the host machine can be protected from any of its malicious VM guests. Particular virtual machine can only access the files and devices with the corresponding labels assigned to that VM. A malicious virtual machine is not allowed to access host home directory, for instance, though the VM may be granted root privileges to run on the physical host system.

In spite of raising expectations, this kind of protection does not restrict one VM from targeting another VM in the same host. One needs a way to label the domains and the image files with the same TYPES. Yet, on top of that, needs to prevent VM1 (running as *svirt_t* label) from corrupting VM2, which would also be labeled as *svirt_t*.

2. Isolation between guests with Multi Category Security (MCS)

MCS is an added optional feature in SELinux, with the option of labeling user files with relevant categories. In most cases, categories are simply text labels, for example 'HR_Classified' or 'Employee_Attendance'. First the categories are configured by system admin, which is then assigned to the users as & when needed.



The configuration of categories can be accomplished by editing the file `/etc/selinux/targeted/setrans.conf`. It maps the human-readable formats to its corresponding internal MCS categories.

The following snippet shows the typical contents of the `setrans.conf` file:

```
s0:c0= HRclassified
s0:c1= EmployeeAttendance
s0:c2=Miscellaneous
s0:c3=SharedSecret
```

From the above snippet, the left hand side of the 'equals to' sign is the MCS security level, with the corresponding human readable value on the right.

In virtualized environment each VM is assigned unique MCS categories. SELinux security policy does not permit different categories interact among each other, even if they have similar values. The corresponding Device/image for each VM is labelled with the VM's MCS label, to restrict VMs from interacting with each other's resources.

Table 1: Example VMs and Corresponding Labels		
Name	VM Process Label	VM Image Label
VM 1	system_u:system_r:svirt_t:s0:c325,c239	system_u:object_r:svirt_image_t:s0:c325,c239
VM2	system_u:system_r:svirt_t:s0:c303,c519	system_u:object_r:svirt_image_t:s0:c303,c519

In the example of Table 1, libvirt creates two virtual machines(VMs) with dynamically generated SELinux labels

To obtain the SELinux labels for a running VM process, following command is used:

```
# ps -eZ | grep qemu
system_u:system_r:svirt_t:s0:c325,c239
```

SELinux labels for the image, is obtained as follows:

```
# ls -lZ /var/lib/libvirt/images/vml.img
system_u:object_r:svirt_image_t:s0:c325,c239
```

In the above snippets, the MCS labels are to be noted. Here, s0 indicates the 'Sensitivity Level', and c325 & c239 indicate the 'category of the data'. Administrators can define these during configuration time.

If someone tries to change the SELinux context (using the chcon command) of the VM image file, it will throw up SELinux AVC (Access Vector Cache) denials to be evaluated by the administrator.

SELinux prevents VM1 (system_u:system_r:svirt_t:s0:c325,c239) from accessing VM2's image file (system_u:object_r:svirt_image_t:s0:c303,c519) -- the VMs can't read/write each other's data, either residing on disk or in memory. Therefore it's very hard to attack each other.

VI. Inference

The main experiment done here is the application of SELinux TE and MCS/MLS techniques for implementing greater security and reliability in virtualized environment.

These two methods are discussed which are instrumental in implementing guest-to-host isolation and guest-to-guest isolation respectively -- validating the use of SELinux for security in virtualization technologies.

References

- [1] XEN Vulnerability <http://invisiblethingslab.com/resources/misc08/xenfb-adventures-10.pdf>
- [2] Redhat Enterprise Linux 7, Virtualization Security Guide, published under Creative Commons License
- [3] Dan Walsh blog <http://danwalsh.livejournal.com>
- [4] sVirt: Hardening Linux Virtualization with Mandatory Access Control, James Morris, Red Hat Security Engineer, Linux.conf.au, 2009, Hobart, Australia
- [5] System Administration, Second Edition, Sven Vermeulen, Packt Publication, December 2016, ISBN978-1-78712-695-4
- [6] Fedora Project wiki, http://fedoraproject.org/wiki/Features/SVirt_Mandatory_Access_Control
- [7] SELinux Project documentation https://selinuxproject.org/page/NB_Overview

Acknowledgement

- Images created by <https://www.draw.io> online free tool.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

ANALYSIS ON ROAD ACCIDENTS DATA TO IMPROVE ROAD SAFETY

Nidhi Kalra¹, Kriti Saroha²
School of Information Technology (SOIT)
CDAC-Noida, India

Abstract: Road accidents are serious issues to the society, they could cause both physical and economic damage to society. Infrastructure is also distorted due to road accidents. Road accidents could lead to both fatal and non-fatal consequences. Many algorithms of data mining have already been used on road accident dataset, to analyze the data for determining patterns and reasons which lead to most of the road accidents. It is estimated that 1.2 million deaths and 50 million injuries occur worldwide each year due to road accidents. Early detection of accidents prone area could help to reduce road accidents by taking appropriate measures of security and rules. It could be beneficial to provide appropriate services for accident prone areas.

Keywords: Road accident mining, clustering, info-gain evaluator, association-rule mining, data mining.

I. Introduction

It has been estimated by National crime record bureau ministry of home affairs that 4,00,517 accidents per thousand of population in 2013 and 4,51,757 accidents per thousand population in 2014 has been occurred in India. Rate of accidents is increasing in India with the passing years, it was recorded 32.6% in 2013 and 36.3% in 2014. People of age group between 30-45 years have been found to be more involved in road accidents [4]. World Health Organization (WHO) has described Traffic accidents report which states, the whole world suffers 20 to 50 million injuries and 1.24 million deaths each year due to that road accidents. Moreover, cost of 100\$ billion is spent in road accidents every year, as estimated by the Centers for Disease Control and Prevention. Every 10 seconds emergency room gets a patient injured in road accidents. Almost 40,000 deaths are caused due to road traffic accidents [2].

Data mining have proven to be useful in mining road accidents dataset. It finds patterns in the dataset, which could infer information to evaluate rules on problems. Data mining could infer highly accident prone areas and help to propose the possible solution to the problems identified. All this would help to decrease economic and social cost of road accidents [6].

Extracted knowledge and patterns would be used to infer results and improve road security. It would be beneficial to collect maximum number of information from road accidents scene to improve analysis.

II. Literature Survey

Various studies have been carried out on the different aspects of Road Traffic Accidents (RTA's). This section briefly describes and compares the techniques and results of different researchers.

Sachin Kumar and Durga Toshniwa [1] have given a novel approach to mine road accidents dataset. Authors used dataset from Dehradune, Uttarakhand from year 2009 to 2014. Dataset contains 15574 road accidents records and 13,09,640 accidents records have been used after preprocessing. Two of data mining techniques have been used by authors: k-mode clustering algorithm and association rule mining technique. Initially dataset is clustered into various clusters using k-mode algorithm with increasing values of k. After the value of k reaches 6, k-means gives maximum accuracy and hence it is selected as working value for implementation. Authors then used association mining technique to generate rules for every six cluster separately. This was proved to be beneficial as to understand patterns and problems for every cluster and generated better rules. Authors have described every cluster separately formed after k-mode algorithm, and rules generated for clusters. Example: cluster first states hilly areas are more prone to road accidents. The results can be utilized by the concerned officers of traffic and road safety department of India to put some accident prevention efforts in the areas identified for different categories of accidents to overcome the accidents.

Ait-MloukAddi, Agouti Tarik and Gharnati Fatima [2] gave an approach to mine road accidents dataset from Morocco. Dataset used was from year 2002 to 2014, with 11 attributes and 50 records. Authors suggested an approach and divided it into two modules, first were association rules extraction and second multi-criteria

analysis. Association rules extraction module, selects one of the association rule algorithm for extracting rules from dataset. Multi-criteria analysis was used for selecting the most relevant rules from the generated rules. Multi-criteria works on three problems that are: selection, sorting and ranking. Electi-tri method was selected to work on ranking problem. It reduces the number of rules generated by Apriori association. Originally 14 rules were generated out of which 12 rules were selected and 2 redundant rules were eliminated by Electi-tri method. Authors in this study have focused, on the factors influencing occupational accidents in Morocco. They have concluded their results with accuracy of 85%.

František Babi and Karin Zuskáová [3] suggested an approach to mine dataset for UK. Authors used dataset of UK from year 2005 to 2015. Dataset included 1 million records and 67 attributes. Dataset was divided into three sub-datasets: accidents, casualties and vehicles. After, removing redundant and not related attributes. 32 attribute out of 67 were selected: 16 of accidents, 7 of causality and 9 of vehicles. Authors suggested two alternatives to mine the dataset one is descriptive analysis through Decision tree algorithm and other is predictive analysis through Apriori algorithm. Authors have used three variations of decision tree for analysis. Results are discussed as: 18.24% error with Random Forest, 14.47% error with Gradient Boosted Classification and 14.63% with Random Forest Big Data. For descriptive analysis authors used Association Mining algorithm. Among three algorithms used for implementation of predictive analysis, the best accuracy achieved was 85% by using Gradient Boosted Classification and Random Forest for Big Data. Descriptive analysis generated many simple and complex rules.

Suvarna Gothane and Dr. M. V. Sarode [4] gave an approach on Indian road accidents. Dataset used was from year 2013 to 2015. Dataset includes 13 attributes with 155 instances. Authors have used two data mining algorithm in paper: first is Info-gain Attribute Evaluator that filters out the non-relevant attributes to the problem, initially dataset had 13 attributes out of which one attribute class was eliminated. After that authors used Apriori algorithm to generate rules. Authors have discussed their results with 85% accuracy. Basic aim of this paper was to eliminate those attributes which are not related to the road accident evaluation using attribute reduction and then generate best rules to evaluate patterns and problem.

Rui Tian and Zhaosheng Yang and Maolei Zhang [5] suggested an approach to mine dataset of china. Dataset used was from year 2005 to 2009; they have selected 100 records from dataset and used two approaches: Rough set theory and Apriori algorithm. Rough set theory works on the fuzzy boundaries, it sets limit to the boundaries and shape the data. Dataset in this paper is divided into four factors of accidents: Vehicle factor, Road factor, Environment factor and Driver factor. Authors have discussed their result with 86% of accuracy.

III. Comparison table

This table describes the comparative analysis of all the papers discussed above.

Table 1: Comparison of different papers in Literature Survey

	A data mining approach to characterize road accident	An approach based on association rules mining to improve road safety in morocco	Descriptive and predictive mining on road accident data	Analyzing Factors, Construction of Dataset, Estimating importance of factor and generation of association rules for Indian road Accident	Method of Road Traffic Accidents Causes Analysis Based on Data Mining
Algorithm	Clustering Association rule	Apriori Multiple criteria analysis	Classification Association rule	Info gain attribute evaluator Apriori algorithm	Rough set theory Association rule
Dataset duration	2010-2014	2002-2014	2005-2012	2013-2015	2005-2009
Accuracy	-	85%	-	85%	86%

IV. Proposed Approach

This section describes the approach that would be used to mine road accident dataset.

The approach clusters the dataset using a clustering algorithm on the basis of various parameters like frequency of accidents in particular region, drunk-driver, and person involved in accidents e.t.c. Then Info-gain Evaluator is used to reduce the no. of attributes. This would remove the redundant and on-relevant attributes. Clusters are then re-formed according to reduced attributes. After the clusters are generated, association mining would be performed on every cluster separately to generate rules for every cluster. This would be beneficial to understand the problems associated with every cluster separately. Flow chart for the proposed approach shown below in Figure1:

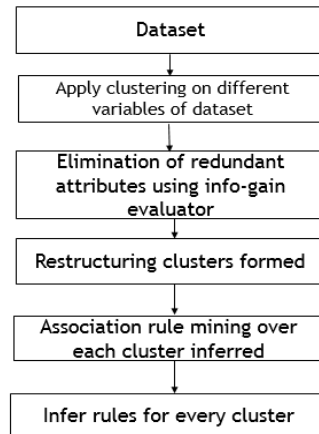


Figure 1: proposed approach

V. RESULTS

Some of the results are discussed in this section. Initially, clusters were formed using the variables like: Frequency of accidents in a state, No. persons involved in an accidents, Drunk Drivers, weather and etc. Clustering was performed using three algorithms and the results/ performed was compared for all the algorithms. The comparison is shown in Table 2.

Table 2: description of results through clustering

Algorithms	Cluster instances	Mean square error value	Log likelihood	Incorrectly clustered instances
K-means	7801(60%) 4010(31%) 1115(9%)			75.1044%
EM clustering	4782(37%) 3052(24%) 5092(39%)	16228.212		83.011%
Farthest first	8047(62%) 2(0%) 4877(38%)		-155.77449	79.5451%

After the comparison of results, k-means algorithm was selected for further implementation. Clusters were formed according to the variables like number of Persons and Drunk people involved in accidents. Value of k was selected as three.

Some of the clusters are discussed below:

Clusters formed according to no. of persons involved are:

Cluster1 includes accidents where less than ten people are involved. Cluster2 describes set of instances where exact of eleven people are involved. Cluster3 includes instances where more than eleven people are involved.

Clusters formed according to drunk drivers involved:

Cluster1 includes instances of less than three drunk people involved. Cluster2 groups instances of less than five people and Cluster3 for more than or equal to five people involved.

After performing clustering, Info-gain evaluator was used to perform ranking over the dataset attributes, and hence non-relevant attributes were removed. Re-forming clusters after reduction of attributes reduced the mean-square errors in clusters.

After that association rule mining was applied on every cluster separately. Result of association according to person involved is:

Best rules found:

- HIT_RUN=0 FATALS=1 11256==> LABEL=0 11218 <conf:(1)>lift:(1) lev(0) [15] conv(1.36)
- FATALS=1 11787==> LABEL=0 11745<conf:(1)>lift:(1) lev(0) [13] conv(1.29)
- HIT_RUN=0 12372==> LABEL=0 12318<conf:(1)>lift:(1) lev(0) [4] conv(1.06)
- WEATHER=1 11417==> LABEL=0 11366<conf:(1)>lift:(1) lev(0) [2] conv(1.04)
- PEDS=0 11187 ==> LABEL=0 11129<conf:(0.99)>lift:(1) lev(0) [-5] conv(0.89)
- LABEL=1 12865==> HIT_RUN=0 12318 <conf:(0.96)>lift:(1) lev(0) [4] conv(1.01)
- FATALS=1 LABEL=0 11745 ==> HIT_RUN=0 <conf:(0.96)>lift:(1) lev(0) [-23] conv(0.95)
- FATALS=1 11797==> HIT_RUN=0 11256<conf:(0.95)>lift:(1) lev(0) [-25] conv(0.95)

9. FATALS=1 11787==> HIT_RUN=0 LABEL=0 11218<conf:(0.95)>lift:(1) lev(0) [-14] conv(0.97)
10. LABEL=0 12865 ==> FATALS=1 11745<conf:(0.91)>lift:(1) lev(0) [13] conv(1.01)

Rules states that, fatal accidents doesn't involve hit-run. Accidents that include less than ten people doesn't have hit-run involved but result in fatal accidents. Cluster1 doesn't include accidents with pedestrians.

Association rules according to drunken drivers are:

1. WEATHER=1 11417 ==> WEATHER1=1 11417 <conf: (1)>lift (1.13) lev (0.1) [1332] conv (1332.84)
2. WEATHER1=1 11417 ==> WEATHER1=1 11417 <conf: (1)>lift (1.13) lev (0.1) [1332] conv (1332.84)
3. WEATHER=1 label=0 11349==> WEATHER1=1 11349 <conf: (1)>lift (1.13) lev (0.1) [1334] conv (1324.9)
4. WEATHER1 =1 label= 0 11349 ==> WEATHER1=1 11349 <conf: (1)>lift (1.13) lev (0.1) [1324] conv (1324.9)
5. HIT_RUN=0 FATALS=1 11256==> LABEL =0 11211 <conf: (1)>lift (1) lev (0) [28] conv (1.69)
6. FATALS=1 11787 ==> LABEL=0 11748 <conf: (1)>lift (1) lev (0) [28] conv (1.69)
7. HIT_RUN=0 12372==> LABEL =0 12303 <conf: (0.99)>lift (1) lev (0) [1] conv (1.01)
8. WEATHER1 =111417==> LABEL=0 11349 <conf: (0.99)>lift (1) lev (0) [-2] conv (0.95)
9. WEATHER1 =111417==> LABEL=0 11349 <conf: (0.99)>lift (1) lev (0) [-2] conv (0.95)
10. WEATHER1 =1 WEATHER =1 11417==> LABEL=0 11349 conf: (0.99)>lift (1) lev (0) [-2] conv (0.95)

Rules conclude that most of the fatal accidents occur in rainy season. If accident involve less than three drunk people then accidents are fatal but doesn't involve hit-run.

VI. Conclusion

Mining on Road accident dataset is beneficial as early detection of the highly accident prone areas could help to take appropriate safety measures. This would help to reduce loss in property and people. Safety measures can be taken by government agencies to reduce accidents because of the identified reasons in accident prone areas. This paper concluded importance of Info-gain evaluator to find relevant attributes for analysis and association rule mining found out best rules to road accidents.

VII. References

- [1] Sachin Kumar, DurgaToshniwal "A data mining approach to characterize road accident locations" Journal of Morden Transport, Springer-2015
- [2] Ati-Mlouk Addi, Agouti Tarik and Gharanti Fatima "An approach based on association rules mining to improve road safety in morocco" Technology for Organization Development IEEE-2016
- [3] Frantisek Babic and Karin Zuskacova "Descriptive and predictive mining on road accident data" International symposium on applied machine intelligence and informaticsIEEE-2016
- [4] SuwarnaGothane, Dr. M. V. Sarode "Analyzing Factors, Construction of Dataset, Estimating importance of factor and generation of association rules for Indian road Accident" International Advanced Computing Conference IEEE-2016
- [5] RuiTian, Zhaosheng Yang and MaoleiZhang "Method of Road Traffic Accidents Causes Analysis Based on Data Mining" Computational for Organizational Development IEEE-2010
- [6] Han J. and Kamber M, "Data Mining: Concepts and Techniques", Academic Press.

VIII. Acknowledgment

I want to acknowledge the department and institute for their guidance and support throughout the work. I would also like to express my gratitude to the members of department for their kind assistance and cooperation for completion of work.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Security of Data Store for E-Commerce Portal

Nidhi Gahlot¹ and Sanjay Ojha²

School of Information Technology (SoIT)

Centre for Development of Advanced Computing (C-DAC), Noida, India

Abstract: In present scenario, at third-party site that deal with data of e-commerce companies, the employee has direct access to the sensitive data of customers without the intervention of his senior official. This problem leads to compromise of security; the employee may get his hands dirty in committing any card related fraud or may leak the information out of the organization. There are many methodologies like access control, encryption, strengthening of authentication, and anti-virus that are practiced by organizations but there are incidents that prove data leakage and identity theft within the organization. In this research paper an improvised approach has been proposed to improve the security within the organization by using the concept of encryption and access permission.

Keywords: ECDH; MD5; access key; private key; internal leak; competent person.

I. Introduction

In present scenario, data leakage and identity theft are common issues that are faced by the commercial industry. Now a days when online business is at its boost, its the prime duty of the governing companies to safeguard the sensitive information of the customer that he provides while buying any product. There were several incidents reported where data theft was committed by an employee within the organization and lead to huge economic loss to the organization, when the employee himself is responsible for stealing the data is called internal leak, this data can be personal details of customers or other employees or may include company's secrets.

Many authors have earlier proposed the techniques that can prevent the internal leak and data theft [3].

Some loop holes are present in previous approaches that still make the data access procedure less secure for an employee to steal the data. Previous approaches include USB scan algorithm[8], Data Security Strengthening by Combining Fingerprints[3], Data Leakage Prevention System with Time Stamp[9], Secured user Authentication by using Enhanced ECDH Algorithm[5], An ECDH-based Light-weight Mutual Authentication Scheme[4].

The problems seen in previous approaches were, the algorithm was only efficient to save the data when the mode of data theft was a USB drive, the algorithm implemented two finger fingerprint technique by calculating minutiae positions of the fingers for authentication process so that in case of data loss the thief cannot create a false finger to access the data but the problem was what if the authorized person is the culprit himself, the algorithm categorized the data into confidential and non confidential but was efficient only for cases when data need to be confidential for some amount of time (hours, days or year), an key agreement algorithm was used for data access process, respectively.

The problem that was highlighted among all the mentioned approaches was that the data can be accessed directly without the permission or intervention of a competent person i.e the person who may not be able to view the data himself but can allow the access permission to other employees.

This research paper has introduced the enhanced security system to safeguard the data that is collected during in transaction at a third-party site by including some encryption algorithms and improvised access procedures. The approach not only focuses on access procedure where competent person is involved but also stores the data in encrypted form in the database, also a tractability feature is included to keep a check on log-in history of the employee. Algorithms that are used are ECDH and MD5, their usage is described in the upcoming chapters.

II. Methodology

To give the live feel for data collection the work was divided into three modules i.e. the shopping the as web application, a payment gateway and the third party application where the data security feature has been applied. The third-party application is divided into phases, one as log-in phase and other as data access phase their functionalities are described further with figures.

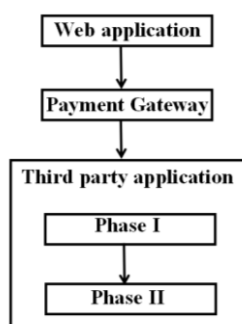


Fig 1. Modules of the overall research work.

For the third-party module the first phase is described as below.

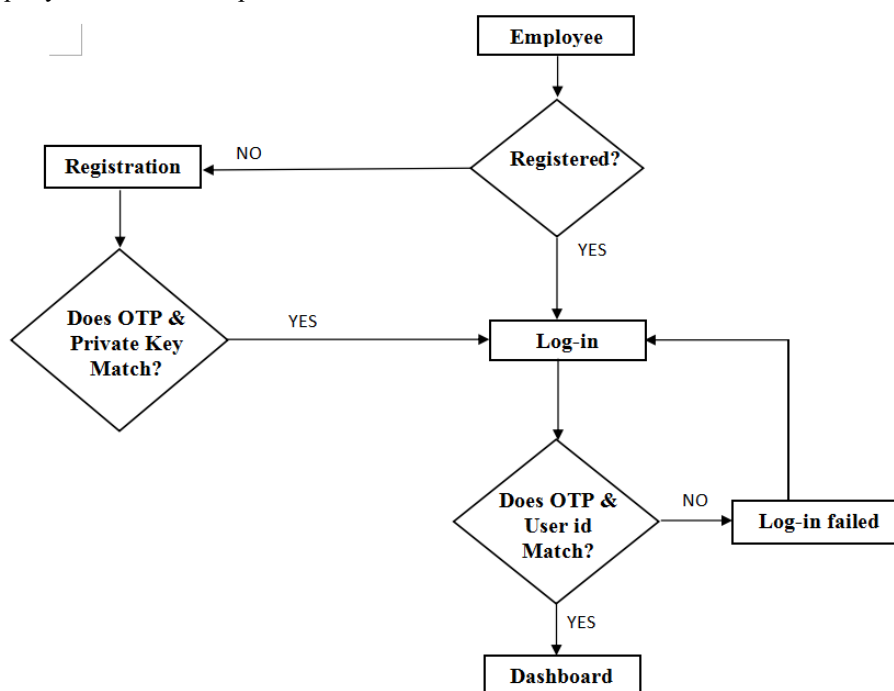


Fig 2. Phase I (log-in phase)

The phase begins with the process of registration if the employee is not registered. For the registration process the person needs to enter his personal details if all the credentials are verified then an private key need to be entered by the employee. This private key is assigned by some competent person of the organization. This key is generated using the ECDH algorithm. The usage of including this key in registration process is that on unauthorized person get registered with the system. When this key is entered correctly an OTP is sent to the employee's mail box which he can enter to complete the registration process and his user id gets created. Further if the employee is already registered he simply needs to log-in. For log-in process user id is required and an OTP is required. This OTP is system generated and is combination of first 3 digits of use id when converted into big integer and first 3 digits of a random number. Only when both user id and OTP are matched the employee is able to log-in and he is directed to dashboard. User id of the employee is saved in encrypted form in the data base using the MD5 algorithm.

Dashboard includes my account, daily transactions and client data. With this proposed approach the employee cannot simply view the data once he completes log-in rather he need to enter the access key which is part of data access phase and is explained in next phase.

Once the employee is able to log-in he is able to see the daily sales of the client he is allocated with. He can see his previous log-in history. Further, if he wants to access the data he need to generate the data access request that is forwarded to some other competent person. The decision is taken that competent person if to allow the access or to deny the permission that has been requested from the employee.

The employee generates the request, at the admin panel all the details of the employee will appear along with system details, the authorized person either can deny the request or accepts it, if he denies the request a dialog box will appear at the employee's panel with the message saying "access denied".

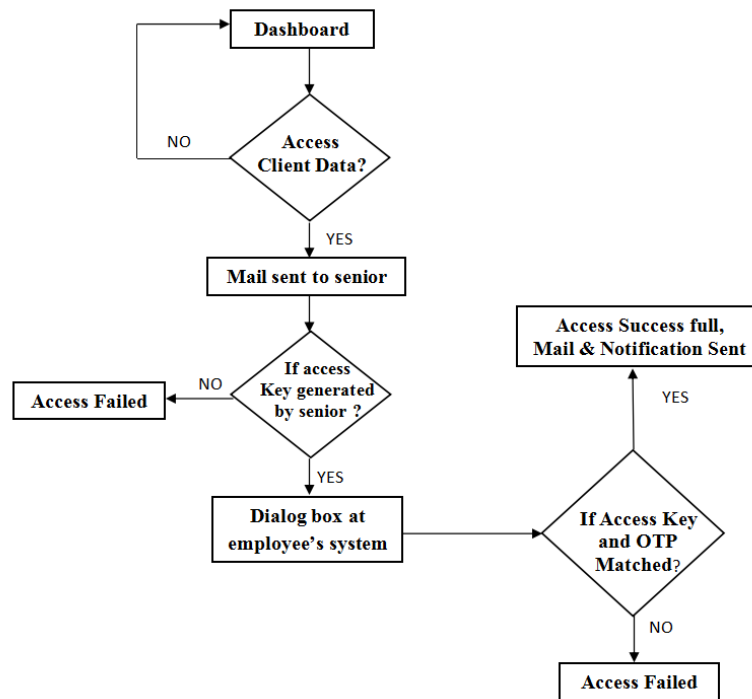


Fig 3. Phase II (data access phase)

If the authorized person grants the request a dialog box will appear at his panel asking or the access key, he is required to enter any alphanumeric 6 digit number, it is then converted to an encrypted form using MD5 algorithm and the forwarded to the employee for his data access process. The advantage of applying MD5 algorithm is that in case some suspicious person seeks into the transmission he is not able to view the actual access key. Now the dialog box will appear at the employee's panel asking for OTP and an access key that was generated by his senior. If the information entered by the employee is matched i.e. if the OTP and access key is correct the attempt is successful and a notification is sent to his dashboard with his log-in history otherwise if the information was incorrect the access is failed.

The advantage of sending the mail and notification to senior and employee respectively is that in case of any suspicious activity in the organization keeping track of employee's activity becomes easy i.e. tractability becomes easy and fast. Even the chances of an employee making an attempt to data theft are reduced because of his mind set as he is aware that someone is always keeping a check on him.

III. Result

In the following research work efforts are done to make the system more secure by using few encryption techniques. The overall security concludes OTP and at the registration level with the private key that was provided by the officials to the employee who want to register himself, later at the time of log-in another OTP is required along with user id of the employee. Employee is able to log-in only when he enters the right information. As mentioned above user id of the employee is saved in encrypted form in the database and the access key required for registration is generated using ECDH algorithm.

Further when the employee wants to log-in he is asked to enter the access key that is generated by some senior official. Due to this feature employee can never be able to view the sensitive or confidential data when he is not permitted by the senior. Also the log-in history is send to senior as well as the employee so that both of them are aware of the employee's activity. The access key that is generated by the senior is also saved in encrypted form in the database.

The system seems to be a good preventive measure to reduce data theft, internal leak and identity theft up-to some extent.

IV. Conclusion

The motive of this research work was to give an improved solution the problem of data theft, identity theft and internal leak. Before this work was started many other approaches of some other authors[1-11] were take as

subject of reading to see through the latest trends and practices that were previously followed. After the study it was concluded that there were some loop holes in the previous methodologies which needs to be work on. The main problem that was noticed was that the employee has direct access to the data without the intervention of any senior authority. In this research work an approach has been proposed that can limit the unnecessary access of the employee and can allow the access of authorized employees only when permission is granted by the senior. As there were some loop holes in the previous approaches but their work was best until someone miss used those loop holes. So it can be concluded that if we speak of security nothing is 100% secure only preventive measures can be taken to minimize the risk of letting the data steal by some intruder or an employee form the organization itself.

References

- [1] Abdulrahman Alruban; Nathan Clarke; Fudong Li; Steven Furnell, "Proactive biometric enabled forensic imprinting", International Conference On Cyber Security And Protection Of Digital Services(Cyber Security),IEEE 2016.
- [2] Amir Harel, Asaf Shabtai, Lior Rokach, and Yuval Elovici, "M-Score: A Misuseability Weight Measure", IEEE Transactions on Dependable and Secure Computing (Volume: 9, Issue: 3, May-June 2012).
- [3] Athira Ram A, Jyothis T S, "Data Security Strengthening by Combining Fingerprints" National Conference On Parallel Computing Technologies (PARCOMPTECH), IEEE 2015.
- [4] inhee Seo, Jihong Park, Young Jun Kim, "An ECDH-based light-weight mutual authentication scheme on local SIP", 2015 Seventh International Conference on Ubiquitous and Future Networks, IEEE 2015.
- [5] Nikhil Gajra, Shamsuddin S. Khan, Pradnya Rane, "Private cloud security: Secured user authentication by using enhanced hybrid algorithm", International Conference on Advances in Communication and Computing Technologies (ICACACT), IEEE 2014.
- [6] Nishakumari Lodha; Prasant Rewagad; Yogita Pawar, "Comparative Analysis of PAVD Security System with Security Mechanism of Different Cloud Storage Services", Fourth International Conference on Communication Systems and Network Technologies (CSNT), IEEE 2014.
- [7] Raja Mukhopadhyay, I. Mukhopadhyay "Data encryption in virtual machine transfer over cloud network", 7th Annual Conference on Information Technology, Electronics and Mobile Communication (IEMCON), IEEE 2016.
- [8] Saurabh Verma; Abhishek Singh, "Computer Forensics in IT Audit and Credit Card Fraud Investigation - for USB Devices ", ICAC (Fifth International Conference on Advanced Computing), IEEE 2014.
- [9] Subhashini Peneti, B. Padmaja Rani, " Data Leakage Prevention System with Time Stamp", International Conference On Information Communication And Embedded System (ICICES 2016), IEEE 2016.
- [10] Shabtai, Elovici, Rokach, "A survey of data leakage detection and prevention solutions", Springer Briefs in Computer Science, Springer, 2012.
- [11] Yucheol Cho, Sangjin Lee, "Detection and Response of Identity Theft Within a Company Utilizing Location Information", International Conference on Platform Technology and Service (PlatCon), IEEE 2016.
- [12] http://articles.economictimes.indiatimes.com/2015-06-25/news/63831555_1_reliance-jio-infocomm-ltd-bharti-airtel-mukesh-ambani-owned-jio (last accessed on 15/04/2016)
- [13] <https://www.americanbazaaronline.com/2016/04/15/tata-group-hit-940-million-trade-secrets-verdict-stealing-epic-systems-corp-s-soft-ware> (last accessed on 15/04/2016)
- [14] <http://news.softpedia.com/news/retiring-sysadmin-fakes-cyber-attack-to-get-away-with-data-theft-507992.shtml#ixzz4M8hbVr4e> (last accessed on 6/09/2016)
- [15] <http://www.dailymail.co.uk/indiahome/indianews/article-3820477/Fake-shopping-portals-dupe-online-pers-luring-unrealistic-discounts-festive-season.html#ixzz4M8itGmN5> (last accessed on 3/10/2016)
- [16] <http://www.amazon.in/> (last accessed on 15/02/17)
- [17] http://www.ebay.in/?aff_source=Google_cpc&site=Brand_New_Exact (last accessed on 13/02/17)
- [18] <http://timesofindia.indiatimes.com/city/hyderabad/cyber-crime-cases-shoot-up-postdemonetisation/articleshow/56192040.cms> (last accessed on 13/02/17)
- [19] <http://timesofindia.indiatimes.com/prosecuted-U-S.html> (last accessed on 01/02/2017)
- [20] http://www.dailymail.co.uk/news/http://articles.economictimes.indiatimes.com/2015-06-25/news/63831555_1_reliance-jio-infocomm-ltd-bh-a-rti-airtel-mukesh-ambani-owned-jio (last accessed on 15/04/2016)
- [21] <https://www.americanbazaaronline.com/2016/04/15/tata-group-hit-940-million-trade-secrets-verdict-stealing-epic-systems-corp-s-soft-ware> (last accessed on 15/04/2016)
- [22] <http://news.softpedia.com/news/retiring-sysadmin-fakes-cyber-attack-to-get-away-with-data-theft-507992.shtml#article-3796339/Five-MILLION-Brits-cancel-credit-cards-fraud-year-combined-2bn-shock-survey-reveals.html> (last accessed on 01/02/2017)
- [23] <http://fox6now.com/2014/12/16es.com/city/noida/Exec-falls-prey-to-Rs-1L-credit-card-fraud/articleshow/56192040.cms> (last accessed on 01/02/2017)
- [24] <https://www.justice.gov/usao-sdoh/pr/final-defendant-31-million-credit-card-fraud-scheme-sentenced> (last accessed on 01/02/2017)
- [25] <http://www.dailymail.co.uk/news/article-2378322/Feds-say-hackers-stole-160-Million-credit-card-numbers-largest-data-theft-cas-ten-charged-with-unlawfully-using-over-3800-credit-card-numbers> (last accessed on 01/02/2017)
- [26] <https://www.nowsecure.com/blog/2010/08/31/departing-employees-and-data-theft/> (last accessed on 04/9/2016)
- [27] <https://www.entrepreneur.com/article/272319> (last accessed on 04/9/2016)
- [28] https://www.symantec.com/about/newsroom/press-releases/2013/symantec_0206_01 (last accessed on 04/9/2016)

Acknowledgment

I take this opportunity to acknowledge all those who have guided me during this project. I express my earnest gratitude towards Mr. Sanjay Ojha, my project guide for his valuable encouragement and guidance. I would also like to express my gratitude to the faculty members for their kind assistance and cooperation during the development of our project.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

An Efficient Plagiarism Detection System using Boyer Moore Algorithm

Harsha Gupta¹ and Sanjay Ojha²

School of Information Technology (SoIT)

Centre for Development of Advanced Computing (C-DAC), Noida

B-30, Sector 62, Noida, Uttar Pradesh, INDIA

Abstract: Plagiarism is basically using someone else's ideas, or words, and then using them as if they are our own. As Internet is now accessible to most of the people whether they are students, employees or professionals. They can easily use Internet as source of information. Plagiarism is a serious issue in professional environment as well as in the education system. Plagiarism of report files, research papers is being done by the students without proper citing or referencing the authors of the documents. Sometimes they may result to legal problems like copy infringement.

Since then, a new kind of system has emerged, the plagiarism detection systems. There are several types: they can use databases (of thesis, books, or articles), Internet or comparison between files. This project is to check plagiarism in 2 type of search: Phrase search, searching where we don't consider the preprocessing steps and check for exact phrase and Text search, which does not contain stop words and stemmed words. Comparison between 2 files either in .doc or .docx format or pdf format can be done by frequency calculation, removal of stop words, stemming and then search string algorithm Boyer Moore. Also efficiency is calculated in terms of similarity between the documents using cosine similarity.

Keywords: Phrase search; Text search; String matching; Stop words; Cosine similarity; Boyer Moore.

I. Introduction

What is Plagiarism? Plagiarism is basically using someone else's ideas, or words, and then using them as if they are our own. This can be of many types. Some of them are as follows: Copying the text as it is without proper quotation marks and without proper citing of the source. Reordering the elements of the source text without citation. Paraphrasing without citation. This is paraphrasing plagiarism, where different words are paraphrased having similar meanings. When we use idea of someone else and then use it in our own way, this comes under Idea plagiarism. When we use our own content without proper citing in our new work, then it is known as self plagiarism. Phrase search, a searching where we don't consider the preprocessing steps and check for exact phrase. This is called Phrase search. Removing of stop words and changing the root words into their other root forms is also a form of plagiarism and this type of plagiarism is known as Text search.

II. Literature Review

Many research papers have been analyzed in order to improve the result of plagiarism detection system. Research part is that this paper have included the Phrase search altogether with text search. Some papers have focused on paraphrasing plagiarism detection [2] where similarity is considered not only in lexicographic terms but also in semantic terms where meaning of words is also considered. This paper basically evaluates paraphrase between the sentences. It is applied on PAN corpus and produced the result with help of SVM classifier. If output comes out to be +1 it means it is plagiarised if -1 it means it is not plagiarised.

This paper also focuses on Paraphrasing plagiarism detection. It makes use of syntactical information in documents and then it computes the similarity that is between them using word similarity measures which are based on WordNet and lexical databases [3]. Text's semantic similarity is obtained from the texts when document are converted into a semantic structural model which is unified. The Microsoft Research Paraphrase (MSRP) is a Corpus which is used in here for the analysis of the system.

Another paper focuses on the n-grams technique [4]. Text is divided into n-grams and then it is matched with the

document which is there in the storage. Stop words are removed and then stemming is also done which are preprocessing steps. It has used both term frequency and later inverse document frequency.

Term frequency (tf): Number of times a particular term is occurring in the document.

Document Frequency (idf): It is the number of documents in which it occurs.

Product of tf and idf is done to calculate weights, document vector is calculated and then cosine similarity is done using query vector and document vector.

III. Methodology

Many Plagiarism detection systems available and in reach to us but not all reveal out their methodologies and how they are doing the plagiarism detection task. This paper mentions the methodology used in the implementation.

Input that we want to check for plagiarism can be copy pasted text, doc files in .doc format or .docx format or pdf files. This paper will focus on basically 2 types of searches. One is Phrase search and other is Text search. Phrase search, a searching where we don't consider the preprocessing steps and check for exact phrase and Text search, which does not contain stop words and stemmed words. Stop words are the words that generally come in the sentences and are frequently occurring. Stop word list is made accordingly as there no specific list for it. This can be made according to your preferences and type of documents that you want to check for plagiarism. This paper uses Van Rijsbergen stop word list, words here are traditionally extracted by frequency analysis of all words in large corpus.

For Phrase search, documents are uploaded and submitted Boyer Moore algorithm which is an search string algorithm is applied and then similarity between 2 strings present in the document is calculated using cosine similarity.

Boyer Moore algorithm: It is a search string algorithm which works more efficiently when alphabet is moderately sized and pattern is relatively long. Unlike other search string algorithms it scan from right to left apart from left to right. We shift the pattern 'value times', and thus here comes the role of having lage pattern length.

Text: hardtooth

Pattern: tooth

Index: tooth

01234

Length of pattern is 5.

A bad match table is made and value is calculated by this.

Value = Length-Index-1.

Letter: toh*

Value: 4355 and later updated to 1255.

We shift the pattern 'value times'.

```
hardttooth
tooth|
tooth|
tooth|
tooth|
tooth|
```

For Text search, documents are uploaded and submitted. First of all Tokenization is done so that tokens of words are obtained from the text. Next step is to remove stop words like a, an, the, on, under etc. Further stemmed words are to be removed. These are the words which are obtained by changing the root words by appending with ed, ing etc. Porter stemmer is used here. Tokenization, Stop word removal and Stemming are the preprocessing steps in the process. Frequency calculation is also done to see the frequency of words that occurs in the document. Here after the

preprocessing steps Boyer Moore algorithm and cosine similarity is applied to calculate the similarity between 2 strings present in the document. Output is obtained in the form of match string of non match strings.

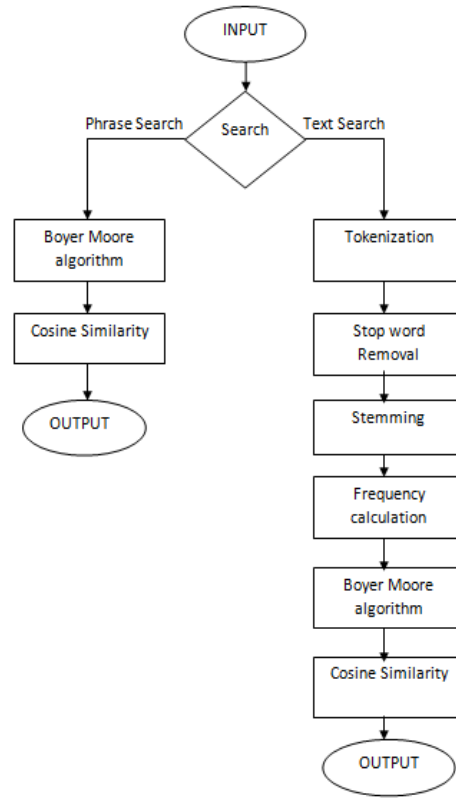


Fig 1. Proposed Approach

IV. Conclusion and Future work

For calculation of efficiency of Boyer Moore algorithm in Plagiarism detection system, it uses a corpus of around 15 documents. By comparing with a variety of doc files and well as pdf files efficiency comes out to be 96.68 by considering exact match words, phrase search and modifications done to the root word.

Basically this is an offline plagiarism detection system. This work can be extended to online documents checking. Not much efficient for short substrings, there comes the existence of Brute force approach and Knuth Morris Pratt algorithms for binary strings. Not applicable for copyright and protected pdfs and doc formats.

V. References

- [1] Sachin V. Shinde , Sangram Z. Gawali ,Devendrasingh M. Thakor “MAS a scalable framework for research evaluation by unsupervised machine learning - Hybrid plagiarism model ” International Conference on Peravasive Computing (ICPC) IEEE 2015
- [2] P.Vigneshvaran,E. Jayabalan, A.Vijaya Kathiravan “An Eccentric Approach for Paraphrase Detection Using Semantic Matching and Support Vector Machine” International Conference on Intelligent Computing Applications,IEEE 2014,pp 431-434.
- [3] Vaishnavi V I, Saritha M, Milton R S “Paraphrase Identification in Short Texts using Grammar Patterns” International Conference on Recent Trends in Information Technology (ICRTIT),IEEE 2013,pp 472-477.
- [4] Urvashi Garg and Vishal Goyal “Maulik: A Plagiarism Detection Tool for Hindi Documents” Indian Journal of Science and Technology,Vol9(12),DOI:10.17485/ijst/2016/v9i12/86631, March 2016, pp 1-11.
- [5] Jianjun Zhang , Xingming Sun, Jin Wang “Semantic Keyword-based Text Copy Detection Method “Advanced Science and Technology Letters 2014.
- [6] Abhay Nitin Pai,Chinmay Neelmadhav Bhusari “Plagiarism Detection System”International Journal of Innovations in Engineering and Technology (IJET),2013
- [7] Kusum Lata Pandey, Suneeta Agarwal, Sanjay Misra Rajesh Prasad “Plagiarism Detection in Software Using Efficient String Matching” Springer 2012.

- [8] <https://www.rff.com/flowchart-shapes.htm> (Last accessed on 28 February 2017)
- [9] <https://www.microsoft.com/net/tutorials/csharp/getting-started> (Last accessed on 25 February 2017)
- [10] <https://stackoverflow.com> (Last accessed on 25 February 2017)
- [11] <https://www.tutorialspoint.com/asp.net> (Last accessed on 25 February 2017)
- [12] <https://en.oxforddictionaries.com/definition/plagiarism> (Last accessed on 25 February 2017)
- [13] <https://www.codelooker.com> (Last accessed on 22 March 2017)

VI. Acknowledgement

It is my privilege and pleasure to express my profound sense of respect and indebtedness to Mr. Sanjay Ojha CDAC, Noida for his inspiration, guidance, cogent discussion, constructive criticism and encouragement throughout this dissertation work. In spite of his busy schedule, he made himself available to me in every possible way he could.

I also want to express sincere thanks to my family and my batch mates for providing immense help and without these I could not have done so well.

Restoration of Mural Images

Gunjan Mishra¹ and Tushar Patnaik²

School of Information Technology (So IT)

Centre for Development of Advanced Computing (C-DAC), Noida

Uttar Pradesh, India

Abstract: A mural is a wall painting, an artwork that is painted or applied directly on the wall or any other large surface area. Old mural images can deteriorate, get distorted, develop cracks, fade away and may even peel out due to various reasons including social, climatic, environmental, historical factors. An approach to virtually restore these mural images using the Digital Image Processing technology which tries to generate the original image. The suggested approach consists of four major steps which are described further in this paper. An Edge Enhancement process is implemented followed by K means clustering and averaging is performed. The final step is to perform Histogram equalization and its comparative analysis with other enhancement techniques like Sharpening and Adaptive histogram equalization. The results of the experiment are good and are providing improved images based on image restoration parameter Peak signal to noise ratio (PSNR). This implemented approach can also be used in future to restore faded deteriorated mural images.

Keywords: K means clustering, averaging, histogram equalization, adaptive histogram equalization, PSNR.

I. Introduction

Digital image processing has been popular for detection and recognition of different images. The different algorithms used in digital image processing solve the purposes for detecting and cleaning the noises and pixels in various image.

Wall painting is a human creation which depicts a culture expressions which exist from the ancient day till present day. Any kind of deterioration or destruction of these mural paintings can cause significant harm to our cultural heritage. Due to different climatic changes and other external impact of environment the wall paintings sometimes get distorted, the intensity of the colors of painting seems to be faded and since the perception capability of our eye is limited it becomes difficult to identify the details of faded image.. To preserve the history and diverse culture, the mural images are required to be restored as the original.

The mural painting recovery i.e the wall painting restoration to originality is a major challenge. Since lot of algorithms and filtering techniques have been developed for restoration of such type of images different research papers have given different approaches and results on different types of images but the look of images to the originality requires lot of analysis and computations. The degraded image was detected keeping ground truth data in view. The process of reconstructing a blurred, damaged or a noisy image to provide an uncorrupted image is termed as image restoration.



Figure 1. Few Deteriorated Mural Image

Restoration of Mural Images can be used for preservation of historical assets. A system that could restore mural images can become a contribution to heritage conservation societies which is built to conserve the national heritage. Building a system that could restore Mural images using digital image processing which is a technical

domain could build a bridge between the technical and art communities. Mural image preservation is not only a contribution to national heritage but also it can be used to infer historical discoveries. A number of old destroyed images carry building blocks for historical discoveries, a mural restoration system could help in reconstructing such images thereby providing an restored image which is close to original image. The restored image could provide the details that the deteriorated mural image failed to reveal

II. Literature review

Karianakis, N., et al. [2] had worked on an approach which was focused on restoring the missing parts of old wall paintings. They applied an algorithm for seamless image stitching of the missing area. In addition to this they also applied TV inpainting. TV inpainting was applied for extracting area and also for repairing the image.

S. Awate ,et al. introduced an adaptive filter, this filter would restore the images by finding out there statistical information. The filter was found to be capable of restoring different types of images.

Bhabatosh Chandra, et al. [1] proposed a patch matching technique. This technique will use a database which will have clean paintings , wherever there will be a patch that is to be filled that patch will be replaced by another patch which will be the best patch found out of all.

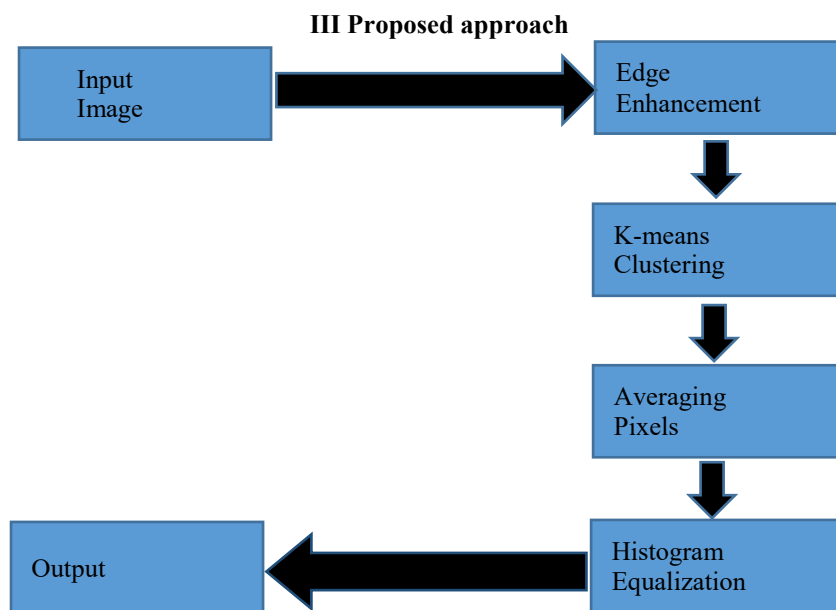


Figure 2. Proposed Approach

A. Input image

Input images are deteriorated mural wall paintings which are to be restored Given the ground truth data and if we have visual comparison with the deteriorated images the wall paintings are totally unclear and extremely deteriorated. The actual content of the image is totally hidden hence the restoration methods should adaptively select the low intensity pixels and enhance them.



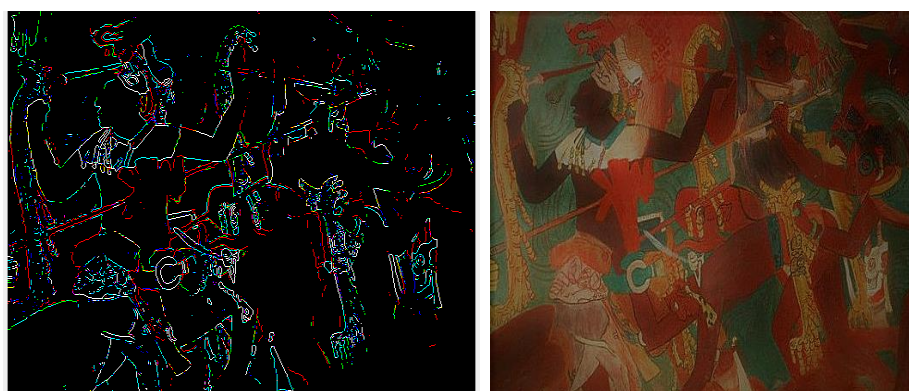
Figure 3. (a) Deteriorated Mural Images



Figure 3.(b) Ground Truth Data

B. Edge Enhancement

Since edges are the boundary of an defined structure colored image the first step to detect an image is the edge enhancement. For Edge enhancement edges are first extracted from the input deteriorated mural image. In this step we need to detect lines, these lines can be edges or other parts of the image. The next step was to use a standard operator sobel to detect and enhance edge. Sobel operator computes the approximation of gradient of the image intensity function. It uses a set (usually two) 3×3 kernels which are then to be convolved with the original image and this is done to calculate the approximations of derivatives.



C. K-means Clustering and Averaging pixels

K-means Clustering subdivides an image into multiple clusters based on the colors. A mean value is calculated for each cluster and averaging is performed by assigning the mean value to all the pixels of the cluster. The smoothening of colors is done by segmenting the image into different clusters using k-means clustering. All of the pixels that are present in each of the cluster are then replaced by the calculated k-mean value of the cluster. At the end of clustering and averaging pixels a smooth image is obtained in which colors are also restored.



D. Histogram Equalization and Adaptive Histogram Equalization

Histogram equalization is applied to enhance and improve mural image for restoring. Histogram equalization is used to enhance mural image for restoring details, colors and contrast further. A sharpening mask can also be applied to the image at the final step but it didn't give good result. The best results were given by adaptive histogram





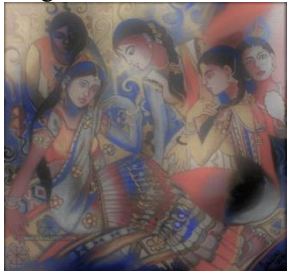







equalization. Histogram equalization is a digital image processing technique that is used for manipulating intensity of the images, it is applied for contrast enhancement. This method is useful in images with backgrounds and foreground that are both bright and dark.

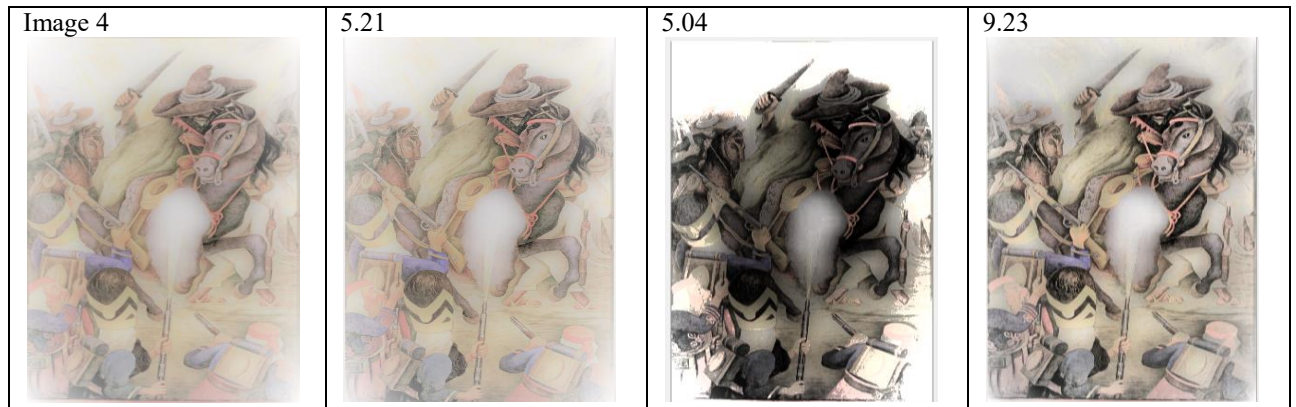
Since the evolutions of the result is measured through PSNR, the increment of PSNR value was not consistent hence Adaptive Histogram Equalization was used. The result turned out to be better i.e histogram equalization increased PSNR for 18 mural images out of total 44 mural images whereas adaptive histogram equalization increased PSNR for 35 mural images out of total 44 mural images.



III. Results

In this paper we have applied methodology to restore deteriorated mural images. The images are successfully restored using the above implemented methodology.

Mural Image	Sharpening Mask	Histogram Equalization	Adaptive Histogram Equalization
Image 1 	18.20 	17.02 	19.20 
Image 2 	15.14 	15.46 	16.21 
Image 3 	15.17 	16.07 	15.5 



IV. Conclusions

The approach improves the quality of the distorted image. The lines, colors, contrast and the structure of the image can be restored by this methodology. The image restoration success rate is concluded by the image restoration parameter PSNR (Peak Signal to Noise Ratio) adaptive histogram equalization is improving PSNR of most of the images. The adaptive histogram equalization algorithm is restoring 80% whereas conventional histogram equalization restored only 40% deteriorated images.

The implemented approach is restoring most of the deteriorated images except the images that are completely washed out, had wide cracks or missing parts.

V. References

- [1] Bhabatosh Chanda, Dhruv Ratna, B.L.S. Mounica, "Virtual Restoration of Old Mural Paintings using Patch Matching Technique", 2012 Third International Conference on Emerging Applications of Information Technology (EAIT), 978-1-4673-1827-3/12, 2012 IEEE.
- [2] Karianakis, N., & Maragos, P., "An integrated system for digital restoration of prehistoric Thera wall paintings" IEEE International Conference on Digital Signal Processing, pp. 1-6, 2013.
- [3] S. Awate and R. Whitaker, "Unsupervised, information-theoretic, adaptive image filtering for image restoration", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.-28, NO.-3, March 2006.
- [4] Rohit T. Pushpalwar, Smriti H. Bhandari, "Image Inpainting Approaches – A Review," Department of Computer Science and Engineering Walchand College of Engineering, Sangli, 2016 IEEE 6th International Conference on Advanced Computing, India, 978-1-4673-8286-1/16, 2016 IEEE..
- [5] S. C. Pei, Y. C. Zeng and C. H. Chang, Virtual Restoration of Ancient Chinese Paintings Using Color Contrast Enhancement and Lacuna Texture Synthesis, IEEE transactions on image processing, Vol. 13, pp.416429, 2004.
- [6] B. Chanda and Pulak Purkait, Digital Restoration of Damaged Mural images, The 8th Indian Conference on Vision, Graphics and Image Processing, 2012.
- [7] A. Buades, B. Coll, and J.-M. Morel, A non-local algorithm for image denoising, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 2, June 2005.
- [8] B. Chanda and D. Dutta Majumder, Digital Image Processing and Analysis, PHI Learning, New Delhi, 2011.

Text Extraction and Recognition from Images

Priti Gangania¹ and Tushar Patnaik²

School of Information Technology (SoIT)

Centre for Development of Advanced Computing (C-DAC), Noida

Uttar Pradesh, India

Abstract: Ever since Logo has been important identification for different organizations and companies, extraction of text from logo have played very crucial role. Text extraction from logo is very complex and a lot of research is going on. Sometimes the text in the logos are unreadable because of fonts, background colour, foreground colour, orientations. This paper mainly focuses on the extraction of text from logo. Since logo contains both Graphics and text, it is difficult to differentiate among text and graphics because of inconsistent background. Text have several properties like colour, size, contrast, orientation etc. Text extraction from logo helps to prevent in trademark cases. We extract relevant text from images

Keywords: Median filter, text extraction, CC (connected component), Recognition, Segmentation

I. Introduction

Logo is a symbol used by different organization to identify their products. They are commonly used by organizations and individuals to aid and promote. Color is an important key to differentiate logos. Variation of size, alignment in logos makes the automatic text extraction more challenging. Trademark images leads to infringement of text in logo if they contains similar text to the original registered logos.

Due to rapidly increase in multimedia images via mobile phones, satellite images etc. The image indexing and text extraction becomes more important to researchers.

There are various kind of logo images. Text and graphics logo: text logo is the type of logo which are comprised of a few letters, usually a company's initials. It follows simplicity of text. Graphics logo contains text with graphics or symbol of company name. Text present in logo contains useful information



a. Text logo



b. Graphics logo

Figure 1 types of logo

Median filter is used after preprocessing of image as it reduce noise on an image. It is a technique to remove noise and improve efficiency in the results. As it retain the edges, and is widely used in image processing. Edges are of critical and importance part to the visual and appearance of images, median filter gives effective result in salt and pepper noise that's why it is widely used in image processing.

The text present in images can be extracted with help of text localization and detection. There are two techniques: region based, CC (Connected component based). Region based uses the text properties such as color or other background differences. This approach mainly uses bottom up method which combine small components or group together into a single form till all text regions are correctly identified. CC labelling algorithm is commonly used because to its lower computational cost and effective implementation. Most of the CC-based methods consist of four processing stages: (a) preprocessing, which includes color clustering and noise reduction, (b) CC generation, (c) filtering out non-text components, and (d) component grouping

Text extraction is applicable to extract text data and recognition of each characters and numerals. The critical role of OCR (Optical Character Recognition) has been demonstrated and used for recognition of text from images. Due to demand of storing paper document information in form to computer storage.

A simple method to store paper document into computer readable form is to acquire digital images by camera or scan via printer. OCR is used for conversion of handwritten, scanned, typed images, a scene photo (ex: text on hospital board), subtitle superimposed on image. OCR is an instrumental technique to extract image text (text within images) that is successively indexed and available to search images .OCR is a field of research in artificial intelligence , pattern recognition, and computer vision.

Some components, such as multi-leveled equations and graphics, are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. Literature review

Rama et.al [6] introduced text extraction approach which was unaware of noise, skewness and orientation of text, intensity or color, layout from heterogeneous images. Various edge detection operations performed by using basic mathematical morphology and then connected component candidate text were found using the edges. These connected components labeled to identify various components of the image. For each connected component, variance was found considered only gray scale of components .Later text was extracted by choosing only those connected component whose variance less than threshold.

Liu et.al [7]'s handled printed/scanned text images and scene text images. He introduced edge based approach by using different variables like edge density, orientation, strength and variance as differentiate text characters embedded in image to form a feature map .Multiscale edge detector handles text extraction in this approach and morphological dilation operator is used for text localization.

III. Proposed approach

Proposed approach is shown below in figure 1

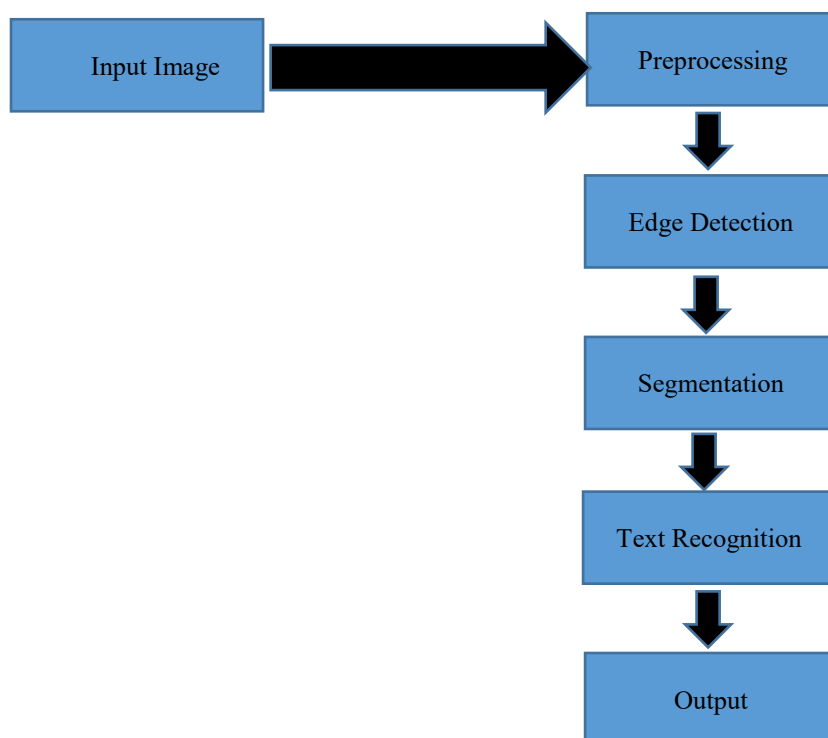


Figure 2 flowchart of proposed approach

This paper represents contribution in segmentation and text recognition for which it shows variation of color and orientation.

A. Input image

Input images are acquire through scanned documents.

Preprocessing

RGB or colored image is transformed to grayscale images and further into binary image. Median filter used to remove noise and preserve edges of an image.



Figure 3(a)Original image



Figure 3 (b) binary image

B. Edge Detection

A standard operator canny has been used to detect edges of an image. To find broad edges, it uses multi stage algorithm. It detect edges with low error rate.



Figure 4 Edge detection

C. Segmentation

Subdivide an image into multiple segments. It perform line by line, character by character segmentation. Text extraction of an image depend on the performance of segmentation.

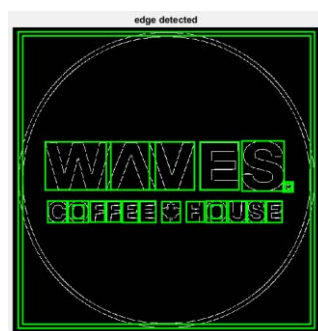


Figure 5 Segmentation



Figure 6 Extracted text

Connected Component Algorithm

Connected Component algorithm has been used to find the connected regions. Connected pixels are identified with the help of neighboring pixels. It scan image and grouping connected pixels into single connected component. It also used for blob discovery, blob extraction and region labelling.

D. Text Recognition

After segmentation, segmented character is now given as input in OCR (Optical character Recognition) for character recognition. To recognize text in logo OCR method is applied.

IV. Results

In this paper we have applied methodology to extract text from logo. We have successfully extracted and recognized text in Text Logo.



Input	Output
	Text: 'heart...'
	W1\VES@ COFFEE 4' HOUSE

Table 1 output of images

V. Conclusion

In this paper we have implemented our approach for extraction and recognition of text from logos. This approach gives better result for segmentation. Curved text in logo cannot be completely extracted. Due to different orientations of text it was difficult to extract text. For future work OCR could be improved to give better result in graphics logo images and could be able to differentiate among text and graphics. Total 80 images of data set has been considered. The experimental results the following results has been observed:

Types of images	Accuracy
Around 40 Images(consist of both text and graphics)	70-80%
Around 40 Images(mostly text logo images)	Above 80%

Table 2 result

References

- [1] Anita Pal & Dayashankar Singh, "Handwritten English Character Recognition Using Neural Network", International Journal of Computer Science & Communication, Vol. 1, No.2, July-December 2010, pp. 141-144.
- [2] M. Sundaresan and S. Ranjini, "Text extraction from digital English comic image using two blobs extraction method," *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on*, Salem, Tamilnadu, 2012, pp. 449-452.
- [3] Afritha Farhath K, Amruthavarshini R, Harshitha S, Saranya A and Velumadhavarao R, "Development of shopping assistant using extraction of text images for visually impaired," *2014 Sixth International Conference on Advanced Computing (ICoAC)*, Chennai, 2014, pp. 66-71.
- [4] G. G. Devi and C. P. Sumathi, "Text extraction from images using gamma correction method and different text extraction methods — A comparative analysis," *Information Communication and Embedded Systems (ICICES), 2014 International Conference on*, Chennai, 2014, pp. 1-5.
- [5] P. Tripathi and A. K. Indoria, "Extraction and recognition of multi-oriented text from trademark images," *Cognitive Computing and Information Processing (CCIP), 2015 International Conference on*, Noida, 2015, pp. 1-5.
- [6] Miriam Leon, Veronica Vilaplana, Antoni Gasull, Ferran Marques(2010),"Region-Based Caption Text Extraction",11th International Workshop On Image Analysis For Multimedia Interactive Services (Wiamis).
- [7] Yu Zhong, Hongjiang Zhang, And Anil K. Jain(1999),"Automatic Caption Localization In Compressed Video", International Conference On Image Processing, pp: 96 - 100 Vol.2.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Big Data Fusion: A Survey

SumedhaSeniaray¹ and Nidhi Jain²

Computer Science & Engineering

Centre for Development of Advanced Computing (C-DAC),

B-30, Sec-62, Noida-201309, INDIA

Abstract: With the rise of the Big Data era, large volumes of structured, semi-structured and unstructured data are one of the major problems, which comes in from various different heterogeneous sources. Because of different forms, these data are often considered apart from each other. If a query involved both structured data and its unstructured form of data, it is not efficient and feasible to execute it separately. Despite the widespread of diffusion of structured data, there still remains a huge amount of data in the unstructured form, be it system-generated or human-generated form of data. For such growing data resources, it is necessary to mend the gap between the varieties of knowledge domains and provide a homogeneous interface for the ease of accessing the information. Thus, this paper discusses various different technologies used to integrate data stored in remote locations.

Keywords: Big Data; Structured Data; Semi-structured Data; Unstructured Data; Data Fusion; NoSQL

I. INTRODUCTION

Big data is the current buzz among the researchers, scientists and retailers, which are trying to make use of the huge amount of data to reach to a conclusion which improves not only their decision making ability but profitability, productivity and power to peep into the past to look upon the future to make it much more good and efficient. In the process, over the years, organizations are learning that Big Data is a trending across many areas of business and technology and not only a single technology, technique or initiative. 'Big Data' describes data sets that are so enormous and complex that it is impractical to manage using conventional software tools. Big data is not all about Exabyte's or petabytes of data. When the amount of data that is needed to be processed is more than the capacity of the system, then it referred to as Big Data. For example – document of 100mb need to be attached to email but email system does not support an attachment after a particular size. So the attachment with respect to an email is referred to as big data.

Big Data phenomenon is often described using 5 V's: Volume, Velocity, Variety, Veracity and Value.

1. Volume

It refers to the large volumes of data that is produced every second. The size or magnitude of such data defines its value and determines if it can be considered as "Big" data or not.

2. Velocity

The rate at which the data is produced and the speed at which data moves around, processed to deal with the demands and challenges for the growth and development.

3. Variety

It is the different types of data that is currently available worldwide. Earlier, we only had structured data that fits into tabular form or in relational databases, but now with big data technology we can now deal with differed types of data like structured, that is, data present within the databases or csv files, semi-structured, like XML data, data present within emails, and unstructured data, such as Word, PDF, text, image, audio or video files.

4. Veracity

It is the trustworthiness or messiness of the data. Thus, the quality of acquired data can vary to a great extent, which affects the accurate analysis.

5. Value

Value is another V to consider when working with Big Data. Accessing big data is great but unless we can turn it into value it is useless.

Types of Data

There are **three types of data** that come with Big Data, structured, unstructured, and semi-structured data. Of these, the last two are new to Big Data.

1. Structured Data:

Structured data is the Data which can be stored in database SQL, that is, a table with rows and columns. They have relational key and are mapped into pre-designed fields. Currently, this data is the most worked on in development and is the most effective way to organize the information. This data is thus visualized as the data where everything is identified, labeled and easy to access. An advantage of relational database applications is the prevailing tools and web frameworks that help in the development of database-focused applications. Some types of structured data can be machine generated, such as data that comes from medical devices (heart rate, blood pressure), manufacturing sensors (rotation per minute, temperature), or web server logs (number of times a page is visited). Structured data can also be human generated, such as age, zip code, and gender.

2. **Unstructured Data:**

Nowadays, 80% of data is considered to be unstructured data. The term “unstructured” is due to the fact that there is no structure available for such data. Unstructured data can also be identified as data that cannot be stacked up in a relational database within columns and rows. It often includes text and multimedia content. Unstructured data is everywhere. Unstructured data, similar to structured data, can be machine generated or human generated. For example, unstructured data that are machine-generated are Satellite images, scientific data, and radar or sonar data etc. Human-generated unstructured data is the data such as, text documents, records, surveys, social media data, website content etc. Unstructured data is fairly detailed and is usually text heavy. Unstructured data can be found in documents, presentations, audio, images, videos, messages, and books. Unstructured data can also come from social media sites such as Facebook, LinkedIn, Twitter, Tumblr, Flickr, Yelp, YouTube, and Pinterest. If the data which is to be stored has no tags (metadata about the data) and has no established schema, ontology, or consistent organization, then it is unstructured.

3. **Semi-Structured Data:**

Data is called semi-structured data if it is not included in a relational database but which has some organizational properties that is helpful to analyse such data. Semi-structured data can be considered to be a form of structured data that like relational databases do not have a formal structure, but it contains tags (like XML extensible markup language used for documents on the web). Schema definition is not necessary for semi-structured data, but this does not mean that it is not possible to have a schema definition, it is rather optional. Therefore, it is also known as self-describing structure. Some of the data that appears to be unstructured but are actually semi-structured are, text, web server logs and search patterns and sensor data. Examples of semi-structured data are JavaScript Object Notation (JSON) and Extensible Markup Language (XML).

Table 1: Comparison of structured, unstructured and semi-structured

	Structured	Unstructured	Semi-structured
Technology	Relational database tables	Character and binary data	RDF/XML
Transaction Management	Mature transaction management, various concurrency techniques	No transaction management, No concurrency	Transaction Management taken from RDBMS
Flexibility	Schema-dependent, rigorous schema	Very flexible, Absence of schema	Flexible schema
Scalability	Scaling DB is difficult	Very scalable	Schema scaling is simple
Robustness	Very robust	-	New technology, not widely spread
Query Performance	Structured Query allows complex joins	Only textual queries possible	Queries over anonymous nodes are possible

NoSQL

NoSQL, also called Not Only SQL, has a many different database technologies that help in management of data which is useful for very large sets of distributed data present today. It helps to resolve performance and scalability issues of big data that the databases weren't meant to do. NoSQL is of great use when it comes to an enterprise that needs to access and analyse the vast amount of unstructured data or the large amount of data that is remotely stored on many different virtual servers in the cloud. Therefore, NoSQL provides a medium for storing and retrieving data which is modelled in other way than the SQL database schema. Big data and real-time web applications require NoSQL databases. Relational databases were not capable of handling the scalability and agility challenges, or take advantage of processing and storage power available today, hence, NoSQL databases helps solve this problem.

Some of the NoSQL database types are:

- **Document database:** These databases store records as documents which are actually a stock of key-value pairs, similar to a key-value store, but it provides some structure and encoding of managed data.

- **Graph stores:** These databases are widely used for interconnected data. They store information relate to networks of data, like social connections. Such networked database uses nodes and edges to store and represent data. Example of Graph stores are Neo4J and Giraph.
- **Key-value stores:** These are the most simple NoSQL databases which pairs keys to its values. Each item in this database is stored as a 'key' (or an attribute name), together with its value. Examples include Dynamo, Aerospike, OrientDB.
- **Column stores:** such as Cassandra and HBase, serializes all the values of a particular column together on disk, which speeds up the data retrieval. Thus, these are used to query large datasets, and store columns of data, instead of rows.

Why NoSQL?

In comparison to SQL databases, NoSQL databases has more scalability and provides better performance, and its data model can address various issues that the relational or SQL model is not designed to do:

- Fast changing large volumes of structured, semi-structured and unstructured data.
- Agile sprints, quick schema iteration, and frequent code pushes
- It uses Object-oriented programming that is user-friendly and flexible

Table 2: Comparison of SQL and NoSQL databases

	SQL	NoSQL
Types	Only 1 type od database (SQL or Relational DB)	Many different types like key-value stores, document DB, etc.
Data Storage Model	Individual records are stored in rows and columns in tables. Related data are stored in a single table.	It varies based on database type. Eg. Key-value stores have only 2 columns 'key' and 'value'
Schema	Structured and datatypes are fixed in advance	It's structures are typically dynamic in nature
Scaling	Vertically	Horizontally
Development Model	Mix of open-source and closed-source	Open-source

II. BIG DATA ISSUES

Few major challenges of Big Data include:

1. **Scalability:** Scalability of the data volumes and the number of sources participating.
2. **Storage insufficiency:** Where can I keep the data? Data increases day by day thus, adding petabytes and zettabytes of data which is difficult to store.
3. **Data management:** Heterogeneity of the nature of sources in terms of data model and accessing of data.
4. **Meeting the need for speed:**Getting quicker answers across large data sets. Accessing the data faster and within a shorter timeframe.
5. **Infrastructure:** How such huge amount of data can be stored. Integration with existing tools is also a difficult task due to various forms of data sets across various sources.
6. **Security:** There is almost no security virtually for all big data tools. Once you get an access, you get access to everything.

Other Big Data challenges include:

- Capturing data
- Curation
- Searching
- Sharing
- Transfer
- Analysis
- Presentation

III. DATA FUSION

The terms data fusion and data integration are typically employed as synonyms; but, the term data fusion is used for synthesizing raw data from several sources to generate more meaningful information representation and data integration refers to combining data and knowledge from various sources into a single platform.

Data fusion is the procedure which integrates variety of knowledge and data depicting the real-world object into a uniform, precise, and convenient manner. Combining relevant knowledge or information from multiple data sources into a single one in order to provide a more accurate and meaningful description or representation of data is the main goal of data fusion.

Data fusion has three broad goals: increase completeness, conciseness of data and being consistent with the data available to the users and applications. Increased completeness can be achieved by incorporating multiple data

sources into the system. By discarding redundant data, merging duplicate entries and fusing common attributes into a single attribute, increased conciseness is achieved. A consistent data consists of all the tuples from various sources that are consistent in accordance to a specified set of integrity constraints. Thus, data fusion aims at producing a complete and consistent data from various sources.

Sensor data from the Internet of Things (IoT) is one of the examples of an application that is mended together to develop an integrated view based on how complex distributed system like an oil refinery performs.

Some of the **Data Fusion Challenges** are:

1. Scalability:

Multiple resources is integrated with data from existing systems to the upcoming data is a scalability issue. To provide a high performance computing experience by collecting huge data sets at higher data streaming rate is done by using IBM mainframe technology which is integrated with big data tool. Thus, this makes it adaptable to the new technology Hadoop which delivers this high computing performance by performing the operations on huge data sets. The variety of data is a feature of Hadoop which gives a better focus on data forms like structured data, unstructured and semi structured

2. Data inconsistency:

The data levels could be inconsistent when different data formats come from different sources; therefore, for optimization of the unstructured data, more resources are required.

3. Incompatible data:

The biggest challenge for Big Data is the technical implementation of integrating data from disparate often incompatible sources. Data from such sources maybe incompatible, that is, there is “mismatch” of data types. E.g. data from database and from YouTube will be of different formats, thus being incompatible, or new video system might be incompatible with existing ones.

4. Incomplete data:

The conclusion drawn from the incomplete data will give drastic impact is one of the common problem. Missing data or incomplete data may occur due to nonresponse, that is, no information is provided for one or more items or for a whole unit.

5. External data reliability:

The origin of the data may contribute to how reliable the data is perceived in the organization. Thus, if the origin of the data is unknown it is difficult to rely on.

6. Synchronization of data across:

A consistency needs to be maintained among the data from the source to destination data storage and also vice versa. This may also jeopardize the security of the data and the sources from where the data is coming from during the process of data integration.

IV. BIG DATA FUSION STRATEGIES

Some of the relevant techniques and methods are discussed in the next section.

1. Mediator-based Data Integration

Mohamed Salah Kettouch *et al.* (2015)^[1] discusses the drawbacks of few existing tools and solutions. The author also proposes an architecture for data integration, done using a mediator and centralized global schema, along with semi-structured heterogeneous data sources with huge amount of structured data, particularly Linked Open Data cloud. But, this approach integrates structured and semi-structured data only and includes a lot of human involvement for successful implementation. Reference [2] proposes a mediator to preserve the autonomy of the participating sources. A Mediator can be defined in integration systems as a homogenous interface and bridge for clients to access a number of data sources. It commonly includes a query processor, which generally follows global schema rules, to reformulate user queries into sub queries and distribute them. Its main role arguably is to hide the complexity from the end user and facilitate the process of accessing and reading information from various sources. The layer architecture in [3] illustrates a middleware system to integrate time-dependent data, or sensor data, with the LOD cloud. It unifies and publishes streaming raw data, coming from various sources. In [4], the authors described the Mediator/Wrapper based architecture for integrating semi structured and structured data MOMIS (Mediator enviroNment for Multiple Information Sources).

2. Agent-based Data Integration

Vincenza Carchiolo *et al.* (2015)^[5] proposes an agent-based system for integrating health-related data retrieved from both structured and semi-structured sources. Structured data being the HIS and websites or social networks are the semi-structured data sources. The integration module does not specify a mapping function so that the system can be trained a-priori to associate different data for the same person. Reference [6] describes an agent-based framework which helps acquire and process distributed, heterogeneous data collected from diverse sources. [7] discusses integration approach used in ONTOFUSION which is an ontology-based system drafted for biomedical database integration which is built on two processes: mapping and unification. Yintang Dai *et al.* (2006)^[8] presents a Multi-Agent based Data Integration (MADI) framework to integrate distributed data

sources across the Internet. It unifies the data warehouse model and middle-ware model by combining them in one system.

3. Combined Index based Integration

Chunying Zhu *et al.* (2015)^[9] presents a novel index structure which helps make amalgamation of structured and unstructured data. The combined index is a joint index over structured database and unstructured document, depending on the existing entity co-occurrences. It is also known to be a semantic index which depicts the semantic relationships between entities and their multiple resources. HS3 [10] is a hybrid semantic search system that uses a combined index to grab admissible unstructured data, complementing structured data stored in a Knowledge Base without the need for two separate indexes. A joint index over text and ontologies is proposed in [11], which aims to Efficient Semantic Full-Text Search. The index consists of two kinds of lists: lists containing text postings, called context lists, and lists containing data from ontology relations, called relation lists.

4. Semantic ETL approach

Srividya K Bansal (2014)^[12] proposes a semantic Extract-Transform-Load (ETL) framework that integrates data coming from numerous sources as Open Linked Data using semantic technologies. It deals with a semantic data model which helps integrate and understand knowledge from multiple sources; RDF as the graph data model is used for a distributed Web of data; useful information and knowledge is extracted from the integrated data using the semantic query language, that is, SPARQL. [13], [14] describes a semantic approach to ETL technologies.

V. CONCLUSION

The basic objective of this paper was to discuss about the big data integration methods and techniques. It reviews various methods and techniques used by the big data researchers to gain knowledge by processing and analysing the big data. This paper also focuses on the challenges faced by researchers while performing big data integration or fusion. Many advances have not occurred for the big data fusion techniques but they are likely to develop in the near future. The development of these techniques will help us to fuse the big data more efficiently and hence producing more efficient and beneficial results.

VI. REFERENCES

- [1] Mohamed Salah Kettouch, Cristina Luca, Mike Hobbs, Arooj Fatima, "Data Integration Approach for Semi-Structured and Structured Data (Linked Data)", IEEE, 2015, Industrial Informatics (INDIN), 2015 IEEE 13th International Conference, July 2015.
- [2] P. Pagano, L. Candela and D. Castelli, "Data Interoperability," Data Science Journal, vol. 12, pp. GRDI19-GRDI25, 2013.
- [3] D. Le-Phuoc, H. Q. Nguyen-Mau, J. X. Parreira and M. Hauswirth, "A middleware framework for scalable management of linked streams," Web Semantics: Science, Services and Agents, no. 16, p. 42–51, 2012.
- [4] M. Vincini, D. Beneventano and S. Bergamaschi, "Semantic Integration of Heterogeneous Data Sources in the MOMIS Data Transformation System," Journal of Universal Computer Science, vol. 19, no. 13, pp. 1986-2012, 2013.
- [5] Vincenza Carchiolo, Alessandro Longheu, Michele Malgeri and Giuseppe Mangioni, "Multisource agent-based healthcare data gathering", IEEE, 2015, Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 1723–1729, ACSIS, Vol. 5, September 2015.
- [6] Łukasz Faber, "Agent- based Data Integration Framework", Wydawnictwa AGH Computer Science Magazine, Vol. 15(4), pp. 389-410, 2014.
- [7] D. Pérez-Rey, V. Maojo, M. García-Remesal, R. Alonso-Calvo, H. Billhardt, F. Martín-Sánchez, A. Sousa, "ONTOFUSION: Ontology-based integration of genomic and clinical databases", Computers in Biology and Medicine 36 (2006), pp. 712–730, 2006.
- [8] Yintang Dai, Shiyong Zhang, "Multi-Agent based Data Integration in Real-world", Proceedings of The Sixth IEEE International Conference on Computer and Information Technology (CIT'06), Sept. 2006.
- [9] Chunying Zhu, Qingzhong Li, Lanju Kong, Song Wei, "A Combined Index for Mixed Structured and Unstructured Data", IEEE, 2015, Web Information System and Application Conference (WISA), September 2015.
- [10] M. Gärtner, A. Rauber, and H. Berger, "Bridging structured and unstructured data via hybrid semantic search and interactive ontology-enhanced query formulation," KnowlInfSyst 41(3), pp.761-792, 2014.
- [11] H. Bast, and B. Buchhold, "An Index for Efficient Semantic Full-Text Search," CIKM, pp.369-378, 2013.
- [12] Srividya K Bansal, "Towards a Semantic Extract-Transform-Load (ETL) framework for Big Data", IEEE, 2014, International Conference on Big Data Congress, July 2014.
- [13] D. Skoutas and A. Simitsis, "Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data," International Journal on Semantic Web and Information Systems, vol. 3, no. 4, pp. 1–24, 34, 2007.
- [14] S. Bergamaschi, F. Guerra, M. Orsini, C. Sartori, and M. Vincini, "A semantic approach to ETL technologies," Data & Knowledge Engineering, vol. 70, no. 8, pp. 717–731, Aug. 2011.

VII. ACKNOWLEDGMENTS

I take this opportunity to acknowledge all those who have guided me in this project work. I express my earnest gratitude towards Ms. Nidhi Jain, my project guide for her valuable encouragement and guidance.

I would like to thank the faculty members of C-DAC, NOIDA for their constant support and guidance during my research work. Their motivation and suggestions were very valuable for my project work. I am grateful to them for their most cooperative attitude and suggestions, without which I could not have been able to do this work.

Lastly, I wish to thank my mother and father who were always there for me by giving everything they have, my brother and friends for their love and support.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Weather Prediction (Rainfall) Using Multiple linear regression along with adjusted R^2

Gunpreet Singh¹ and Kriti Saroha²

School of Information Technology (SoIT)

Centre for Development of Advanced Computing (C-DAC), Noida
Uttar Pradesh, India

Abstract: *Precipitation expectation is a vital piece of climate forecast. Contrasted with regular techniques foreseeing rainfall rate, the approach applying authentic records and data mining innovation indicates clear advantage in processing cost. Precipitation forecast issue has been one of the significant issues in the area of securing water security. Exact rainfall forecast can be effectively put to use by the agro based economy nations regarding long term prediction.*

Keywords: *-Multiple linear regression, Data mining, Monsoon, p-test and Rainfall.*

I. Introduction

In agriculture based countries like India, exclusive water management is essential for providing water for farming activities and for efficient flood prevention and drought prevention. One of the major problems encountered during water management is to accurately predict the rainfall. Accurate rainfall prediction will provide accurate and timely management of crop selection for farmers, road safety at high mountains etc. Rainfall prediction depend on many complex factors such as wind speed humidity, geographical locations etc.

A weather condition is a natural process that takes place because of various atmospheric factors and which have an adverse effect on environment. Water received due to rainfall can provide solution for dryness, shortage of water for farmer's use and increase in amount of reserves other than groundwater. This in turn would help in decreasing the death rate of plants that occurs due to shortage of water. On the other hand, the other effects brought due to high rainfall are hazards such as floods and landslides. Considering rainfall as an important source of water supply, efficient rainfall prediction model is needed to avoid such disasters.

There are different types of prediction which may be classified as:-

- Large:-Forecasting 14 or more days in advance.
- Medium:-Forecasting 5 to 7 days in advance.
- Short:-Forecasting 1 to 3 days in advance.

The rest of the paper is organized as follows: section 2 describes the related work in the area. Section 3 describes the approach used. Section 4 summarizes the results and conclusions.

II. Literature Survey

Various studies have been done on the different aspects of prediction of rainfall. Studies have been undertaken to analyze data locally as well as globally, with results frequently varying depending on the geographical area, altitude, pressure etc. This section briefly describes and compares the techniques used by different researchers with their conclusions. Some of them are described in this section.

Chowdhari K.K et.al collected data from Indian metrological department, Pune. Authors presented the work on novel algorithms and explored different mining techniques to study about weather and climatic changes.

In this paper, the authors have initially transformed all the data into uniform format. The authors have then applied clustering technique and analyzed the clusters, thus obtained. The authors performed cluster analysis, grouped and analyzed data in the clusters as follows,

Cluster 1:- Regions discovered with precipitation in the scope of 0.1 to 3mm.

Cluster 2:- Regions discovered with precipitation in the scope of 3mm to 5mm.

Cluster 3:- Regions discovered with precipitation in the scope of 5mm to 10mm.

Cluster 4:- Regions discovered with precipitation in the scope of 10mm to 25mm.

Cluster 5:- Regions discovered with precipitation more prominent than 25mm.

The groups consequently show the precipitation received on that day.[1]

Pinky Saikia Dutta *et-al*, proposed an approach using Multiple linear regression on the dataset obtained from the Economical Statistical Department of Guwahati and Assam. The Dataset consists of data ranging from year 2007-2012 with eight attributes. The authors first performed the p-test to find the necessary attributes required for prediction. In the p-test, relative humidity attribute failed the test and hence it was discarded. The authors performed multiple linear regression on the reduced attributes and predicted the rainfall for the month of October with an accuracy of 63%.

The authors have also suggested that if some additional attributes are also considered, then there are chances to obtain the results with better accuracy.[2]

A.Geetha, *et-al*, have proposed an approach using Decision trees and collected the dataset from ncdc.noaa.gov. The dataset consists of 20 attributes, but authors interviewed the scientists to gain knowledge about the relevant attributes and thus the total count of attributes was reduced to 12. The authors trained the system with dataset of year 2013 and used decision tree on dataset of year 2014 to predict the rainfall. The accuracy thus obtained was 80.67%.[3]

Valmik B Nikam, *et-al* have proposed method based on Bayesian approach. The authors collected the data from the Indian Metrological Department, Pune. The dataset consists of 36 attributes, but the authors ignored the less relevant attributes and selected only seven attributes. The authors first preprocessed the data and then applied Bayesian approach. The accuracy of the results obtained on sample data was 81.66%.[4]

Anif Hanifa, *et-al* have proposed an approach using adaptive neural fuzzy inference system on dataset collected from Meteorology Climatology and Geophysics Agency Jakarta. The authors have used three attributes which are air temperature, air humidity and wind speed for predicting rainfall. Authors have first defined the fuzzy and membership function and then defined the rule base for prediction. The authors concluded with qualitative ability to predict for the category 0(rain) and 1(no rain).[5]

The work in the area is extended further in this paper by considering some more attributes and applying Multiple linear regression to improve the results.

III. Proposed approach

This section presents the proposed approach that would be used to mine rainfall dataset. The approach includes applying the Multiple linear regression on the dataset containing one more attributes, namely wind direction in addition to the attributes used in [2]. Fig 1 Shows the approach.

The accuracy thus obtained is compared with the other models. It is evident from the results obtained that the accuracy is improved with the inclusion of wind direction in the dataset.

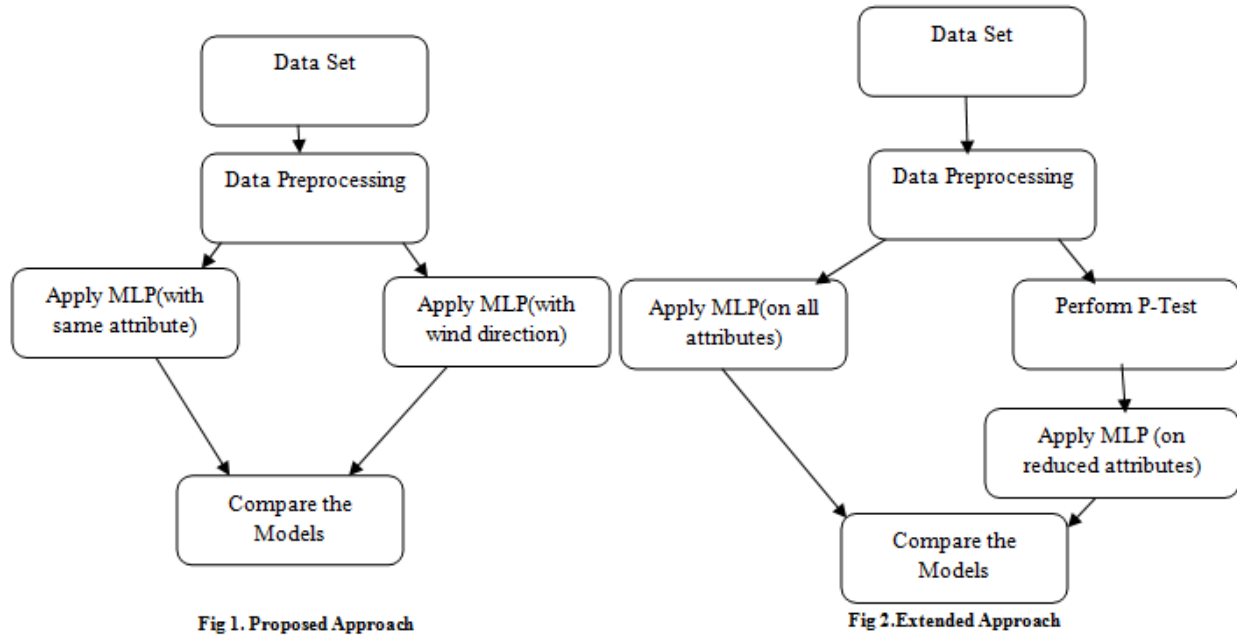
The accuracy obtained with inclusion of wind direction is 75% which is better as compared to 63% obtained earlier in[2].

The above approach is further extended by considering some more attributes and selecting the relevant ones. Multiple linear regression is applied on the dataset containing 13 attributes and accuracy thus obtained is compared with the models proposed by others. Table 3 shows the comparisons.

Further, attribute reduction is done using p-test to find the minimal set of relevant attributes to be used for prediction. Multiple linear regression is then applied to reduced set of attributes and accuracy thus obtained is compared with the models proposed by authors. Fig 2 shows extended approach.

The accuracy obtained when all attributes are considered is 75%, and with reduced set of attributes is also 75%.

It is evident from the Multiple R and R^2 values as shown in Fig. 3(a), that maximum correlation between the attributes is obtained, when we consider all the attributes for prediction of rainfall.



IV. Results

Implementation is done using MLR and it is found that the model explains 75% of rainfall prediction. Fig. 3(a) explains the correlation between attributes, values of R^2 , errors in models.

<i>Regression Statistics</i>	
Multiple R	0.7587
R Square	0.575626
Adjusted R Square	0.55194
Standard Error	0.245521
Observations	228

Fig. 3(a)

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	12	17.57955	1.464963	24.30237	8.63E-34
Residual	215	12.96034	0.060281		
Total	227	30.53989			

Fig. 3(b)

Fig. 3(b) shows the value F-test is 8.63E-34 which means that the variables used are sufficient for prediction.

In Fig. 3(b) the various values have the following meanings:

Where df stands for degree of freedom

SS stands for sum of squares

MS stands for mean of square

F stands for f-statistics

Significance of f tells us the significance of the model

The performance in terms of accuracy when monthly rainfall is predicted for one complete year is shown in Table 1 as follows:-

Case 1:-MLP applied on attributes used in[2].

Case 2:-MLP applied with addition of one more attribute (wind direction) in case 1.

Case 3:-MLP applied on all 13 attributes of the dataset used in this paper.

Case 4:-MLP applied on the reduced set of attributes obtained after applying p-test.

Accuracy obtained (in different models)

Cases	Accuracy
Case 1	58.6%
Case 2	75%
Case 3	75%
Case 4	75%

Table 1

Table 1 shows the accuracies obtained in various models/cases.

Rainfall predicted (Predicted rainfall vs Actual Rainfall)

Cases	Predicted rainfall(inches)	Actual rainfall(inches)
Case 1	7.9	12.73
Case 2	9.213	12.73
Case 3	12.48	12.73
Case 4	9.427	12.73

Table 2

Case 1:-MLP applied on attributes used in[2].

Case 2:-MLP applied with addition of one more attribute (wind direction) in case 1.

Case 3:-MLP applied on all 13 attributes of the dataset used in this paper.

Case 4:-MLP applied on the reduced set of attributes obtained after applying p-test.

Table 2 shows the amount of rainfall predicted when different cases/models are considered

It is evident from the values given in Table 2 that the prediction in case 3(MLP applied on all 13 attributes of the dataset used in this paper) is closest to actual rainfall.

Accuracy of 75% is achieved in the prediction of rainfall when all 13 attributes in the dataset are used for prediction, and maximum correlation is also found between attributes in case when all attributes are used for computation.

Total Number of Attributes	Attributes which failed p-test	Reduced Set
Year	Month	Year
Month	Maximum Temperature	Dew Point
Maximum Temperature	Minimum Temperature	Humidity
Minimum Temperature	Pressure	Visibility
Wind Speed		Max Gust
Wind Direction		Cloud Cover
Max Gust		Wind speed
Cloud cover		Wind Direction
Dew Point		
Humidity		
Visibility		
Pressure		

Table 3

After normalization of data is done, p-test is applied for ranking of attributes according to their selective class as shown in Table 3. In the p-test, the attributes:-month, maximum temperature, minimum temperature and pressure failed the test and hence they were discarded from the dataset. Table 3 shows the reduced set of attributes after p-test was applied.

V. Conclusion and Future work

The proposed model considers max gust, cloud cover, visibility, dew point, humidity, wind speed, Mean sea level as predictors for predicting rainfall for a region. 75% accuracy is achieved in variation of rainfall for the proposed model. The model can predict monthly rainfall. Some more predictors like wet day frequency, vapour pressure which are not included in the current dataset due to constraints on data collection can also be included and may give

more accurate results. The resulted rainfall amounts are focused to help farmers in making decision regarding their crop selection.

References

- [1]. Chowdhari K.K ,Girisha R ,K C Gouda “A Study Of Rainfall over India Using Data Mining” International Conference on Emerging Research in Electronics, Computer Science and Technology-2015.
- [2]. Pinky Saikia Dutta ,Hitesh Tahbilder “Prediction Of Rainfall Using Data mining Technique Over Assam” Indian Journal of Computer Science and Engineering-2014.
- [3]. A.Geetha ,G.M.Nasira“Data Mining for Meteorological Applications: Decision Trees for Modeling Rainfall Prediction” IEEE International Conference on Computational Intelligence and Computing Research-2014.
- [4].Valmik B Nikam, B.B.Meshram “Modeling Rainfall Prediction using Data Mining Method A Bayesian Approach” Fifth International Conference on Computational Intelligence, Modelling and Simulation-2013.
- [5].Anif Hanifa Setyaningrum, Praditya Megananda Swarinata”Weather prediction application based on ANFIS(Adaptive neural Fuzzy Inference system) Method in West Jakarta Region”- International Conference on Cyber and IT service management-2014.
- [6].http://en.wikipedia.org/wiki/Weather_Forecasting.

A Brief Survey on Detection of Wormhole Attack in MANET

Jyoti shokhanda¹ and Rekha saraswat²

School of information technology (SoIT)

Center for Development of Advanced Computing (C-DAC), Noida

B-30, Sector-62 Noida Uttar Pradesh INDIA

Abstract: MANET is a self arrangement, self formed and self established network of wireless movable nodes without having any fixed layout. In MANET there is no central point of coordination so MANET are vulnerable to many security attack and Wormhole attack is most severe security attack in which two or more than two malicious nodes attracts the traffic of network and transmit this traffic across the malicious nodes. Wormhole Attack disrupts the routing mechanisms completely by dropping packets, changing or stealing information. Wormhole Attack can be easy to implement but very difficult to identify in the network. It violates all the cryptographic mechanisms such as authentication confidentiality, availability and integrity. This results due to the cooperative and trusted nature of nodes in the MANET.

Keywords: MANET, Wormhole, RREQ, RREP, AODV

I. Introduction

A Mobile Ad hoc networks is a collection of movable nodes (PDAs, laptops, sensors etc) interfacing without coordination of any centralized system. The movable nodes which are in transmission area of each other are free to directly communicate and nodes that are not under transmission area with each other interface through the in-between nodes. Each node in MANET behaves as router or host simultaneously to forward data due to the lack of any centralized monitoring system. Movable nodes can join and leave the network which causes the dynamic topology. The nodes are cooperative in nature for data forwarding processes. Mobile Ad hoc networks are completely distributed and implemented at the place where setting up an infrastructure is not possible examples military services areas, disaster areas, hill stations etc.

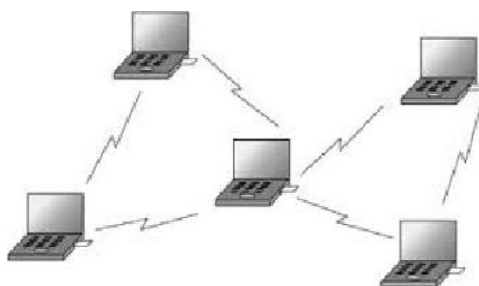
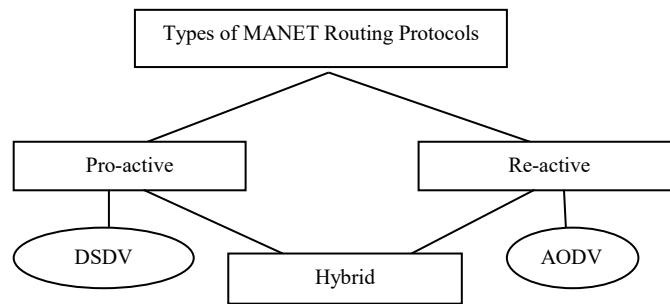


Figure: Mobile Ad hoc Network

- 1). Dynamic layout: Dynamic topology means nodes are free to relocate in any direction with different velocity. Due to this network topology may change at any point of time. These movable nodes are dynamically arranged in the network. They can join or leave the network.
- 2). Framework less: MANET doesn't depend on any pre established architecture or centralized coordination system. Every node arranged in apportioned associate mode act as independent router system and host system in which they send or received data packets simultaneously. This is on the spot created network and does not have any base station.
- 3). Multi hopping: Wireless network in which two or more hops transmit information from source to destination. In multi hopping many paths are possible between source and destination called multi path hopping. Multi hopping has multiple paths instead of one.
- 4). Scalability: Scalability is the feature of the Mobile Ad hoc networks in which it involved large coverage area with thousand of nodes so scalability is critical to the successful set up of these networks. Scalability is the capability of the MANET to handle large network area.

III. Types of Routing Protocols in MANET

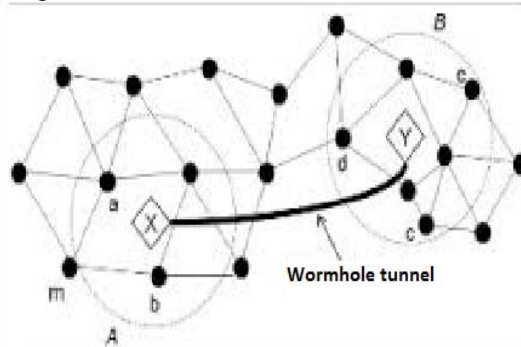


There are three types of routing protocol in MANETs.

1. Pro-active routing protocol: In proactive pathing protocol every node drives a table which contains details to its one hop neighbors even without requiring it. This is called table driven routing protocols. Examples: DSDV, OSLR, WRP etc.
2. Re-active routing protocol: In this category of routing protocols a route is established whenever required for communication process that's why Re-active routing protocols are called Ad hoc on appeal protocols. Each node maintains a route cache table which involves in the route discovery from source node to destination node. Examples: DSR, AODV TORA etc
3. Hybrid routing protocol: Hybrid routing protocol is the consolidation of pro-active and re-active protocols. It has some features of proactive and reactive routing protocols examples ZRP, ZHLS, TORA. TORA is a reactive protocol with some proactive features where a path established between source and destination through directed acyclic graph.

IV. Wormhole Attack

Wormhole attack can be introduced by intruder without understanding the structure of network or harming any authentic nodes. In this Attack the intruder puts the malicious node at some powerful transmission position in the network so that it attracts maximum traffic to pass through the wormhole node. Wormhole nodes create a fake path which is shorter than original one inside the network which is called "tunneling". This confuses the routing system to follow that path because it is shorter towards the destination. A wormhole attack can steal, change or drop the data packets or passes to other wormhole node at some distant position.



In the given Figure X and Y are the two malicious nodes which are connected via a link called tunnel. This tunnel is high bandwidth wireless channel or any wired medium which is used to transfer data through the colluding nodes. The tunnel can be composed by the in-band and out of band medium or through the immense communication energy. In the figure shown above, 'a' is the source node which records packets from the neighboring nodes transmit it to the destination 'd'. whereas 'X' is malicious node which show a shorter route to source towards destination and starts recording packets from the source and transmits these packets to the other malicious node 'Y' which is nearer to the destination.

V. Types of wormhole attack based on the visibility of attacker

1. Exposed attack: In the exposed attack, the colluding nodes put their identity into the RREQ packet as a legitimate node. Hence other nodes are aware about the availability of malicious node present in the network, but they do not identify wormhole nodes. In exposed attack the colluding node does not alter contents of the data but they share between the malicious nodes.
2. Hidden attack: In the Hidden attack, the colluding node does not put their identity into the RREQ data packet. So the legitimate nodes do not know the malicious attackers. Hidden attack is hard to detect in the network because it does not reveal their identity in the network.

VI. Related Work

1. **Ning song and Xiangfang Gi:** In this paper the authors proposed a detection technique on statistical analysis to identify wormhole attack and to identify malicious nodes. A Statistical analysis method is a detection technique to detect routing inconsistency as we have sufficient information about paths for multi path routing.
2. **Yih-Chun Hu:** In this paper a detection technique is proposed called packet leashes for identifying and violating against wormhole attack. There are two kinds of packets leashes: geographical and temporal leashes. A geographical leashes assure that receiver of the packet is within a limited distance from the sender. And Temporal leashes assure the packet have an upper bound time to pass from source to destination.
3. **H.Sun.Chiu and KingShan.L** gives a method called DELPHI by recognize the lag from different route to the destination. This method does not require any timer clock and hardware system.
4. **Mohammad R.A, King Su.Ch. :** gave a new method called RTT-TC, which is based on the time to send a packet and the time at which acknowledgement is received and topological comparisons. This method first relies on RTT calculations to identify malicious node and then compare to topological comparison to remove the authentic neighbor from malicious one.
5. **Ronggong Song Ming Li:** In this paper authors developed a technique which is based on signal processing technique. This method observes the traffic which passes towards the receiver node. The destination node is analyzed by converting the received time into signal and converting this signal into the frequency domain using the Fast Fourier transform.
6. **Sun Ch., Doo Kim, Do. Lee, Jae J:** In this paper an algorithm is proposed, called WAP algorithm without any special hardware. This algorithm uses a method called one hop neighbor node controlling method in which every node monitors its one hop neighbor node to detect the wormhole attack.
7. **Juhi Biswas, Ajay Gupta, Dayashankar Singh** implemented an algorithm called WADP which is implemented using AODV, in which node validation method is used to expose the colluding node and remove the false positive problem as exposed attack is launched in the network. This algorithm is basically improvement to the WAP which suffered from the false positive problem.
8. **Piyush Kaneria, Dr. Anand Rajavat** proposed a trust pathing protocol based on AODV routing algorithm in which trust value is calculated using trust function to detect the malicious node based on reliability.

VII. Comparison and details

Technique	Detection	Limitation
Packet leashes : geographical and temporal packet leashes	This method used the loosely synchronized clock to detect malicious nodes using GPS	Equipped to hardware.
SAM: statistical analysis method	statistical calculation of the paths between the source and the destination	Multi hop path protocols in this method are not supported.
Delphi: Delay per hop	This method observes the delay between the nodes through different paths between source and the destination.	Cannot detect the exact position of the malicious node.
RTT-TC: round trip time with topological comparison	This method first depends on the rounded run time mechanisms to detect the malicious node and the use time topological comparisons to remove authentic node from the suspected list.	Accuracy is not sufficient.
WAP: Wormhole attack prevention	Node monitoring method is used to detect malicious node exposed attack is also detected	false positive problem
WADP: Wormhole attack detection and prevention	Removes false positive problem	Not able to detect Hidden attack
Trust based AODV approach	To detect malicious node on the tangent based function	Modified in AODV

VIII. Conclusion and future work

In this paper, different detection techniques for wormhole attack in wireless network are analyzed. We discussed wormhole and its types and how it can be launched in many ways. Various techniques to detect and prevent wormhole attack are also explained. Wormhole attack is hard to detect because it affects the network without knowing the cryptographic techniques used in implementation. Most of the techniques work on identifying wormhole nodes within the networks. An efficient intrusion detection and prevention can be performed by properly authenticating the incoming nodes. If nodes are authenticated before joining the network, then detection as well as prevention may become more effective and network may work with less risk and more efficiently.

IX. Reference

- [1] Yih-chun Hu. "Wormhole Attacks in Wireless Networks" IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, VOL. 24, NO. 2, FEBRUARY 2006
- [2] Hon Sun Chiu and King-Shan Lui "DELPHI: Wormhole Detection Mechanism for Ad Hoc Wireless Networks "IEEE 2006.

- [3] Ronggong Song. "Enhancement of frequency based wormhole attack detection", 2011 - MILCOM 2011 Military Communications Conference, 11/2011.
- [4] Juhi Biswas, Ajay Gupta, Dayashankar Singh "WADP: A Wormhole Attack Detection And prevention Technique in MANET using Modified AODV routing Protocol" 9th IEEE International conference on Industrial and Information system (ICIIS), IEEE 2014.
- [5] Mehdi enshaei, Dr. Zurina Bt hanpai, "A review on wormhole attack in MANET" Journal of theoretical and Applied information technology (JTAIT) Vol.79.no1, 2015
- [6] Sminesh C N, Anuj J, "An Improved Clustering-based Approach for Wormhole Attack Detection in MANET" 2014 3rd International Conference on Eco-friendly Computing and Communication Systems, IEEE 2014.
- [7] Piyush Kaneria, Anand Rajavat. "Detecting and avoiding of worm hole attack on MANET using trusted AODV routing algorithm", 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 2016
- [8] Mohammad Rafiqul Alam, King Sun Chan. "RTT-TC: A topological comparison based method to detect wormhole attacks in MANET", 2010 IEEE 12th International Conference on Communication Technology, 2010.
- [9] Jae-il Jung. "WAP: Wormhole Attack Prevention Algorithm in Mobile Ad Hoc Networks", 2008 IEEE International Conference on Sensor Networks Ubiquitous and Trustworthy Computing.
- [10] D.B. Johnson. "Packet leashes: a defense against wormhole attacks in wireless networks", IEEE INFOCOM 2003 Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE 2003.
- [11] Chan, King Sun, and Mohammad Rafiqul Alam. "TCBWD: Topological comparison-based Byzantine wormhole detection for MANET", 2011 IEEE 7th International Conference on Wireless and Mobile Computing Networking and Communications (WiMob), 2011.
- [12] Aarti, Dr. S.S tyagi "Study of MANET: Characteristics, Challenges, Application and Security Attacks" IJARCSSE, 2013.
- [13] Akansha shrivastava and Rajni Dubey "Wormhole attack in mobile Ad hoc Network: A survey" International Journal of security and its Applications (IJSIA) vol.9 no.7, 2015



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Efficient Detection of Black Hole in Mobile Adhoc Networks

Deepak Sharma and Munish Kumar
Computer Science Department
C-DAC, Noida, India

Abstract: Mobile Adhoc networks are unstable and vulnerable to attacks so their security issues are more critical to design and implement than for the wired networks. Here we proposed a trust based approach to solve the black hole problem in Mobile ADHOC network, in which all the nodes store the trust value of neighbouring nodes and maintain them in a table. Work done so far using a trust based approach is worthy having a problem of identifying friendly node as black hole during congestion. In this paper we claim to eradicate this problem by analysing the average packet dropped. This paper will provide a thorough knowledge of efficiently detecting a Black Hole in MANET.

Keywords: Mobile Adhoc Network (MANET), AODV Protocol, Black Hole, Promiscuous Mode, Congestion

I. Introduction

MANET consist of many self-arrange networks of moving nodes. These movable nodes communicates without any infrastructure. Mobile adhoc network is largely used in military function, disaster areas, small personal network (PAN). There are large number of affairs about MANET just as security problem, precise transmission bandwidth, broadcasting messages, dynamic link establishment.

In recent years there are large number of security affair studied by the researchers. Some of them are snooping intrusion, wormhole intrusion, black hole intrusion, routing table overflow, denial of service (DoS), distributed DoS intrusions, etc. The utmost popularized security hazard such as black hole attack. Some researchers propose their secure ideas to solve this very important affair, but the security problem is still unable to eradicate completely.

MANET has some very important affair. Some of them are IP addressing, routing protocols, security, mobility management affairs, QOS. In all of these research affair, security has been a peak affair for analysts.^[1]

Due to lack of any central agency in MANET, the node are only responsible for routing and hence works as router in transmitting the data.^[1]

AODV protocol is a reactive routing protocol which is used for routing in mobile adhoc network. This protocol helps in finding route from a source to destination using a cycle of route request packet (RREQ) and Route reply packets (RREP). A route discovery process starts whenever a node wants to communicate with the other.

This process consists of a series of multiple RREQ (Route request packet and RREP packets. Source node also called as agent floods RREQ to all its neighbors, nodes having valid route to the destination on receiving a RREQ packet replies with a Route reply packet which shows that that node have a route which leads to destination. After discovering a route to destination data transmission starts from that route.

Blackhole and Grayhole attacks fall in the class of most common attacks in DOS attack category. In Blackhole attack, intruder node claims having a valid route to the destination by sending fake RREP packets and after admitted to the data path purposely drops the data packets.

In this paper we are going to substantiate the following claims:

- Provide a relatively good scheme that analyzes the behavior of nodes to eradicate the black hole node problem.
- Completely eradicate the problem of identifying friendly node as black hole during congestion using a trust based protocol.
- Solve the black hole problem with trivial compromises.

II. Proposed Approach

IDEA BEHIND THE APPROACH

This work proposed a trust based protocol to solve the black hole detection in mobile ADHOC network, in which all the nodes store the trust value of neighboring nodes and maintain them in trust table.

There are two types of trust values corresponding to every neighboring node:

1. Absolute Trust (based on node's behaviour)
2. Indirect Trust (based on node's behaviour in comparison to other neighbouring nodes)

Nodes are divided among three classes on the basis of their trust values T ($0 < a < b < \infty$):

1. Highly Trusted Node (Trust values (T) lies from $b < T \leq \infty$)
2. Moderate Node (Trust values (T) lies from $a < T \leq b$)
3. Risky Node (Trust values(T) lies from $0 < T \leq a$)

WORKING PRINCIPLE

- Trust values are calculated on the basis of node's behaviour in the network and network performance.
- In initialization every node assigns a maximum value of trust for all its adjoining nodes.
- Every time when a mobile node M send a data packet it sets a timer of N second and hear the wireless link.
- After expiration of timer M node verifies that the same data packet is received from the node to which it is forwarded ($M+1$). This is done by hearing to the medium in promiscuous mode.
- In case if the node has not heard the packet then node M reduced the trust value for the node ($M+1$).
- Sender selects the Route on the basis of Total Trust values of the replying node.
- Here the nodes have different classes Sender identifies the class of the node to which it belongs on the basis of the trust value.
- Select the most reliable node, if two nodes belong to same class then select any one randomly.
- If any node's trust value reaches a threshold (0) then node considered as black hole node and further communication block from that node.

CALCULATION OF TRUST VALUES

Absolute Trust

Value of absolute trust is calculated on the basis of node's behavior in the network. Following parameters are considered:

- (1) Packet Successfully delivered (P_{SD})
- (2) Packet Dropped (P_{DP})

Indirect Trust

Value of indirect trust is calculated on the basis of node's behavior in comparison to other neighboring nodes. Following parameters are considered:

- (1) Average Packet Dropped (PD_{AVG})
- (2) Packet Dropped (P_{DP})

Total Trust

Sum of both the Absolute Trust and the indirect trust is stored as total trust.

FORMULAS FOR CALCULATIONS OF TRUST VALUES

Absolute Trust

$$AT = AT + \{P_{SD} \times (0.01) - P_{DP} \times (0.1)\}$$

Indirect Trust

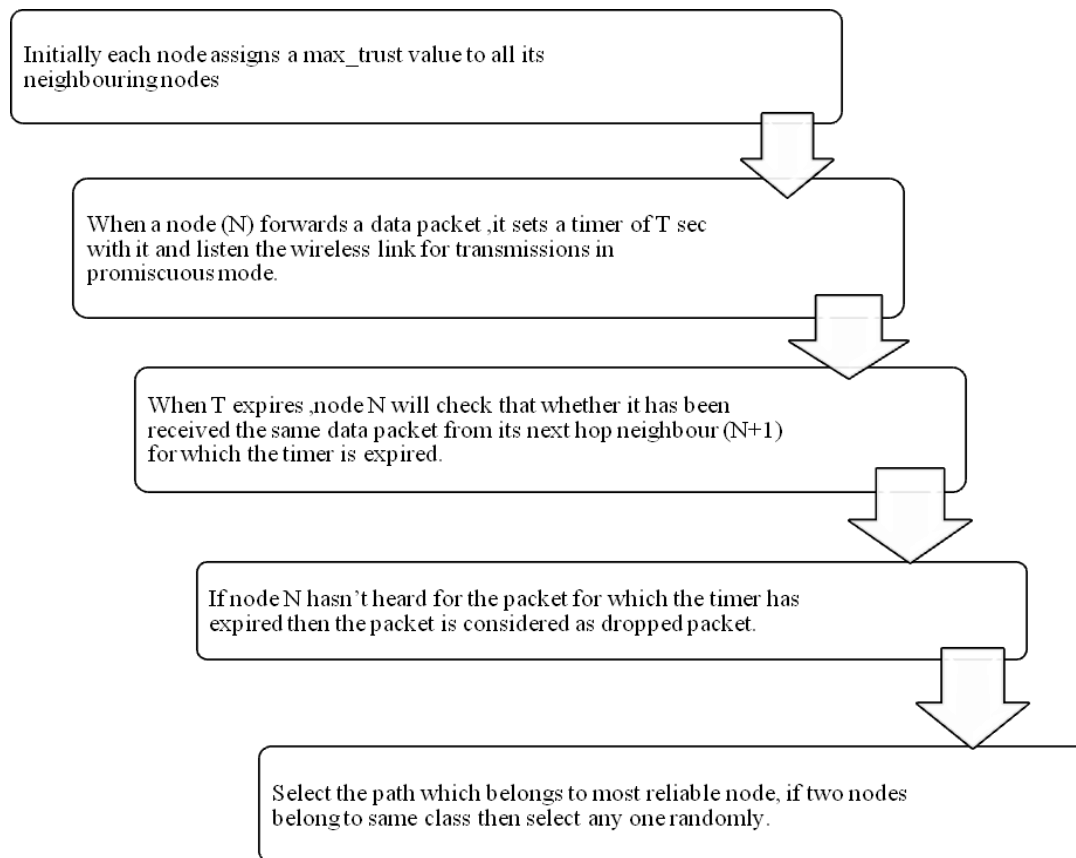
$$IT = IT + \{PD_{AVG} - P_{DP}\} \times 0.1$$

Total Trust

$$\text{Total trust} = AT + IT$$

FLOW GRAPH

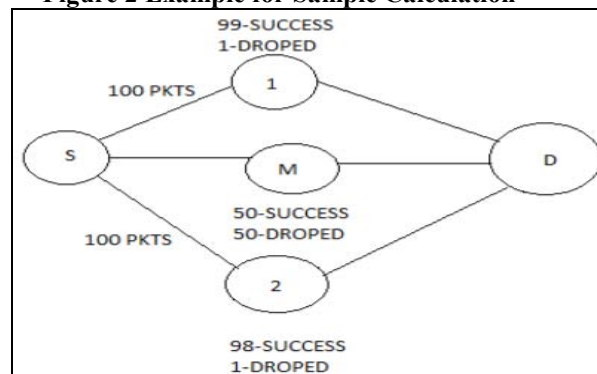
Figure 1 flow graph



Sample calculations:

Sample calculations are shown in the table for network represented in figure 1. Take initial values of Absolute and Relative trust as

Figure 2 Example for Sample Calculation



Calculation For Node -1	Calculation For Node -2	Calculation For Node -M
$AT = AT + \{P_{SD}x(0.01) - P_{DP}x(0.1)\}$ $AT=1.8$ $IT = IT + \{PD_{AVG} - P_{DP}\}x0.1$ $IT=2.66$ $Total\ trust = AT + IT$ TOTAL TRUST=4.46	$AT = AT + \{P_{SD}x(0.01) - P_{DP}x(0.1)\}$ $AT=1.78$ $IT = IT + \{PD_{AVG} - P_{DP}\}x0.1$ $IT=2.56$ $Total\ trust = AT + IT$ TOTAL TRUST=4.346	$AT = AT + \{P_{SD}x(0.01) - P_{DP}x(0.1)\}$ $AT= -3.5$ $IT = IT + \{PD_{AVG} - P_{DP}\}x0.1$ $IT= -2.24$ $Total\ trust = AT + IT$ TOTAL TRUST= -5.74

Table I Sample calculations

III. Implementation

SIMULATION ENVIRONMENT

Operating System: Ubuntu 14.04

Simulator: Network Simulator-2 (version 2.35)

Table 2 Simulation Model

Simulator	Network Simulator 2.35
Transmission Range of Node	200m
Duration	40seconds
Mac Type	IEEE 802.11g
Number of nodes	25
Antenna Type	Omni Antenna
Propagation Type	Two Ray Ground
Network Area	2000 m X 2000 m
Routing Protocol	AODV, DS_AODV
Transport Agent	UDP
Packet Format	CBR
Number of Blackhole nodes	3

STEP FOR IMPLEMENTATION OF PROPOSED APPROACH

- Formulating Simulation Model
- Understanding Structure of AODV Protocol
- Traffic Generation
- Inception of Blackhole node in Network
- Add and maintain Trust Table at each node for storing and maintaining Trust values.
- Activate Promiscuous mode on each node.

IV. Results and Analysis

QUANTITATIVE METRICS

- *Packet Delivery Ratio*: It is the proportion between the packet received and the generated packets.
- *Throughput*: Numerical measures of performance. The measures of routing policy effectiveness how well it does its job.
- *Route Acquisition Time*: The time taken to successfully discover a route from the time it is requested.

Table 3 Comparative Analysis

Metrics	AODV	DS_AODV
Packet Delivery Ratio (PDR)	22.2256 %	60.5157 %
Throughput	0.0341304 Mbps	0.0613186 Mbps
Acquisition Time	0 Second	0 Second

V. Conclusion and Future Work

CONCLUSION

This work proposed a trust based protocol to solve the black hole detection problem in mobile ADHOC network, in which all the nodes store the trust value of neighboring nodes and maintain them in the trust table. This will help to eradicate the black hole problem in MANET. According to this approach, we claim to find the black hole in the network sooner.

FUTURE WORK

- Calculate Results for different scenario by varying
 - Communication Range
 - Trust Threshold
- Compare with AODV Protocol for analysis for following parameters
 - Number of Packet Dropped
 - End to end Delay

REFERENCES

- [1] Rutvij H. Jhaveri, Sankita J. Patel, Devesh C. Jinwala, "DoS attacks in Mobile Adhoc Networks: A Survey", Second International Conference on Advanced Computing & Communication Technologies, IEEE, 2012.
- [2] Fidel Thachil, K.C. Shet, "A trust based approach for AODV protocol to mitigate Blackhole attack in MANET", International Conference on Computing Sciences, IEEE, 2012.
- [3] Nital Mistry, Devesh C. Jinwala and MukeshZaveri, "Improving AODV Protocol against Blackhole Attacks", International Multiconference of Engineers and Computer Scientists 2010, vol. 2, March 2010.
- [4] Vishnu K, Amos J. Paul, "Detection and Removal of Cooperative Black/Gray hole attack in mobile adhoc networks." , International Journal of Computer Applications, Vol. 1, No. 22, 2010.
- [5] SoufieneDjahel, FaridNa"Ot-abdesselam, and Zonghua Zhang, "Mitigating Packet Dropping Problem in Mobile Ad Hoc Networks: Proposals and Challenges", IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 13, NO. 4, FOURTH QUARTER 2011.
- [6] Tamilselvan L, Sankaranarayanan V, Prevention of Black hole Attack in MANET. Paper presented at the 2nd International Conference on Wireless Broadband and Ultra Wideband Communications, Sydney, Australia, 27-30 August, 2007, IEEE
- [7] NidhiChoudhary, Dr.LokeshTharani, "Preventing Black Hole Attack in AODV using Timer-Based Detection Mechanism" SPACES-2015, Dept. of ECE, K L UNIVERSITY, IEEE, 2015
- [8] Al-Shurman M, Yoo S-M, Park S, Black Hole Attack in Mobile Ad Hoc Networks. Paper presented at the 42nd Annual ACM Southeast Regional Conference (ACM-SE'42), Huntsville, Alabama, 2-3 April 2004
- [9] Anand A. Aware,Kiran Bhandari, "Prevention of Black hole Attack on AODV in MANET using hash function" .Published in: Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 3rd International Conference on 8-10 Oct. 2014,IEEE
- [10] Apurva Jain, UrmilaPrajapati, PiyushChouhan, "Trust Based Mechanism with AODV Protocol for Prevention of Black-Hole Attack in MANET Scenario". Published in:Colossal Data Analysis and Networking (CDAN), Symposium on18-19 March 2016, IEEE.
- [11] T. Clausen, J. Dean, C Adjih, "Generalized Mobile Adhoc Network Packet/Message Format", RFC-5444, July 2015.
- [12] Aarti,Dr. S.S Tyagi "Study of MANET: Characteristics, Challenges, Applicationand Security Attacks", International Journal of Advanced Research in ComputerScience and Software Engineering, Volume 3, Issue 5, ISSN: 2277 128X May 2013

Emotion Detection through Facial Expressions

Sanika Singh¹, Tushar Patnaik²

School of Information Technology (SoIT)

Centre for Development of Advanced Computing (C-DAC), Noida
Uttar Pradesh, India

Abstract: Lot of research work is done in the human computer interaction. To make computers more intelligent, so that it can understand human emotions too, more naturally as human can understand of another human. Since human-human interaction is very natural, so to make human computer interaction also natural many researches are going on in this field. Images of different persons posing different expressions have been taken and system is been trained with these images. So that system can detect emotion correctly and recognize it exactly. This paper uses the method of recognizing emotions of human being by using Euclidean distances. In this paper we have considered all seven basic emotions such as happy, angry, sad, fear, surprise, disgust and neutral. Total 60 facial expression images are considered and trained by using Euclidean distance. The images are trained for each emotion by calculating mean and subtracting from original image. The last step is to compare the features from edge detection and Euclidean distance with test image. Any emotion is combination of six primary emotions like- sad, surprise, anger, fear, disgust, happy.

Keywords: HCI, Euclidean distance

I. Introduction

Nonverbal communication plays major role in human interaction and along this nonverbal communication large portion is in form of face expression. For human being to show emotion face expression is the natural means. New challenge of Human computer interaction is capability to recognise face expressions. Recognising the facial expressions will result in to identify the basic human emotion i.e. anger, fear, disgust, sadness, surprise and happiness. The expression may vary individual to individual. Movement of facial features produces facial expressions. Facial expressions not only used to express emotions, but it is also used to provide essential communication clues while interacting socially. While communicating the non-verbal message, face gives the basis, the capability of reading the facial emotions becomes an essential part of emotion intelligence [2]. Charles Darwin had given first idea of expression of emotions in work build from his theory of evolution. Then Ekman and Friesen in their studies showed that there are six emotions —happy, sad, anger, surprise, fear and disgust which become very popular and are called as six basic expression.



Figure 1: Types of expressions

In my approach many algorithms are used to outputs the features of face such as lips and eyes, mouth. First of all the input image is loaded in train folder as well as test folder. Then after input images are analysed through various algorithms in order to improve the input image, to improve intensity, to remove noise of image. Secondly

edges of images are detected. Variety of distances between edge points is calculated and from that features are evaluated and algorithm called PCA known for dimensionality reduction is used so that data reduction can take place. In next algorithm face is detected. Emotions are now classified as correct emotions on distances basis.

II. Related Work:

With the help of automatic face Emotion detection system human-like robots and machines can be created. Automatic face emotion detection system helps to deal with the issues such as categorizing face images in different classes of expressions. Lot of progress is done in last few years. The bunch of multilayer neural net system is used for the categorization based facial emotion detection system. To extract the features of the Cohn Kanade dataset Logarithm Gabor filter is used and the feature vector is reduced by Principal Component Analysis [1]. Other study of face emotion detection system tells that linear programming algorithm is used for feature extraction [4]. In order to present facial expression MPEG-four method is used to give parameters for facial action. By using these facial action parameter, researchers have developed automatic facial emotion detection system [9]. And classification is done by maximum likelihood calculation, for that multi stream HMM is used to generate maximum likelihood. In the recent time period automatic emotion detection systems have gained popularity.

III. Proposed Approach:

Using variety of methods and techniques, here we have proposed an automatic emotion detection system from face expression from still images. Basically there is a need for camera to take face images of various kinds of expressions. After gathering face image it is supposed to go through preprocess step in order to remove environmental variations and others too from face image. It involves steps like color transformation, brightness adjustment etc. Now the feature is being extracted, feature of a face image are eye, mouth, nose, so these feature are detected. Further with the help of these features of face such as eye, lip categorization of emotions is done into six basic emotions and one is neutral.

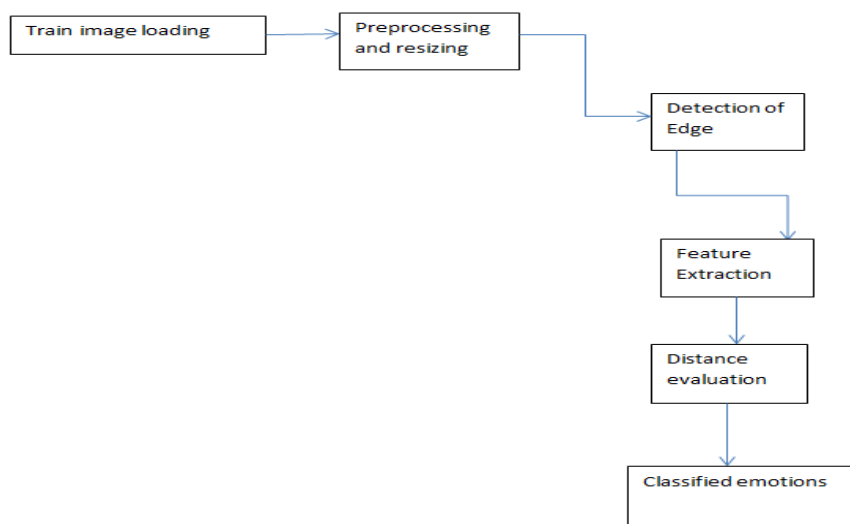


Figure 2: Flow chart of proposed approach

IV. Input Image preprocessing:

4.1 Color transformation and brightness adjustment



Figure 3: Original input image and Light adjusted input image

Light compensation based on luminance; in which we calculated average luminance Y_{avg} . First, we compute the average luminance Y_{avg} of input image. $Y_{avg} = Y_{i,j}$, where $Y_{i,j} = 0.3R + 0.6G + 0.1B$.

4.2 Skin extraction:



Figure 4: Output from preprocess stage

4.3 Noise removal:



Figure 5: Output of noise removal

V. Edge Detection and dimensionality reduction:

5.1 Edge detection: By edge detection corner points of face features are achieved such as eye, lip, & nose.

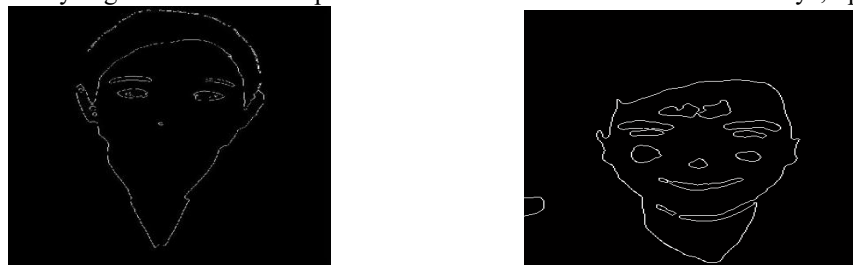


Figure 6: Output of edge detection

180x280x3 double

37.0540	39.2938	37.1618	22.3607	20.6155	16.2788	16.4012	8.2462	12.6491	16.1245	5.6569	11.0454
21.2132	19.2354	21.0000	21.9317	23.6008	20.2485	17.4642	5.8310	15.5242	32.3883	39.0128	31.2570
71.5681	59.2368	44.7214	26.1725	9.0000	34.2053	56.1427	72.0069	75.0000	67.0075	52.0384	45.0444
71.1758	74.2496	97.4166	109.1788	130.0500	136.5284	139.0036	128.4718	116.5204	108.0740	107.2287	122.0164
154.9322	134.6180	124.0363	116.9273	115.4340	97.0155	88.2780	92.5419	107.0047	119.3399	127.0984	119.0378
154.8289	43.4166	135.9301	158.7104	163.3799	71.1688	32.2025	30.0832	9.4868	21.8403	40.7185	59.8415
101.5529	37.4833	90.8240	71.8401	116.2454	76.3675	80.5047	98.0000	113.2166	124.1491	127.9062	125.7657
63.8905	75.6902	60.9590	61.6847	103.4843	118.9622	150.2065	144.0868	126.3210	114.3372	86.3481	68.6586
66.2193	29.6142	14.3178	44.9444	82.0244	81.4985	74.1080	63.2851	53.2353	40.0500	20.2237	8.9443
30.0832	16.4012	14.3178	33.2415	50.0100	62.1691	76.1577	90.4489	102.7035	110.1681	120.7021	129.7575
121.0496	125.6025	136.0919	148.2161	164.0030	168.1458	175.0343	171.0731	165.3663	158.6190	143.6837	136.1947
154.3794	147.5025	136.3672	128.4718	118.3427	106.3814	88.4590	69.5845	52.0384	32.0156	8.9443	23.7697
223.5173	206.8864	165.1696	120.2082	81.4371	43.5660	16.2788	4.2426	5.0000	8.0623	14.5602	9.2195
186.0000	223.0874	246.0772	276.2191	297.2423	318.6283	316.3100	297.5450	262.0115	224.0558	193.5820	157.2037
82.4924	49.6588	26.0192	48.3011	74.9466	110.1681	138.0580	156.2850	177.1384	189.2511	189.2749	151.7168

Figure 7: Feature vector after edge detection

Since, it is too big in size so further dimensionality reduction algorithm is used to reduce size of feature vector by normalizing mean to zero.

VI. Face feature extraction

Process used for it is to extract the shape and size of the eye, nose, and lip and through distance differentiating the faces and parts of face. Mainly five feature points are being used; every feature is in the terms of distances.

- Feature 1- width of left eye
- Feature 2 -width of right eye
- Feature 3 -width of nose
- Feature 4 -width of mouth corners
- Feature 5 -width of face

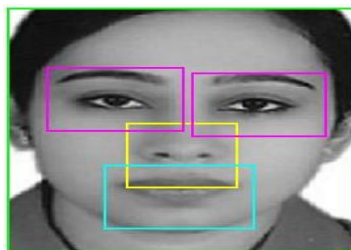


Figure 8: Output from extracting eye, lip, mouth and nose (face features)

VII. Face Detection

The motive behind face detection is to detect presence of human face in any given input images. The process is difficult because the parameters such as shape, size, color, texture of every image is different. Here in order to detect face skin region is separated by using threshold.



Figure 9: Face detection result

Wholes obtained after edge detection is filled to obtain connected region, then the skin color blocks are combined to get complete face skin color block, then finally face is identified in the given query image.

VIII. Distance Measurement:

8.1 Euclidean Distance computation: Euclidean distance formula if there is a feature vector of size n dimensions is given by:-

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

125	132	137	156	161	170	205	222	247	252	255	255	249	255	245	247	243	252	228	224	208	206	196	192
137	142	153	162	175	202	189	191	204	201	216	228	235	250	255	252	255	245	252	247	231	230	237	226
185	172	155	172	181	183	206	224	219	212	217	208	201	208	154	178	207	218	228	224	226	239	238	239
208	191	182	177	189	205	209	202	213	219	221	207	223	158	29	24	59	68	115	159	198	193	187	186
230	214	195	191	186	196	211	215	221	224	214	221	216	224	185	136	79	97	131	152	86	89	154	190
234	218	215	211	200	203	199	219	219	222	235	235	229	223	227	230	200	111	113	100	62	30	44	152
230	233	226	218	215	209	204	207	216	222	220	231	233	230	237	236	235	238	242	211	211	167	145	203

Figure 10: Features having n-dimensions

E_distance	7701
h	256x1 double
h1	256x1 1x1 double
I	180x280 uint8
J	180x280 uint8

Figure 10: Euclidean distance between any two images

IX. Emotion Detection: Emotion detection is done using Euclidean distance between the train image and the test image. Also the test image is compared with respect to neutral image meanwhile best matching image from database is also searched. The final output is in the form of txt contains emotion corresponding to the test image and how different from neutral the test image is also it give match from the database.

X. Results: Evaluation is done on system by 60 images of different expression of different people. Image size is 600×800 pixels, which was reduce further into 280×180 pixels. The system is tested on 42 images.

Average Total accuracy of the proposed system is: 93.78%.

$$\text{Accuracy} = \frac{\text{Correctly classified test images}}{\text{Total test images}} * 100$$

Sr. No	Emotion Type	Number of Test images	Number of Exact match occurs	Recognition rate
1	Happy	11	9	82%
2	Angry	09	8	88.78%
3	Sad	03	03	100%
4	Disgust	07	6	85.72%
5	Fear	03	03	100%
6	Surprise	03	03	100%
7	Neutral	06	06	100%

XI. Conclusion and future work

Main important contribution of the paper is proposed system can detect edges from given images and using that edges, distance among different features can be computed through Euclidean distance Formula. For every image this distance is different indicating different emotions. Emotions are categorized on this basis. This system can be of use in future for hardware implementation, such as in robotics, in security systems. In the future work, more number of emotions can be included for other prediction. More number of emotional classes can be included. Also work can be done on whether the teary face is of class sad or excitement and joy. Detecting emotion with constraint like mustache on face, sunglass at face, side view of face these things can be taken under consideration in future work. These are the issues on which focus need to be given. More number of features can be included to detect more complex expressions. Complete software database dependent and the camera resolution dependent. So if the good resolution DigiCam or good resolution analog cam issued then accuracy of detection can be improved.

References

- [1] Lajevardi, S.M.; Lech, M.; — Facial Expression Recognition Using Neural Networks and Log-Gabor Filters, Digital Image Computing: Techniques and Applications IEEE 2008.
- [2] H. A. Elfienbein, A. A. Marsh, and N. Ambady, "Emotional Intelligence and the Recognition of Emotion from Facial Expressions" in *The Wisdom of Feelings: Processes Underlying Emotional Intelligence*.
- [3] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan cultural elements in Facial displays of emotion" *Science*, New Series, vol. 164, no. 3875, pp. 86-88, April 4, 1969.
- [4] Guodong Guo, Charles R. Dyer, — Simultaneous Feature Selection And Classifier Training Via Linear Programming, A Case Study For Face Expression Recognition Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03) vol.1 2003, On page(s): I-346 - I-352.
- [5] Neha Gupta and Prof. Navneet Kaur "Design and Implementation of Emotion Recognition System by Using Matlab" *International Journal of Engineering Research and Applications (IJERA)* 2013, Vol. 3, Issue 4, pp.2002-2006.
- [6] Jyoti Rani, Kanwal Garg "Emotion Detection Using Facial Expressions" *International Journal of Advanced Research in Computer Science and Software Engineering*, April 2014, Volume 4, Issue.
- [7] Anurag De, Ashim Saha "A Comparative Study on different approach of Human Emotion Recognition based on Facial Expression Detection" in *International Conference on Advances in Computer Engineering and Applications*, IEEE, 2015.
- [8] T. Kanade, J.F. Cohn, Y. Tian, "Comprehensive Database for Facial Expression Analysis", *Proc. 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 46-53, 2000.
- [9] Devi Arumugam and S. Purushothaman, "Emotion Classification Using Facial Expression", *International Journal of Advanced Computer Science and Applications*, Vol. 2, No. 7, 2011.

Hand Gesture Recognition

Anushka Sharma¹ and Tushar Patnaik²

School of Information Technology (SoIT)

Centre for Development of Advanced Computing, Noida,
Uttar Pradesh, India

Abstract Lot of research work is done on hand gesture recognition to make the computer intelligent enough to communicate with humans. The signs made by hands are used to convey a message to the computer. More research is going on how accurately these signs can be predicted by the computer. This paper describes a method for recognizing Indian sign language alphabets given as input video in the form of hand gestures. In hand gesture recognition hands are used to convey a meaningful message. Communication is an essential form to convey one's thoughts and ideas. Normal people can communicate verbally but dumb and deaf people find it extremely difficult to express themselves so for their aid Sign Language has evolved.

Keywords Indian Sign Language, American Sign Language, Hand gesture recognition, SIFT, HOG, SVM, Background subtraction, Back propagation Neural Network

I Introduction

Automatic detection of Sign Language is an ongoing research area in the field of computer and human interaction. The automatic recognition systems are being developed for the replacement sign language interpreters and create an ease for the dumb and deaf community so that they can communicate to the outside world. Hearing and speech impaired people find it very difficult to communicate with others. According to a survey 5.5-15 million hearing impaired people are there in Indian Sign Language differs from region to region. The signs used for communication in one region may be different from the other region. Sign Language consists of a large dictionary which has different actions for different words. Characters or alphabets in sign language are a part of this dictionary. The formation of these characters differs from region to region. In American Sign Language only one hand is involved in the formation of characters but in Indian Sign Language two hands are involved in the formation of gesture so overlapping issue arises. Eighteen two hand characters are present in the ISL dictionary which are A,B,D,E,F,G,H,K,M,N,P,Q,R,S,T,X,Y,Z.

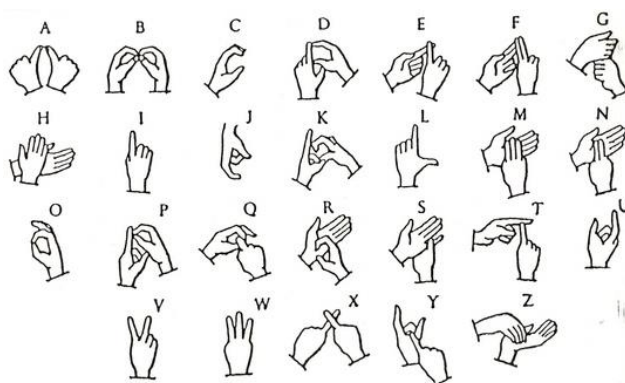


Figure 1 Character in Indian Sign Language

II Literature Survey

In [3] the morphological operation Thinning is used to remove selected foreground pixels from binary images. The database consist of the images of American Sign Language. It is used for obtaining the skeletal image as it reduces all lines to single pixel thickness. In [4] SIFT is used for extracting the keypoints from the hand gesture images. Space incompatibility of SIFT keypoints are space incompatible so bag of feature approach was applied. The vector quantization maps the keypoints to a unified dimensional histogram vector after the application of K-mean clustering. These histogram vectors are to be fed as inputs to multiclass SVM classifier for the recognition of the gestures. In [5] the morphological operation Thinning is used to remove selected foreground pixels from binary images. It is used for obtaining the skeletal image as it reduces all lines to one pixel thickness. The database used was American Sign Language.

In [6] In this paper, gradient direction histogram (HOG) features of gestures are being extracted, then Support Vector Machines is used to train these feature vectors then at the testing time, a decision is taken using the previously learned SVMs. The database used was American Sign Language.

III. Proposed Approach

The dataset used consists of videos of the formation of a gesture which are captured through a webcam. Three sets of videos are taken in this project for 18 two hand gesture. Two sets of videos are used for the training purpose and the third one is given for test. The number of correct predictions in the test videos will give the accuracy of approach.

The input video is converted into consecutive frames and a hybrid approach is applied using the HSI color model based segmentation and background extraction. In the HSI skin color segmentation the skin color pixels in the frame is encountered but it has disadvantage that if any other skin color object is present then it will also be detected and it is illumination variant so we apply moving object detection in the consecutive frames by using Background Subtraction. In Background Subtraction method the background frame is fixed which includes no gestures and the consecutive frames which include formation of a gesture. These consecutive frames will be subtracted from the background frame. So, only the skin color moving pixels will be detected which in our case are the hands.

While the frame is processed one after the other the number of regions in the frame is calculated by the 8-connected component method. Before the formation of a gesture the number of components will be two but after the formation the two hands will be touching each other so the number of connected components will be one. Only the frames in which the number of regions will be one will be considered for the feature extraction phase. Histogram of oriented gradients features are extracted to find the distinct features of each hand gesture. The extracted features are passed through the artificial neural network to identify the gestures. The neural is trained by using back propagation so that the system responds correctly to the test sample. The recognized gesture is given as output in the text format.

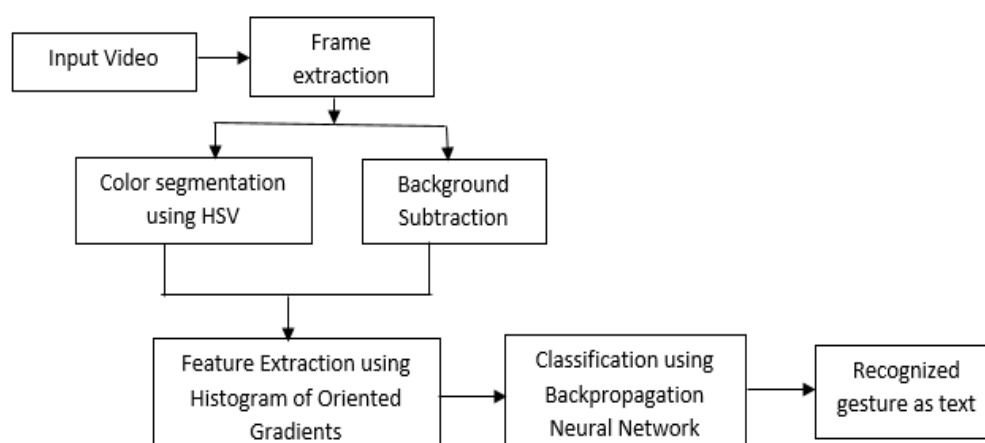


Figure 2 Flowchart of Proposed approach

Segmented Region

Segmentation is done so that only the hand region is extracted from the frame. The Hue Saturation color model is used to find the skin color segmented area and the background subtraction is done for the moving object detection. After the combination of both these techniques the final segmented region is obtained. The results of the segmented frames are as shown below

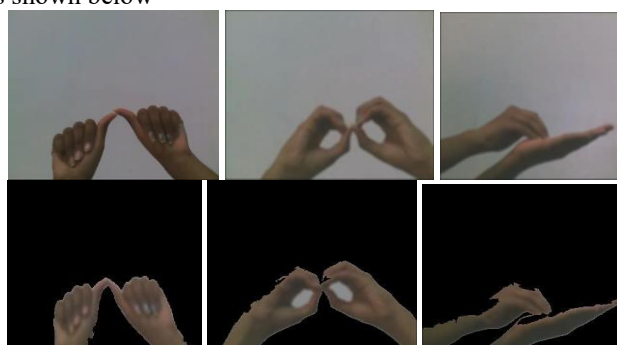
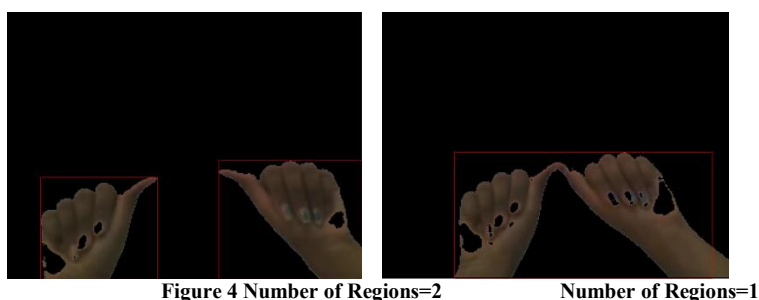


Figure 3 Segmented frames

Number of regions

While the processing of the frames the number of regions in the frames will be determined by the bounding box area. The frames in which the number of regions will be one will only be considered for the feature extraction.



Feature Extraction

Histogram of oriented gradients is used for the feature extraction. The major challenge with extracting features from two handed characters is that they include overlapping gestures. From the segmented hand regions the features required to distinguish one gesture from another will be extracted. The descriptor size should be equal for all the gestures. The smallest bounding box area including only the hand area is cropped and resized. The HOG feature vector obtained is fed as input to the neural network.



Figure 5 Magnitude of vectors in image by applying HOG

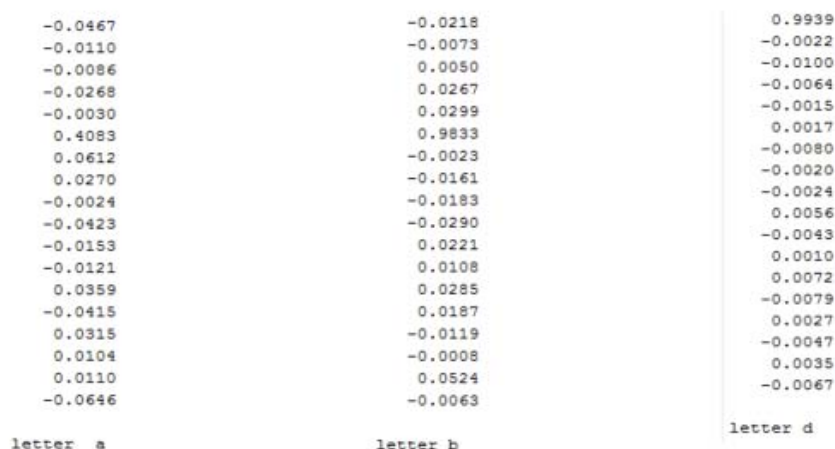


Figure 6 Results after training the HOG features of the characters

Classification

The back propagation artificial neural network is used to classify the hand gestures. During training ,the features of the training gestures are given as input to the network and the result is obtained for the gestures. If gesture is recognized correctly then the bias is updated for the next gesture. The error signal is passed back through the network and the weights or bias are updated.

Training stops when the error is negligible and correct predictions of the gesture are obtained.

IV Results and Analysis

Two videos for the gesture formation are used for training the classifier and one video for each gesture is used in the testing phase.

The misclassified gestures are N as M, R as M, X as E because they have close HOG features. The alphabets which after segmentation have close HOG features are misclassified.



Figure 7 Segmented M

Segmented R



Figure 8 Segmented N

Segmented M

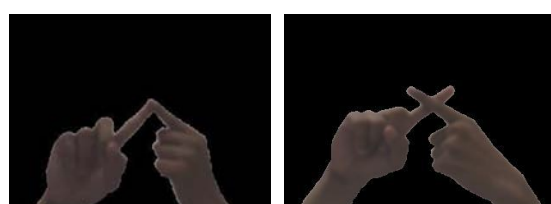


Figure 9 Segmented E

Segmented X

Character	Classified character
A	A
B	B
D	D
E	E
F	F
G	G
H	H
K	K
M	M
N	M
P	P
Q	Q
R	M
S	S
T	T
X	E
Y	Y
Z	Z

$$\begin{aligned}
 \text{Accuracy} &= \frac{\text{Total no. of correctly classified gestures} \times 100}{\text{Total no. of gestures}} \\
 &= \frac{15 \times 100}{18} \\
 &= 83.33\%
 \end{aligned}$$

References

- [1] <http://www.indian signlanguage.org-> Indian Sign Language Images
- [2] M. S. Sefat and M. Shahjahan, "A hand gesture recognition technique from real time video", 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, pp. 1-4, 2015
- [3] S.N. Omkar and M. Monisha, "Hand Gesture Recognition using Thinning Algorithm", ICTACT Journal On Image and Video Processing, vol. 02, 2014
- [4] Nasser H. Dardas and Nicolas D. Georganas, "Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Technique", IEEE Transactions On Instrumentation and Measurement, vol. 60, no. 11, pp. 3592-3607, Nov. 2011
- [5] S.N. Omkar and M. Monisha, "Hand Gesture Recognition using Thinning Algorithm", ICTACT Journal On Image and Video Processing, vol. 02, 2014
- [6] M. S. Sefat and M. Shahjahan, "A hand gesture recognition technique from real time video", 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, 2015, pp. 1-4

Recognition of Fruits and Vegetables from Images

Shanam Afzal¹ and Tushar Patnaik²

School of Information Technology (SoIT)

Centre for Development of Advanced Computing (C-DAC), Noida
Uttar Pradesh, India

Abstract: We have proposed a system which recognises fruits and vegetables from images. This system can be used by many applications like checkout system for supermarkets where it can be used as replacement of manual barcodes, as tool for learning by patients with down syndrome and by small children. This system can also find its application in the development of food quantity and nutrition estimation system. Lots of research work has been done in this area but with some constraints like recognising only single fruit/vegetable at a time or not recognising any cut fruits present in the image. This system recognises multiple fruits and vegetables present from the images and also any cut fruits present if any can also be recognised. Our system first segment the images with multiple fruits/vegetable using k means and then extracts features using bag of surf features and classify each of them.

Keywords: K-means, Bag of features, Recognition, Segmentation, Classification

I. Introduction

The proposed system is developed in order to recognize different multiple fruits and vegetables present in an image. This system can extend its application into development of other systems such as supermarket checkout systems where manual barcodes can be replaced by our system, in food nutrition and quantity estimation system to measure calorie and nutrition system and also as a learning tool for small kids and down syndrome patients. We are classifying 25 different fruits/vegetables through this system. We have collected the data of these 25 categories from the google images out which there are 11 fruits and 14 vegetables that we are recognizing through our system. We have different types of images, some of them have only single fruit/vegetable present, some of them have multiple fruits/vegetables of single category, some of them containing cut and/whole fruit and some containing different multiple fruits/vegetable present. Some examples are shown below.



Figure1.Single fruit (banana)



Figure2.Cut fruit (apple)



Figure3.Multiple onions



Figure4.Different vegetables

Most of the previous works has constraints that either the image contains only single fruit/vegetable or even if they have multiple fruits/vegetables present in the image then they are identifying these items with low accuracy rates. And previous research doesn't consider to classify the cut fruits present in the images. These are some of the constraints that have become the focus of our work and successfully overcome these overcome these constraints with good accuracy rates varying between 70-95% which is shown by experimental results.

The rest of the paper is divided into the following sections: II. Literature Survey related food image recognition describing all the previous work done related to the concerned problem. III. Proposed approach which explains the proposed approach. IV. Experimental Results which shows recognition and classification results. V. Conclusion concluding the paper and containing future work.

II. Literature review

Matsuda et.al [5] introduced an approach which consists of various region detection and segmentation techniques to detect various food items present in the image and uses different features such as color histograms, gabor and sift for extracting features and classify them using SVM classifier but achieved accuracy of 55.8%.

Rana *et al.* [2] worked on features of food such as color, shape, size and texture. Hough lines and circles are used for shape feature of food items. All of these features are fed to SVM to classify food items. Jotou *et al.* [6] recognized single food item at a time with classification rates up to 61%

III. Proposed approach

Proposed approach is shown below in figure 5

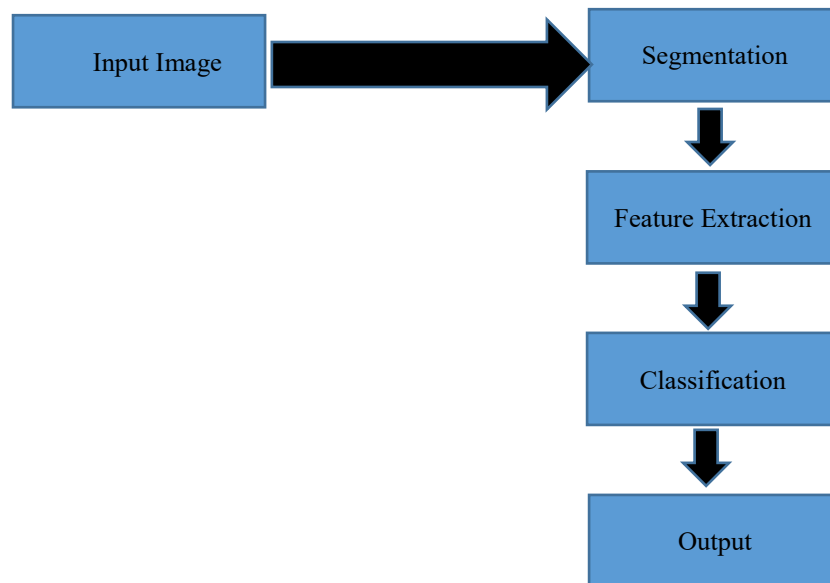


Figure 5. Flowchart of proposed approach

We have proposed the following approach to recognize fruits and vegetables present in the images given as input.

A. Input image

We provide an image as an input which consists of fruits or/and vegetables to the system.

B. Segmentation

Segmentation is a process that separates an image into multiple regions. Segmentation is performed by using k means algorithm. But before applying k means algorithm, we have converted the images from RGB color space to La*b* color space because it is most exact color representation and also device independent. Its accuracy and portability makes it suitable for our work. After applying k means clustering algorithm to divide the image into different clusters. The algorithm of k means is described below:

1. First of all choose k initial clusters that is centroids.
2. Then compute the cluster centroid distances of all pixels to each centroid using Euclidean distance.
3. Each pixel gets assigned to the cluster which has closest centroid to the pixel.
4. By averaging of pixels in each cluster, k new centroid locations are obtained.
5. Repeat steps 2 to 4 until centroids no longer move. This divides the image into separate color clusters.

When there are objects present in the image of same color then we manually crop the image to divide into different segments.



Figure 6. Original Image

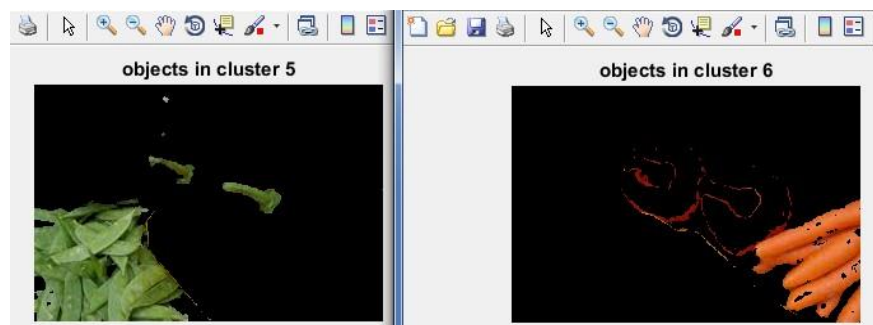


Figure 7. Segmented images

C. Feature Extraction

We have used the bag of surf features for our feature extraction technique. First of all surf features are extracted from all the images present in all image categories.. SURF (Speed Up Robust Feature) descriptors are invariant

of scale, orientation and illumination changes. We build a dictionary of code words and used subsets of the training images to build the dictionary of code words. Then we plot the descriptors obtained from this subset onto high dimensional space and clustered using k means algorithm into 500 clusters.

D. Classification

Features extracted from the images are encoded and feed into multiclass linear svm. We have trained 25 linear svm as we have 25 categories to classify. Linear svm is fast and saves memory. This is used for training as well as testing of the images.

IV. Results

In this paper we have successfully recognized fruits and vegetables from the given images whether images contain single or multiple or whole or cut fruit/vegetable. Some of the results are shown below:

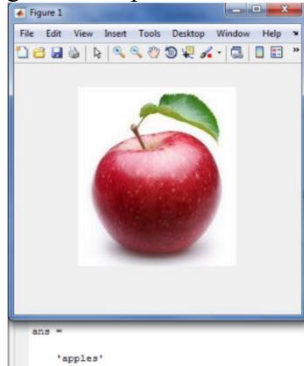


Figure 8. Output of single fruit

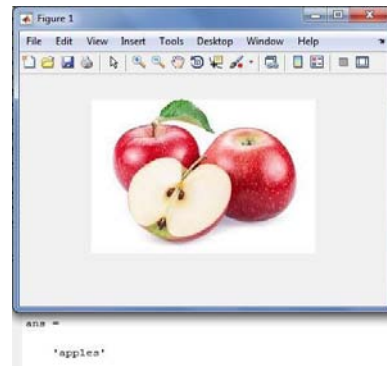


Figure 9. Output when cut fruit is present



Figure 10. Input image

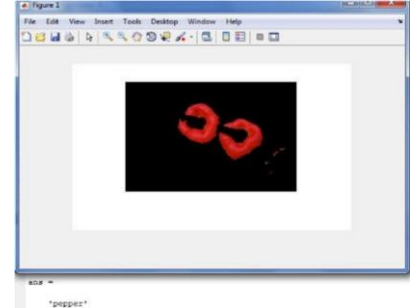
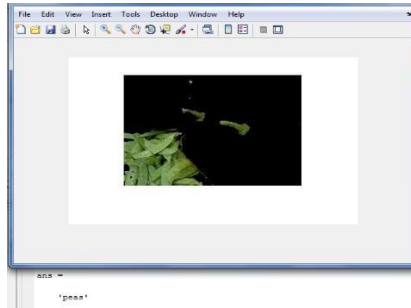


Figure 11(a) and (b). Output from input image

V. Conclusion

In this paper we have implemented our approach recognizing fruits and vegetables from images successfully. The recognition accuracy varies from 70-94% for various type of images. Accuracy has improved from the previous work in this field and also we have considered images with cut fruits/vegetables present in the images. For the future work, the system can be developed for including other food items as well apart from fruits and vegetables that can be further used for developing food nutrition and quantity estimation systems. Improving the segmentation techniques may further improve the accuracy of the system.

VI. References

- [1] P. Pouladzadeh, S. Shirmohammadi and R. Al-Maghrabi, "Measuring Calorie and Nutrition From Food Image," in *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 8, pp. 1947-1956, Aug. 2014.
- [2] I. Boujelbane, S. H. Said and T. Zaharia, "Multi-object recognition and tracking with feature points matching and spatial layout consistency," *2014 IEEE Fourth International Conference on Consumer Electronics Berlin (ICCE-Berlin)*, Berlin, 2014, pp. 355-359.
- [3] Yoshiyuki Kawano and Keiji Yanai, "Real-time Mobile Food Recognition System" In proc. Of IEEE CVPR International Workshop on Mobile Devices, 2013.
- [4] Y. Matsuda, H. Hoashi and K. Yanai, "Recognition of Multiple-Food Images by Detecting Candidate Regions," *2012 IEEE International Conference on Multimedia and Expo*, Melbourne, VIC, 2012, pp. 25-30.
- [5] Taichi Joutou and Keiji Yanai, "A food image recognition system with Multiple Kernel Learning," *2009 16th IEEE International Conference on Image Processing (ICIP)*, Cairo, 2009, pp. 285-288.

Brief Overview & Comparison of Various Energy Aware Routing Protocols in MANET

Sameeksha Kukreti¹ and Rekha Saraswat²

School of Information Technology (SoIT)

Centre for Development of Advanced Computing (C-DAC), Noida

Uttar Pradesh, India

Abstract: MANET is all about wireless mobile ad-hoc network without base station. In MANET various wireless nodes are connected within their transmission/receiving range. Nodes can move freely in MANET and can communicate to other nodes through in between nodes directly or indirectly. As the nodes are operated by battery, saving energy is an important issue in the MANET. Node energy failure can affect the entire network thus to increase the life, routing within the network has have to be energy wise efficient. In order to enhance the lifespan of the network, routing must be energy efficient. This paper gives an overview of different approaches available for energy efficient routing protocol and gives an insight to these protocols.

Keywords:- MANET - Mobile Ad-hoc Network, DSR-Dynamic source routing protocol, RRQ - Route Request ,RRP - Route Reply, RER -Route Error.

I. Introduction

MANET is an independent network with no central controller and without fixed network infrastructure and is being governed by different mobile host referred as nodes. Now-a-days there is increase in the number of portable devices such as laptops, mobile phones etc. They are becoming essential for an individual, since these are easily available, user friendly and cheaper. This leads to Mobile ad-hoc network a new way for communication. In MANET each node is independent, operates as end node and router for other nodes. MANET is a self-administrating, self-healing, self-organizing and self-forming network. MANET usually consists of battery operated computing devices.

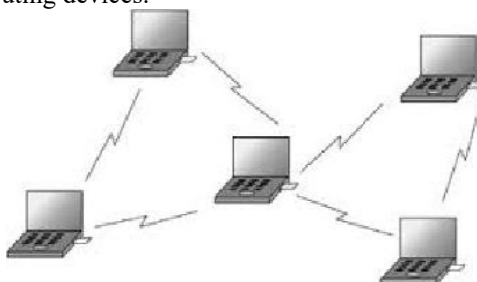


Figure: Mobile Ad hoc Network

Silent features of MANET:

- **Not fixed/changing layout:** as in MANET nodes are mobile thus they are free to move apart from radio propagation and thus can lead to dynamic changing, unpredictable network layout.
- **Limited Bandwidth and changing link capacity:** It has been seen that links, which are wireless has got less capacity as compared to that of wired links owing to multipath propagation, mingling noise and roaming signal interface. Hence it can be concluded that throughput of the system is usually less than wired system
- **Limited energy of mobile nodes:** In MANET nodes are mobile and run by battery. There are many nodes in MANET and are connected to each other through their transmission/receiving range .Enfeeble of node battery would have a large effect on whole network.
- **Multiple hop communication:** MANET does not require any infrastructure, as in MANET nodes which are generally mobile, communicates with each other through their transmission/receiving range. In MANET data packet cannot be send directly to the final destination node, hence in-order to transmit data to the final end node (destination node), starting node (source node) has have to relay the data packets through in-between nodes.

The main concern in MANET, as clearly described above is preservation of energy as nodes are moving and are run by battery. Energy is important asset in MANET. As ad-hoc network allow multiple -hop communication

among mobile nodes which are run by battery, energy preservation becomes very important in-order to preserve the connectivity of the ad-hoc network. Multiple energy aware routing protocols have been proposed by employing miscellaneous approaches such as broadcast power adaptation, adaptive sleeping, multiple hop communication, topology control etc. Out of them many approaches consider routing metric for example hop count and delay. These approaches do not take into account remaining energy of the battery. So efficient utilization of energy is proportional to life span of network therefore preservation of energy is an important task in MANET. This paper gives an overview of various different approaches available for energy efficient routing.

II. Literature survey

To perform an analysis of various existing routing protocols which work for energy efficiency in MANET, brief description of the protocols is given below.

Baisakh, Nileshkumar R Patel, Shishir Kumar. [1]et.al 2012 proposed an approach, which modifies DSR (a well-known routing protocol for ad-hoc network). DSR does not take into account, the energies of the in-between nodes. The author mainly focuses choosing the most efficient optimized energy route from starting (source) node to end (final) node. Author has chosen only those in-between nodes which have residual energy level more than threshold level at a particular time. If energy level of the node is less than threshold level than the new path finding process is again reinitiated. This approach suffers from overhearing and stale route problems.

Madhubala Patil, Sowmiya Raksha, R.K.K. Joshi, V.B.Nikam.[2]et.al2015 proposed an approach, which is basically an alteration to previous approach, in previous approach mobile nodes are chosen based on minimum threshold value. As previous approach has overhearing and stale route problem, which results in packet loss and more energy usage .In this approach the answer to overhearing problem has been solved by making some changes in the previous approach. Author has also used timer for energy preservation.

Baisakh, Nileshkumar R Patel, Shishir Kumar .[3] et.al in this approach which was proposed in 2012, author considers DSR, a well-renowned routing protocol, as the basic fundamental protocol and tried to do some changes in its algorithm which further lead to the formation of new ESSDSR protocol. In this approach main motive of author is to send data from nodes having energy level higher at specific time. In this approach author has changed the DSR in such a way that if mobile node which is sending the data in multiple route arrive at level below or same to the threshold value of its initial energy of the battery, then mobile node ask the adjoining mobile nodes to see different path for sending data as node can die soon because of energy depletion. So this modified approach has also brought not only conservation of energy but also brought up energy survival aspect for mobile nodes with low energy.

Akanksha Meshramt, M.A. Rizvi.[4] et.al 2014 proposed an approach, in which author worked on beating the difficulty of low battery energy in the mobile nodes. In this paper, main aim is to save energy and to bring up new energy awareness scheme in MANET. The author in his approach, has first fix the minimum threshold energy level of every node to a particular value and then the sender node will circulate the data to those adjoining nodes, after examining the threshold level of every path in-between nodes to destination, after that all path to the destination are identified. Abandon all the routes having energy less than threshold level. Only those paths whose energy level is greater than threshold will be examined.

F. J. Ros, P. M. Ruiz,[5]et.al2014 et.al. In this approach author's main focus is on the topology of the network, that various changing layouts of the ad-hoc network can easily be adapted by commanding the transmitting and receiving powers, overall the author has suggested a Distributed protocol named as DBSSP protocol in-order to minimize the overall energy usage of the network and to deal with changing layout at network layer.

Pinki Nayak , Rekha Agarwal and Seema Verma.[6]et.al2011In this author has analyzed the comparative energy usage of two well-known routing protocols in MANETs, such as AODV (Ad hoc On-demand Distance Vector Protocol) and DSR (Dynamic Source Route Protocol) under different mobility cases. Further it is shown in the paper that DSR performance is better under low mobility conditions.

III. Comparative Analysis

In this section, Comparision of all the existing schemes/ protocols are done, that have discussed so far.

TITLE	AUTHOR	ABSTRACT	ADVANTAGE	LIMITATION
ECDSR in MANET	Baisakh, Nileshkumar R Patel, Shishir Kumar.	In this approach author modify DSR protocol a well-known routing protocol for ad-hoc network, which do not take into account, the surplus energy of the node. Here in this paper the main focus of the author is to choose the best optimum route from starting node (source node) to final end node (destination node).	1) Selects the path having nodes with higher amount of energy. 2) Number of dead nodes are less. 3) Enhances the network lifetime.	1) Overhearing problem. 2) Stale route problem. 3) Nodes always remain in active state

Extended ECDSR Protocol for Energy Efficient MANET	Madhubala Patil, Sowmiya Raksha, R.K.K. Joshi, V.B.Nikam .	This approach is basically an alteration to previous approach, in previous approach mobile nodes based on minimum threshold value is chosen. As previous approach has overhearing and over decade path problem, which surpass to packet misplace and more energy usage .In this new approach the answer to overhearing and over decade path has given by indicating changes in previous approach.	1) Solve overhearing problem. 2) Consider the residual energy of the node. 3) Improves the lifespan of the network.	1) Does not solve the over decade path problem. 2)No specific technique to select which node should overhear and which node should not overhear.
ESSDSR in MANET	Baisakh, Nileshkumar R Patel, Shishir Kumar.	In this method which was proposed in 2012, author considers DSR a well-known renowned routing protocol as the fundamental protocol and tried to do some changes on it which leads to the formation of new EESASDSR protocol. Here in this approach main motive of author is to send data from nodes having energy level higher at specific time.	1) Avoids low energy nodes from overusing. 2)Improve the individual node life time .h 3) Enhances the lifetime of the network.	This method work efficiently for 10-12 nodes only
Novel method for trustworthy Communication in MANETs	Akanksha Meshramt, M.A. Rizvi.	In this approach author worked on beating the the difficulty of low battery energy in the mobile nodes. Here in this paper author main aim is to save energy and to bring up new awareness scheme in MANET.	1) Increase in success rate of packet sending by protecting nodes from fading out due to energy failure. 2) Improves the performance of the network.	1) Link breakage problem. 2) This method uses threshold energy, in this method energy remain in the network after threshold are wasted and for better result need to apply optimum threshold.

IV. Conclusion and Future work

The research work explored above shows that MANET is increasingly playing a vital role for efficient and quick data communication, however owing to its wireless nature and multiple hop nodes characteristics, MANET is completely constrained by its energy consumption issues. Thus all research work emphasis on increasing its energy efficiency by discovering routes for data transmission from source to destination using less energy. If routes are identified by taking nodes residual energy into consideration, at the time of route discovery then more energy efficient paths can be identified.

V. References

- [1] Baisakh, Nileshkumar, R Patel "Energy Conscious DSR in MANET" 2nd IEEE International Conference on Parallel, Distributed and Grid Computing, 2012.
- [2] Akanksha Meshramt, M.A. Rizvi "Novel Approach for Reliable Communication Using Energy Aware Routing Protocol in MANET", IT in Business, Industry and Government (CSIBIG) Conference 2014.
- [3] Baisakh, Nileshkumar R. Patel, "Energy Saving and Survival Routing Protocol for Mobile Ad-Hoc Network", International Journal of Computer Applications (0975 – 888) Volume 48– No.2, June 2012.
- [4] Madhubala Patil, Sowmiya Raksha, R.K.K. Joshi, V.B.Nikam , "Extended ECDSR Protocol for Energy Efficient MANET", 2015 International Conference on Advanced Computing and Communication System (ICACCS -2015), Jan. 05 – 07, 2015, Coimbatore, INDIA.
- [5] Pinki Nayak, Rekha Agarwal and Seema Verma, "Effect of Mobility on Energy Consumption in DSR and AODV protocols in Mobile Ad hoc Networks", Journal of Computer Science and Engineering, Vol. 9, Issue 2, PP: 25-29, Oct. 2011.
- [7] Aarti, Dr. S. S. Tyagi, "Study of MANET: Characteristics, Challenges, Application and Security Attacks". International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 3, Issue 5, May 2013.
- [8] F. J. Ros and P. M. Ruiz, "Implementing a New Manet Unicast Routing Protocol in NS-2", University of Murcia, Dept. of Information and Communications Engineering, Spain, December 2004.
- [9] Mohd Tahir, Anas Iqbal, Abdul Samee Khan, "A Review Paper of Various Filters for Noise Removal in MRI Brain Image", International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCCE), Vol. 4, Issue 12, December 2016.
- [10] Sushant Kumar Choudhary, Vimlesh Kumar, "Review on Energy Efficient Routing Protocol in MANET" International Journal of Advanced Research in Computer and Communication Engineering (IJARCCCE), Vol. 5, Issue 4, April 2016.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Sentiment Analysis of Social Media and Web Data using Machine Learning

ShivaKarthik S¹, Lovey Joshi², Krishnanjan Bhattacharjee³, Swati Mehta⁴, Ajai Kumar⁵

^{1,2,3,4,5}Applied Artificial Intelligence Group, CDAC, Pune, MH, INDIA

Shivangi Jain⁶, Aakruti Katre⁷, Nikita Gatagat⁸, Yashodhara Haribhakta⁹

^{6,7,8,9}Department of Computer Engineering and Information Technology

College of Engineering, Pune-5, MH INDIA

Abstract: The aim is to develop a generic machine learning tool which has been depicted through development of sentiment analysis system comprising of polarity calculation, domain identification, language identification, spam/ham identification on natural language text extracted from social media and web data. Various machine learning algorithms/techniques are used to process the extracted data to get the normalized data. This data will be firstly normalized using various preprocessing techniques and then further undergo natural language processing. It will also perform tasks like Sarcasm identification, analysis of Contextual tweets for cumulative sentiments of a given base tweet. The data will be stored and indexed through Apache Solr. It will then be accessed by user query on a graphical user interface. The entire output is depicted through comprehensive visualization of sentiment analysis of social and web media where the efficiency is determined by inherent Machine Learning algorithms and training of the Learning systems

Keywords: Sentiment Analysis; Machine Learning; Domain Identification; Apache Mahout; Apache Solr; Natural Language Processing; Social Media

I. Introduction

As Cognitive Science tries to simulate human intelligence capabilities, the brute force methods of statistical sampling or hard coded methods of language analysis rule based system will not suffice in the age of text analytics and big data where retrieval is not enough but efficient retrieval of information along with intentions hidden within information like sentiment analysis have become very important. Supervised Machine Learning has become one flexible way to handle Big Data where varieties, time and efficiency or relevancy are the keys. Sentiment analysis system has been developed to depict an instance of Machine learning in this project. Sentiment Analysis refers to the determination of attitude of a speaker with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The basic task in sentiment analysis is to classify the polarity of a given text of the document, sentence, or feature or aspect level whether the expressed opinion in a document, a sentence or an entity feature is positive and negative. It uses NLP, Machine learning methods or statistics to extract, recognize or characterize the sentiment content of a text segment. Social media texts provide large quantities of interesting and useful data as well as new challenges for NLP and Machine Learning. Social media texts include chats, online commentaries, reviews, blogs, emails, forums, and other genres. Typically, the texts are informal and notoriously noisy. Thus, many NLP tools have difficulties processing and normalizing the data. To prepare the text for suitable sentiment analysis we need to clean and preprocess the data. Lots of uninformative parts and noise is present in the online text which can cause hindrance in further analysis of the text. One of the benefits of analysis is to be able to identify the current trending topics and discussions among a large user community. It can also help in understanding the public opinion on happenings of the world. E-News data helps us to know the recent happenings around the world. Sentiment Analysis helps us in monitoring the social media and identifying public opinions.

II. Related Work

Sentiment Analysis had very wide applications and several Research Papers analyze it. The main aspect of machine learning is feature selection. [2] Identification of the most salient features for learning, focusing on a learning algorithm on those aspects of the data which will be useful for the analysis and future prediction are the main objectives of the feature selection. The paper [3] states that their learning model predicts that the inclusion of social media information can help in predicting the sentiment to a better extent as compared to the approach using SVM wherein only textual features are used for predicting the sentiment. The use of linguistics is used to analyze the sentiment of twitter messages in this paper [4].

III. Objectives

The objectives are to develop several representative models of Machine Learning. They are as follows:

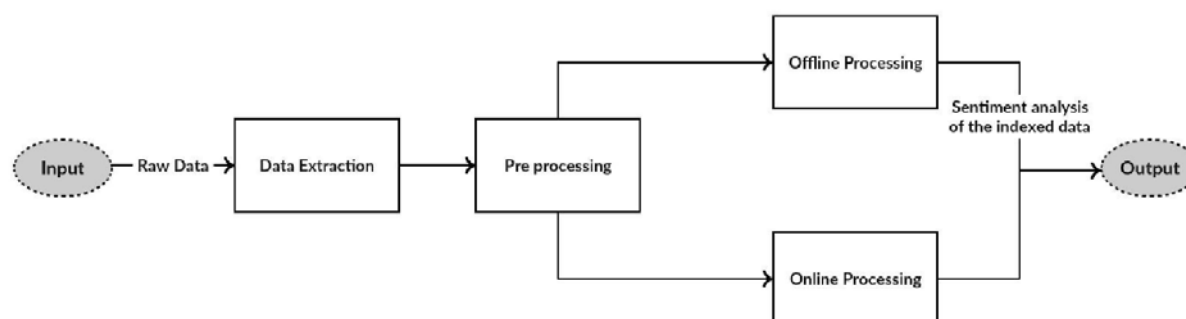
1. To develop an integrated tool which performs task of sentiment analysis using domain identification, language identification, Spam/ham filtering along with multi word extraction using Machine Learning tools, approaches, and trainable data sets.
2. To identify polarity of a sarcastic text segment.
3. To analyze the contextual polarity of tweets based on polarity of the base tweet.
4. To perform dynamic clustering on real time queries to obtain polarity statistics of social media information and web data analytics

IV. Data

The data for Language Identification model has been obtained from Leipzig Corpora collection [6]. 1 Million Sentences of German, French and English language have been used to train the model. For Hindi, manually collected data set of 150 documents is used. For Spam/Ham identification model the data has been manually collected from twitter and Facebook feeds. For Domain Identification, data is collected from two sources, BBC news corpora and manual collection from Google News. For Sentiment Analysis, data set is collected manually from Google news and segregated as positive and negative. The data for sarcasm is collected manually.

V. Design Model

The designed system is divided into different modules. The following diagram depicts them:



VI. Implementation

1.6.1. Data Extraction.

This module deals with the task of data extraction. Data is extracted from three sources namely, Twitter, Facebook, and E-News.

i. Data Extraction from Twitter

Data is extracted from twitter using the Streaming API. Streaming API streams real time tweets based on user query. Twitter4j library is used to access the Streaming API.

ii. Data Extraction from Facebook

Data is extracted from Facebook using Facebook Graph API. The Facebook public pages can only be accessed to extract the data.

iii. E-News Extraction

For extraction of E-News Jsoup library is used. The required information from the web news includes the title, main content, URL, keywords. These things can be extracted using the tags.

1.6.2. Preprocessing.

This module deals with all the preprocessing tasks. The output of this module will be normalized data. The Tasks involved are, splitting attached words, Spelling correction, Expansion of abbreviations, Extraction of Hashtags, Cashtags and Mentioned names, and Expansion of emojis and smileys.

i. Splitting attached words

Due to word limit in twitter and introduction of hashtags a lot of words are combined. For e.g.: words like #IamGoingToMumbai will be separated as # I Am Going To Mumbai. This word splitting will help in making the text readable and will be of use during the next tasks.

ii. Spelling correction

It is necessary that the spelling mistakes are corrected. In this process, initially the words are compared to a dictionary. If the word is valid it is put further in the pipeline for processing, else spelling correction is done. Spelling correction is done based on proximity analysis and using a word probability file. So, an attempt to

correct a word and substitute it with its root word is made. For example: detromed will be corrected to destroy and then put into the pipeline for further processing.

iii. Expansion of abbreviations

E-news as well as social media data contains abbreviations. For abbreviation expansion, a dictionary is made which contains several abbreviations along with their expanded form. The words in data are compared and dictionary lookup is performed for replacing an abbreviation with its suitable expanded form. For e.g.: BRB, which is generally used in Twitter, is expanded to Be Right Back





iv. Extraction of Hashtags, Cashtags and Mentioned names

Hashtags are used to highlight keywords in a social media data. They are denoted by #. Cashtags are used to denote prices, such as stock prices. They are denoted by \$. Mentioned names are used to find names of persons mentioned in the text. They are denoted by @. Extraction of these will help us to identify keywords, prices as well as mentioned people in the text. For this we have used twitter-text java library. A string is taken as input and output obtained is the extracted data and original text. Result will be - Mentioned names: levijedmuxphy, snapchat Hashtag: frustrated Cashtag: null.

v. Expansion of emoticons and smileys

This preprocessing module will remove the emojis character from the social media data and replace the smileys and emojis with its corresponding meaning. This expansion will thus be helpful in analyzing the sentiment of the text and for polarity calculation. For e.g.: will be replaced by :happy face:. The open source emoji4j library is used to obtain the data set for emojis and its corresponding aliases. The emojis.json file contains the list of near about 350 emojis and its meaning.

Table I Examples of some Emojis

Positive	Negative
 , 	 , 
8-)	8-(

1.6.3. Offline Processing.

This module deals with all the processing tasks. These are Language Identification, Spam/Ham Filtering, Domain Identification, Sentiment Analysis.

These tasks have been done using Transformed Weight-normalized Complement Naive Bayes Algorithm(CBayes). It is a machine learning algorithm. The platform used is Apache Mahout. Naive Bayes is simple to implement and is based on the principle of feature independence. CBayes boosts the accuracy and is comparable to state of the art algorithms like Support Vector Machine(SVM).Using CBayes algorithm, text is modeled better through transformations and problems of skewed data bias and inappropriate weight has been solved. Some of the formulas used by this algorithm are shown in the figure 1.

For all the tasks, trainable models have been created. Presently, four languages are identified. They are Hindi, English, German and French. The training data (documents) have been manually selected and then a model is created using the algorithm. There are six domains in total. The text is classified in one of these six domains. Spam Ham filtering is also done. Tweets which are too personal or contain only urls or images are considered as spam tweets. Rest is classified as ham. Sentiment analysis of the text is done. The text is identified as having positive or negative polarity. Also, sarcasm identification is done using a database. Analysis of Contextual tweets for cumulative sentiments of a given base tweet has also been done. It means that the sentiment of each comment is identified based on the sentiment of the base tweet. For this, the base tweets and reply tweets have been retrieved from twitter. They have been retrieved using the Search API. A query is made. For each retrieved result, another query is made of the form toUserIdsinceTweetId. The InReplytoStatusId of resultant tweets are compared to the TweetId of the base tweet. If they are equal then the tweet is added to the result tweets list of the base tweet. Then the polarity of base tweet is identified. Then the individual polarity of reply tweet is identified independent of the polarity of base tweet. Now these two polarities are combined to get the contextual polarity of comment tweet. The combination is done in this way: If the base polarity is positive and tweet polarity is also positive, overall polarity is positive. If the base polarity is positive and tweet polarity is also negative, overall polarity is negative. If the base polarity is negative and tweet polarity is also positive, overall polarity is negative. If the base polarity is negative and tweet polarity is also negative, overall polarity is positive. After the polarity identification, this processed data is indexed into Apache Solr. It is indexed using uploading Xml files from terminal or using Solrj library. Solrj is java library provided by Apache foundation to add documents to Solr. It is an open source storage and search platform. It allows storage of unstructured data. The fields will be id, keywords, title, content(/tweet), domain, polarity. This stored data is then used for online processing.

Figure Error! No sequence specified. θ represents the probability and subscript i is the word in class c . w is the weight of word i in class c

- Let $\vec{d} = (\vec{d}_1, \dots, \vec{d}_n)$ be a set of documents; d_{ij} is the count of word i in document j .
- Let $\vec{y} = (y_1, \dots, y_n)$ be the labels.
- TWCNB(\vec{d}, \vec{y})
 1. $d_{ij} = \log(d_{ij} + 1)$
 2. $d_{ij} = d_{ij} \log \frac{\sum_k 1}{\sum_k d_{ik}}$
 3. $d_{ij} = \frac{d_{ij}}{\sqrt{\sum_k (d_{kj})^2}}$
 4. $\hat{\theta}_{ci} = \frac{\sum_{j: y_j \neq c} d_{ij} + \alpha_i}{\sum_{j: y_j \neq c} \sum_k d_{kj} + \alpha}$
 5. $w_{ci} = \log \hat{\theta}_{ci}$
 6. $w_{ci} = \frac{w_{ci}}{\sum_i w_{ci}}$
 7. Let $t = (t_1, \dots, t_n)$ be a test document; let t_i be the count of word i .
 8. Label the document according to

$$l(t) = \arg \min_c \sum_i t_i w_{ci}$$

1.6.4. Online Processing.

The data gathered from the Preprocessing and offline stages is finally indexed into Solr. The system has 5268 documents indexed as of now. Out of these, 1536 documents obtained from e-news are indexed, 1641 from Facebook are indexed and 2091 documents from twitter are indexed. The general queries while retrieval of documents for indexing are India, Ransomware, Bollywood, Internet, NASA, ISRO, cartoons, Bahubali, Salman Khan, Computers, USA, Zika Virus, CDAC etc. 6000 documents were retrieved for indexing, out of which 732 were in a language other than English or were considered as spam. Results for some queries have been shown in the results and analysis section. The Visualization is done on the 'Banana' dashboard using the data in Solr server. The user can query and relevant results are displayed both in the form of documents and as charts like Sunburst, Bar graph etc. Clustering is done using 'Carrot'. Dynamic query can be entered by the user on browser and the clustered results are obtained. The algorithm used for clustering in the project is K-Means Clustering. [5].

VII. Result and Analysis

The aim is to understand supervised machine learning using sentiment analysis as an instance. To increase the accuracy of sentiment analysis a number of processing steps have been carried out. The data obtained from social media sites is inconsistent. In order to get data in an intelligible format text segmentation, emoji translation, spelling correction have been performed. This normalized data is used for processing. The language of the text is identified as only English language text is desired. Spam Ham helps in clearing out all the unimportant data. Domain Identification is done to identify the domain of the text as a text can have different meaning in different domains. It has been observed that sentiment of a text is domain dependent. This is because meaning of words differs in different domains. e.g.: the meaning of hit in entertainment domain is success(The movie was a hit.), in sports is scoring (The batsman hit a six.) and in politics it is hitting someone(The person hit the minister.). The sentiment of hit for entertainment and sports domains is positive whereas in politics it is negative. This problem is solved by training different sentiment analysis models for different domains. The training set used for these models is domain specific. This has helped in improving the accuracy of sentiment analysis. The tables below show the number of documents used for training of each model and the accuracy obtained for each of the models trained using Naive Bayes Algorithm.

Table II Overall Result

Modules	Accuracy	Precision	Recall	F-measure
Language Identification	85.40%	0.81	0.79	0.80
Domain Identification	83.60%	0.85	0.85	0.85
Spam/Ham Identification	78.00%	0.75	0.76	0.76
Sentiment Analysis	80.34%	0.79	0.77	0.78

Table II shows the overall data which has been used for the training of the models. This data has been collected manually or from online corpus as explained in the data section. In all 9037 documents have been used for training the domain identification models. The accuracy obtained is from testing the models on 1357 documents.

The approach for sentiment analysis of sarcasm is also data driven. Classified training data is supplied for model building along with the other training data in the sentiment analysis models. Some examples of sarcastic sentences put in the data set are: This is too good to be true.(Negative Sentiment) I work 40 hours a week only to be poor.(Negative Sentiment) Sentiments are largely dependent on the context of conversation and the tone of the speaker. Hence, this approach does not add substantially in the accuracy of the sentiment analysis model. The contextual polarity of tweet is analyzed based on polarity of the base tweet. Results obtained for some example texts are:

Table III Contextual Polarity

	Text	Polarity
Base Tweet	Iamenragedbythenewpolicies.	Negative
Reply Tweet	Ifeelthesame.	Negative
Reply Tweet	Idontagree	Positive

Table IV Online Processing

	Number of Documents
E-news	1536
Facebook	1641
Twitter	2091

Table IV shows the total number of documents indexed in the Solr server as of now from each source namely, E-news, Facebook, and Twitter.

Table V Domains of Queries put in our System

	Domains
India	Entertainment, Science&Technology, politics, sports
Internet	Entertainment, Science&Technology, Business, Sports, Health.

In Table V, the results of retrieved documents have been shown. The query is the query which we have put on our banana dashboard, the domains are the domains in which the results of queries have been classified.

Figure 2 Bar chart showing Sentiments.

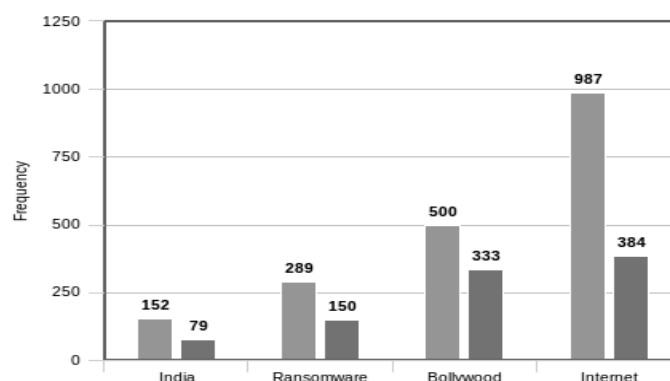


Figure 2 shows the four queries which were done on the system. The columns in the chart represent how many documents were classified as negative and how many as positive in the retrieved results. The left column shows the number of documents classified as having a positive sentiment and the right column shows documents classified as having a negative sentiment.

Table VI Accuracy based on Five Queries

Query	Our System Accuracy	Other System Accuracy
India	96.28%	93.00%
Internet	85.00%	83.88%

In Table VI the query column shows the queries done on the system and the accuracy is of our system is compared with the already made system. The accuracy of our system is determined by comparing the automatic classification by manual classification done on the documents. The accuracy of other system is calculated by determining the sentiment of the retrieved documents on the other system and comparing the results with manual classification.

Figure 3 Sunburst depicting Domains, Sentiments and Ids.

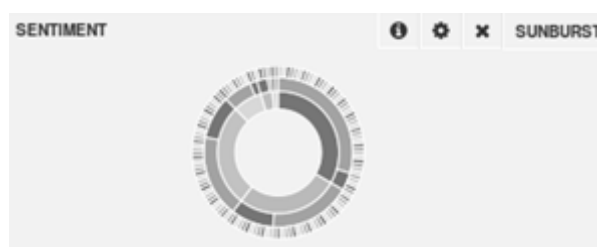


Figure 4 shows the visualization of the data. All the data represented here is about India. Innermost circle represents the domains in which documents containing India have been classified. The outer circle represents the overall negative and positive sentiments in each domain. The outermost circle represents the id of the documents belonging to each domain for each sentiment.

Figure 4 Bar chart showing Sentiments.

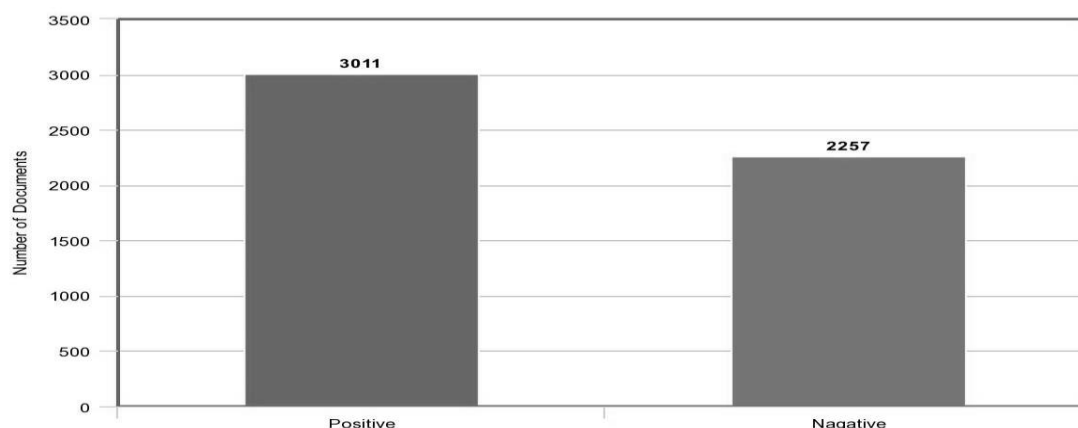


Figure 5 shows the distribution of the sentiments of all the data indexed on Solr. 3011 documents are classified as positive and 2257 as negative.

VIII. Conclusion

This is an integrated tool which has different functionalities. The processing has been done using classification which is a machine learning task. The preprocessing of the documents helps to improve the accuracy of sentiment analysis. The use of machine learning in processing of data helps in creating a dynamic model in the sense that it can be changed by changing the training set. Hence, accuracy can be improved, support for more languages can be added because of the learning ability of models. A trainable model is made to predict the sentiment of English text. Sentiment of any document can be categorized among the two categories that are negative and positive. Some other trainable models made using machine learning algorithms are domain identification, language identification and Spam ham filtering. The accuracy of these models is quite satisfactory and the scope of further improvement always remain.

IX. Future Work

The created sentiment analysis tool uses supervised machine learning. This creation depicts the efficacy of such approach and adds versatility and flexibility to text analytics in big data platform. While talking about Supervised Machine Learning, as depicted in Sentiment Analysis, the sentiments can be further classified based on human emotions like happy, sad, angry etc. Presently we have focused on polarity identification and the text is classified as positive or negative. The domains can be further increased and more specific domain identification can be obtained as a trainable model has been used. Hence the text will be narrowed down to a more specific domain. Also, multiple domains can be identified in a text unit. Support for languages other than English can be given and language specific processing tasks can be performed. This would help in getting better results for processing tasks. The model can learn based on the input dataset given. So, by giving a dataset for some other language, model can be trained to recognize that language. Sentiment Analysis is a part of decision support system. A decision support system is an information system that supports decision making abilities which results in ranking, sorting, or choosing among different alternatives. This project can be further extended to make such a decision support system which takes into consideration the sentiments involved of the users.

References

- [1] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.
- [2] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li, "User-level sentiment analysis incorporating social networks," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011, pp. 1397–1405.
- [3] S. Kamruzzaman and C. M. Rahman, "Text categorization using association rule and naive bayes classifier," arXiv preprint arXiv:1009.4994, 2010
- [4] D. Goldhahn, T. Eckart, and U. Quasthoff, "Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages." in LREC, 2012, pp. 759–765.
- [5] M. Kaur and U. Kaur, "A survey on clustering principles with k-means clustering algorithm using different methods in detail," International Journal of Computer Science and Mobile Computing, vol. 2, no. 5, pp. 327–331, 2013



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

A Brief Survey on Various Spin Protocols In Wireless Sensor Network

Sunita Kumari¹ and Rosy Verma²

School of Information Technology (SoIT)

Centre for Development of Advanced Computing (C-DAC), Noida

B-30, Sector 62, Noida, Uttar Pradesh, INDIA

Abstract: WSN is a wireless network consisting of geographically distributed independent devices, utilizing detector to watch physical or atmosphere lot. A WSN system include a gateway that give wireless connectivity back to the wired world and broadcast nodes. In this paper main focus is on SPIN protocol (Detector Protocols for information via debate) this protocol effectively disseminate news between the detector nodes. In the dissemination process individual sensor nodes check the whole network. In WSN all sensors nodes are known as sink nodes. As the nodes are operated by battery, saving energy is an important issue in the WSN. Node energy failure can affect the entire network thus to increase the life, routing within the network has have to be energy wise efficient. In order to enhance the lifespan of the network, routing must be energy efficient. This paper gives an overview of various spin protocols available for energy efficient routing in WSN.

Keywords: SPIN - Sensor Protocols for information via Negotiation, WSN - Wireless Sensor Network, ADV – Advertisement, REQ – Route Request

I. Introduction

WSN is a infrastructure less network which consists of various small sensor nodes, which calculate environment parameters for example sound, vibration, pressure and some other parameters. WSN is also used in Battlefields. SPIN is a protocol based on negotiation which contains negotiation and resource adaptation. By negotiation it means that each node in WSN communicates with other nodes before forwarding data to other node and each node contains its own resource manager that keep the record of resources used by the node. WSN can also be used for analysis, searching and checking. In WSN every mobile node senses the data from environment and processes it and then finally transfers it to the base station. These mobile nodes are essential feature of wireless network, then make the base station from where user can receive the data. SPIN protocol send the message by the method called negotiating. For this work SPIN uses metadata. First a node transmit the metadata to other nodes, if a particular node wishes to receive actual data then only it sends that data by using following three messages.

ADV: This message consume by a mobile node to advertise metadata among adjacent node.

REQ: This message is used to ask for a data.

DATA: This message used to transmit original data.

This protocol initializes working when a mobile node identifies the new data it send the ADV message storing (incomplete declaration about the data) known as metadata. If any adjacent mobile nodes are interested in the data transmit a REQ message and recover the data with the aid of DATA message to this adjacent node. After that adjacent node do again this process until all the nodes in the wireless sensor network get a copy of the data routing.

SPIN Protocol Family includes:

SPIN-PP:

SPIN-PP stands for point to point in this data sent from one point to another point in a network. SPIN protocol uses point to point network, as point to point network is very simple and its cost is also cheap. SPIN point to point protocol uses 3 steps –advertisement, request and transmission of data in each stage.

SPIN-BC:

SPIN-BC is basically made for broadcast network. SPIN-BC shares only one channel for communications. If a mobile node transmit a data on broadcast channel, then the data is received by every adjoining mobile nodes which are in a range of the node who is sending the data.

SPIN-EC:

SPIN-EC stands for energy conservation. SPIN-EC helps in reducing the energy consumption of the adjoining nodes. In start, SPIN-EC make sure that only those nodes whose energy level is more than threshold can join the network in order to upgrade the lifetime of the network.

SPIN-RL:

It expand the abilities of SPIN-BC to enhance its reliability. This protocol time to time send the ADV and REQ messages and each node keep record of it. If node requested a data from other node and data messages are not arrived to that node on time, then REQ message can be send again .So accuracy can be improved by again advertising metadata after some time.

C-SPIN

C-SPIN is an Cluster version of SPIN. This protocol stop the usage of same smallest path. . It has three stages that is ADV, REQ and DATA stage to do communication many clustering algorithms are used for effective data transmission. This method make sure the data delivery definitely. In this network is partition into clusters and each cluster communicates by their respective cluster head's and each time cluster head is also rotated randomly to save energy. Here at any point of time cluster head become dead. So that is the main limitation of the protocol.

This paper gives an overview of various different SPIN protocols in WSNfor energy efficient routing. As the main concern in WSN, is preservation of energy as nodes are moving and are run by battery. Energy is important asset in WSN. As WSN allow multiple - hop communication among mobile nodes which are run by battery, energy preservation becomes very important in-order to preserve the connectivity of the WSN network.

II. Literature survey

To perform an analysis of various existing SPIN routing protocols which work for energy efficiency in WSN, brief description of the protocols is given below.

1.Luwei Jing, and Feng Liu:In this paper author presented a routing algorithm. The routing algorithm focused on the detector protocol for sending information via debate. To resolve the problem of “blindly forward” and “data inaccessible” in detector protocol for information via debate, a new routing algorithm called detector protocol for information via debate.

2. Mohammed Salamath ,SunithaR:In this paper author developed a data centered routing protocol by modifying SPIN protocol. And make an improvement in SPIN protocol to achieve scalability and increase the lifetime of network.

3.Zeenat Rehena,Sarbani Roy and Nandini Mukherjee:This paper is present benchmark method is proposed which is also known as MSPIN.It's also compare with traditional SPIN protocol to check it's performance.

4. ChetanAmbekar, Gaurang Lakhani, Keyur Shah and KevalBhanushali:This paper proposed a technique on data aggregation the main aim of this technique to collect and combine data in energy sufficient manner so that network lifetime increased.

And this paper also discussed a brief description of sending temperature of source node to destination node to decrease cost of fewer faults.

5 ChintaChokshi&Bijon Desai:This paper reduce the work of stuffy to choose a protocol based on requirements. In this paper a technique is proposed based on combination of data centric and hierarchical method to simplify the task of selecting a routing protocol based on achievement and scalability.

III.Compression Table

Protocol Name	Advantage	Disadvantage
SPIN	SPIN protocol is high energy level based protocol and more effective compare to other protocol in a wireless sensor environment.	In SPIN protocol data advertisement cannot secured delivery of data thus does not suitable in applications where reliability is required.
I-SPIN	I-SPIN is a negotiation process and based on three ways: data broadcasting, data requesting, and data transmission.	In I-SPIN many energy wastage in data broadcasting time.
M-SPIN	This paper is present benchmark method is proposed which is also known as MSPIN.It's also comparing with traditional SPIN protocol to check its performance.	The dissemination of data in the network through SPIN protocol takes long time. Some nodes used multiple times which looses energy

S-SPIN	Secure-SPIN protocol has all the features which described above but in S-SPIN to build up the life time of the network and to achieve scalability, dead nodes includes in cluster heads are dynamically replaced by using recovery algorithm.	In S-SPIN network life time is increase but energy wastage increase some time.
--------	---	--

IV. CONCLUSION AND FUTURWORK

This mainly focuses on various efficient routing algorithms based on SPIN protocol via negotiation technique in WSN. The main aim of this paper is to minimize power consumption of the detector node and enhance the lifetime of the network. In future this work the transmission time of the node can be reduced for disseminating data in the wireless detector network. So that the energy consumption can be further minimize and also maximize accordingly to increase the lifetime of sensor network.

V. Reference

1. Ian F. Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci "A Survey on Sensor Networks" Communication Magazine, IETE Volume: 40, Issue: 8, 2002.
2. Zeenat Rehena, Sarbani Roy, Nandini Mukherjee, "A Modified SPIN for Wireless Sensor Network," Communication System and Network (COMSNETS), 2011 Third International Conference, 2011 IEEE.
3. Luwei Jing, and Feng Liu, "Energy Saving Routing Algorithm Based on SPIN Protocol in WSN ," IEEE Communications Magazine 169-185, 2011.
4. Mohammed Salamath, Sunita .R , "Analysis of the existing SPIN protocol for data dissemination," International Journal of Science and Research (IJSR), Volume 5 Issue 3, March 2016
5. Mohammed Salamath, Sunita .R," A Survey on SPIN Protocol in Wireless Sensor Network" ,International Journal of Science and Research (IJSR) 2014
6. Holger Karl, and Andreas Willig, "Protocol and Architecture for Wireless Sensor Networks," Wiley Publication, 2005.
7. Geetu and Sonia Juneja, "a study on a routing protocol SPIN in WSN", 2012 vol 5, no 2, pp 345-352, IJITM.
8. Kazi Chandrima Rahman, "A Survey on Sensor Network", JCIT, ISSN 2 078-5828 (PRINT), ISSN 2 21 8-52 24 (ONLINE), 1(1), manuscript code: 100715, 2010



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

A Survey of Centralized Key Management Schemes for Secure Multicast Communication

Ekta Garg¹ and Vinod Kumar²

School of Information Technology

Centre for Development of Advanced Computing, Noida

Uttar Pradesh, India

Abstract: Many network applications require transmission of data to all the group members and maintaining security in large and dynamic groups is one of the major challenges for secure multicast communication. This feature is not provided in IP multicast which leads to the need of secure key management. The group key management schemes are divided into three categories: Centralized, De-centralized and Distributed schemes. This paper mainly provides a survey on various existing centralized key management schemes which are applicable in different areas. In addition, the comparison of existing schemes is done based on performance and security parameters. Finally, we discuss some new research directions in the area of key management for secure multicast communication.

Keywords: Key Management Schemes, Multicast Communication, Cryptography, Security, Secure Group Communication

I. Introduction

The tremendous growth in Internet over last few decades is commendable which plays vital role in development of many new technologies, applications and services provided over IP. With this rapid growth, various new applications like iPay- TV, Video on Demand (VoD), Collaborative work and e-learning etc. came into picture. They not only need normal data to be sent via Internet but many sensitive data are being exchanged over it which needs more secure and confidential environment. This secure environment is referred to as Secure Group Communication (SGC) where data is being recovered by intended users only. Thus limiting access to data is done through encryption and decryption of messages. For this process we need keys to encrypt and decrypt the data and these keys are known as Cryptographic keys. Therefore it is difficult for any intruder who is not part of the group to decrypt the message. In group communication this common key, which is used for communication is known as Group Key.

The term key management deals with distributing key securely to all the users and managing situations of user join and leave operations where there is need to change in the group key to maintain security. In addition to providing security, we need to handle scalability which plays major role in real time applications. However, distributing group key to legitimate users need to maintain forward secrecy and backward secrecy both [1]. Henceforth, re-keying a group key after member leaves is must because we need to produce a new group key as old group key is known to leaving member and that cannot be used for further communication, this supports Forward Secrecy. Similarly whenever a new member joins the group, there is need to change the group key so that new member cannot access old data, this provides Backward Secrecy.

Key management entity should preserve security with least computational and communicational cost so as to handle dynamicity efficiently. In order to support such computations at server or user, there are pre-defined architectures based on different requirements and functionalities. Group key management has been further divided into three broad categories: Centralized key management, distributed key management and de-centralized key management[2]. In Centralized key management, single entity is responsible for controlling the whole group, hence these protocols seek for minimum computation and storage capacity at server's end. Whereas, Distributed Key Management focuses on group key generation with contribution of all the members i.e. members themselves are responsible for Group Key generation. Whereas, in de-centralized key management, larger groups are further divided into sub groups so as to minimize the load at single point (server) and removed the single point of failure problem. All these different categories have their own merits and demerits and are used according to the requirements of any application. But, they should maintain proper security even if there is any change in network. This paper has covered Taxonomy of Key Management in Section II explaining intricate details of key management protocols with slight explanation of different architectures. Section III presents an extensive survey of centralized key management protocols with comparative analysis on performance and security basis. In addition, current research directions have been explained briefly in Section IV and then conclusion by end of this paper is provided in Section V.

II. Taxonomy of Key Management

Key management is a vital element of Secure Group Communication (SGC) which ensures access control of group key and establish proper channel for mass communication. The term key management deals with many actions that are handled by group key manager [3]. These actions and functions can be explained in detailed as follows:

- *User Authentication*: This deals with authenticity of member, whether user is genuine user or not. Various authentication mechanisms are used for this process and assure validity of users.
- *Initialization and Key Generation*: Key initialization refers to the sharing of private keys to all the respective members. Similarly after initial join operation, Group key is calculated at Server.
- *Key Distribution*: After key generation, it is being distributed over channel to whole group.
- *Re-Keying*: Whenever there is change in group formation, there is need to change the group key in order to maintain security. This process is known as re-keying.

Even though above functionalities are being handled judiciously by Key Manager, the risk of breach in security is a major issue. Henceforth, key management not only deals with above functionalities but pays attention to various security aspects of group communication. These aspects can be defined in terms of Forward Secrecy, Backward Secrecy, Key Independence, Collusion freedom.

- *Forward Secrecy*: To maintain secrecy in group communication when a member leaves a group. Old member after leaving must not see future discussion. This is termed as Forward Secrecy.
- *Backward Secrecy*: To maintain secrecy in group communication whenever a member joins a group. Newly joined member should not be able to view past communication, is termed as Backward Secrecy.
- *Key Independence*: Keys should not be related to each other so that if one key is leaked, the rest cannot be determined easily by any attack possible.
- *Collusion Freedom*: Already left members share their piece of information and tries to regain access to the group key. This is highly undesirable situation.

Key Management protocols are divided into two main categories: Network Dependent schemes and Network Independent Schemes [4]. Network Independent Schemes are further divided into three main categories: Centralized, De-centralized and Distributed key management schemes discussed in previous section whereas Network Dependent schemes are classified into Cluster based and Tree based.

- *Network Independent Scheme*: The schemes that are not dependent on architecture of network are known as Network Independent scheme. Since, they are independent of features of network; they can work in both wired and wireless environment.
- *Network Dependent Scheme*: Network dependent protocols are dependent on the features of the underlying network foundation for them to operate effectively. Network dependent group key management protocols support mobile multicast communication where communication is unaffected even when members moves around over widespread area.

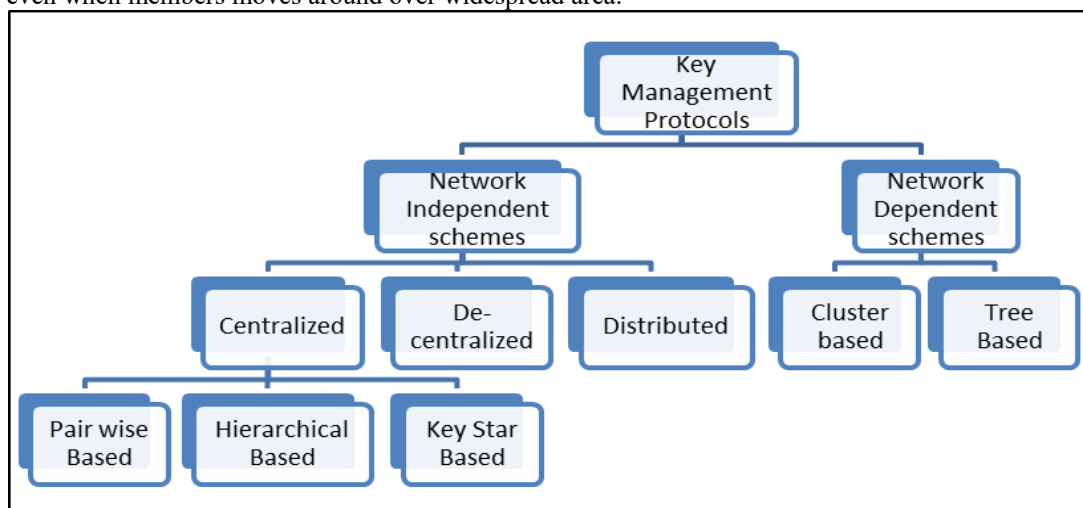


Figure 1 Key Management Protocols

III. Literature Survey

This section describes the rigorous literature survey on centralized key management schemes. In centralized key management systems, server alone is responsible to carry out whole process of group communication securely. This single entity manages key generation, key distribution and authentication of all the members. Some challenges faced by centralized scheme are as follow and there is need to minimize these overhead cost:

- *Scalability overhead*: Overall success of group communication depends completely on the single centralized entity; therefore the task of rekeying becomes an overhead when the group size increases. Therefore, there is need to reduce scalability overhead for better performance.
- *Storage overhead*: The number of keys required to store by each member increases with increase in size of group which is not advisable.
- *Computation overhead*: As single entity server is doing all the work, the number of operations done at server must be in limit that it should not affect the communication.
- *Communicational overhead*: The parameters that are used to convey Group key in encrypted form must be minimized to reduce bandwidth overhead.

In last few decades, many schemes have been proposed and developed, some algorithms aimed at reducing cost at server's end and some focussed on less computation at user's end. Guang-Huei Chiou and Wen-Tsuen Chen [5] proposed secure locking concept based on Chinese Remainder Theorem (CRT). This scheme was used to solve the secure broadcasting problem at that time. Even with less computation, it was not meant for large groups and hence consumed more time for key generation in case of high scalable situation. Muckenhirn and Harney [6] proposed the Group Key Management Protocol (GKMP) where Key Distribution Centre (KDC) creates Group key and sends it to respective members. However, same copy of group key is distributed even if member leaves the group, thus forward secrecy is not maintained. Wong et al. [7] and Wallner et al. [8] proposed the concept of Logical Key Hierarchy (LKH) for Group key generation. In this approach, leaves of tree correspond to members of the group and internal nodes denote intermediate keys which are used for decryption of group key stored at root of tree virtually. Even though scalability is somehow managed but increase in number of members increases the number of keys required to be stored by member. Thus, storage cost is very high for large groups because $((\log_2 n) + 1)$ keys are required to store at one user.

An enhancement in the hierarchical binary tree approach is a one-way function tree (OFT) [9] and is proposed by McGrew and Sherman. This protocol reduced the size of the rekeying messages from $2(\log_2 n)$ to only $(\log_2 n)$ and message size was reduced too. Waldvogel et al. [10] improved their proposed scheme to change the hierarchical tree structure for a flat table (FT) known as Centralized Flat table Key Management (CFKM). This scheme decreased the number of keys held by the Server. But, it was prone to collusion attacks. A set of evicted members, may combine their sets of keys to recover a group key, hence were able to have unauthorised access to group communication. Xinliang Zheng et al [11] introduced new centralized group key protocols based on the CRT method i.e. Chinese Remainder Theorem based Group Key management (CRGK). Even though the load was completely shifted to server but they optimized the number of re-key broadcast messages, user-side key computations and number of key storages. They used Extended Euclidean algorithm for reducing complex computations compared to other protocols. The only disadvantage of this protocol is that load at server is increased drastically with increase in number of members.

Iuon-Chang Lin et al [12] proposed a new RSA based multicast key management scheme to solve the problem of rekeying. This protocol applies a star-based architecture i.e. Start Based Management Key (SBMK) to eliminate the rekeying processes and provided the good performance when the membership changes in a multicast group. In hierarchy based, each member has to hold $\log_2 n$ keys but in star based scheme each member holds only one key. Main demerit of this scheme is increase in computational load on the server. J.A.M. Naranjo et al [13] presented a scheme known as Extended Euclid Algorithm based Protocol (EEAP) to securely distribute a group secret to a set of receivers with only one multicast communication. The Extended Euclidean algorithm provides a fast pace solution to the problem of finding the Greatest Common Divisor (GCD) of two large numbers which is used to distribute key privately. It is suitable for all topologies. But numbers of broadcast parameters are more while sending group key and computational and storage requirements at server is high too.

P. Vijayakumar et al [14] focused on reduction of computation complexity by performing fewer multiplication operations (using an existing Karatsuba divide and conquer approach) and reducing amount of information stored in Group Centre (GC) and group members. This scheme GCD based key distribution Protocol (GCDP) broadcasts 5 values to all users which increases load while transmission of group key. Moreover, calculations at group member's end are complex. The authors P. Vijayakumar et al [15] proposed a Chinese remainder theorem-based group key management (CRTGKM) scheme that significantly reduces computation complexity of the key server. While rekeying only 1 addition and subtraction operation is done but scalability is the major issue with this protocol.

Table 1 depicts performance analysis of existing schemes or protocols that have been discussed so far. Computational cost at Key Server (KS) and users area is analyzed. Also, Storage cost at user area and KS area is explained briefly in addition with communicational cost. Here, 'n' is Number of users, 'A' is cost of Addition operation, 'M' is Multiplication operation cost, H is Hash function operation cost, 'E' is cost of Encryption, 'D' is decryption cost, 'EEA' is cost of Extended Euclid Algorithm to calculate multiplicative modular inverse, 'gcd' represents GCD operation cost, 'mod' is Modular operation cost, 'exp' is Exponentiation cost, 'EEA'' is cost of Extended Euclid algorithm for calculating GCD, 'modinv' is Modular inverse cost, 'S' is Subtraction cost.

Table 1 Performance Analysis

Parameters Schemes	Computation cost (KS)	Computation cost (user)	Storage complexity (user)	Storage complexity (KS)	Communication cost
Secure Locks	$O(1)(A \text{ or } S)$	1 mod	2	$4n + 3$	1 broadcast
GKMP	$2E$	$2D$	2	N	1 broadcast
LKH	$\log_2 n H + 3 \log_2 n E$	$(\log_2 n + 1)D$	$\log_2 n$	$2n-1$	$2 \log_2 n - 1$ broadcast
OFT	$(\log_2 n + 1)H + \log_2 n XOR + 3 \log_2 n E$	$(\log_2 n + 1)D + \log_2 n(H + XOR)$	$(2n - 1)$	$(\log_2 n + 1)$	$2 \log_2 n + 1$ broadcast
CFKM	$2E$	D	$(2n-1)$	$(\log_2 n + 1)$	2 broadcast
CRGK	$O(n)(XOR + A + M + EEA)$	1 mod + 1 XOR	2	$2n + 1$	1 broadcast
SBMK	$O(n)(M + \gcd + EEA + 2 \text{ mod}) + 1 \text{ exp} + 1 \text{ mod}$	1 exp + 1 mod	2	$2n$	1 broadcast
EEAP	$O(n)M + 2 \text{ exp} + 2 \text{ mod} + 1 EEA'$	1 modinv + 2 exp + 3 mod	2	N	3 broadcast
GCDP	$O(n)(M + EEA' + 1 \text{ exp} + 1 \text{ mod})$	1 M + 1 S + 1 mod + 1 exp	2	N	5 broadcast
CRTGKM	$O(1)(A \text{ or } S)$	1 mod	2	$4n + 3$	1 broadcast

Table 2 shows security analysis of already existing schemes in terms of security parameters that are used frequently. These parameters are Group Secrecy (which maintains information should be available to intended users only), Forward Secrecy, Backward Secrecy, Key Independence, and Collusion Freedom.

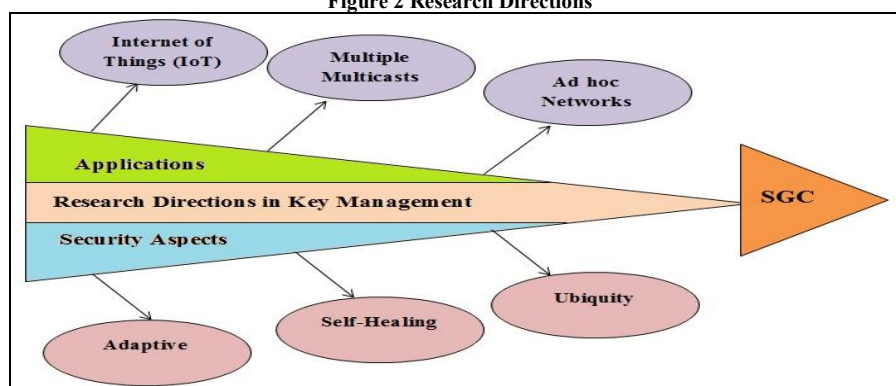
Table 2 Security Analysis

Parameters Schemes	Group Secrecy	Forward Secrecy	Backward Secrecy	Key Independence	Collusion Freedom
Secure Locks	Yes	No	No	Yes	No
GKMP	No	No	Yes	Yes	Yes
LKH	Yes	Yes	Yes	Yes	Yes
OFT	Yes	Yes	Yes	Yes	Yes
CFKM	Yes	Yes	Yes	Yes	No
CRGK	Yes	Yes	Yes	Yes	Yes
SBMK	Yes	Yes	Yes	Yes	Yes
EEAP	Yes	Yes	Yes	Yes	Yes
GCDP	Yes	Yes	Yes	No	Yes
CRTGKM	Yes	Yes	Yes	Yes	Yes

IV. Current Research Directions

The protocols studied so far works best in Areas like Pay-Tv, video-conferencing, video on demand (VoD), database replication, broadcasting stock quotes, e-learning, software updates where re-key operations are not that fast and connection is steady. But in environment where network dependency is up to maximum level, there is need to have such protocols that are adaptive, distributive and reliable with some unique properties like mobile- based and self- healing. The vital function of self-healing property is that users are able to recover misplaced group keys on their own, without requesting additional transmission from the group key server. Thus, extensive research is going in this direction [16].

Figure 2 Research Directions



With emergence of different group based services, multiple multicasts exist where single user subscribes to multiple groups. Broadcasting data in such insecure environment makes network more attack prone which leads to access control mechanism. These mechanisms must ensure confidentiality, accurate delivery, perfect content transmission and reliability. In addition, old security parameters like forward secrecy, Backward Secrecy etc. should be satisfied. Most common example of multiple multicasts is service provider providing three distinct multicasts such as information service, telematics services and TV streaming. In this example, service provider must not suffer from key management overhead due to multiple tasks. Handling such situation with at most security is hot topic. Researchers are working in this direction and new protocols are being introduced but there is no perfect answer to it.

Similarly, various new technologies like Internet of Things (IoT), Ad-hoc networks, etc. are emerging suddenly and novel schemes are required that can handle such situations prominently. The term ubiquity is relatable with IoT as the devices linked in this environment can be operated always from anywhere at any point of time. Since, devices used in IoT are resource constrained (like less memory space, poor battery backup etc.), there is need of light weight solution for secure communication. Authentication plays vital role in such situation as any malicious node can try to access our environment, therefore proper authentication schemes are required that meets all the constraints and can be able to authenticate the network as well. This two way authentication is known as mutual authentication where network not only authenticates users but users are able to find unreliable network while communication.

Likewise Ad-hoc networks are most commonly used technology in military and disaster recovery situations, in environment handling, in daily patient care at advanced hospitals or at homes. With such crucial applicability, security is fundamental part of it and key management provides overall security in form of authenticity, confidentiality and integrity. Since, ad-hoc is wireless in nature, it deals with complex situation like topology change i.e. dynamic connection among nodes and scalability problem as we are unaware of exact group size. As these nodes run on battery, they need efficient energy consumption protocols with self-healing property.

Even in applications like Smart grids where delivering power reliably with improved efficiency is main target but ignoring potential threat is highly undesirable. Therefore, devices used at user end like smart meters and corresponding infrastructure which provides information must be properly secured. For securing the communications in a smart-meter infrastructure, also referred as Advanced Metering Infrastructure (AMI), we require to combine different security technologies. Among these technologies, encryption is a crucial technology, which allows one to transmit data securely across the AMI and authenticate the different parties involved in this interaction. Even though, we have many existing encryption schemes for group key distribution but scalability is major issue as it involves millions of devices. Thus, key management in this technology needs a different architecture that can accommodate such drastic changes in group while communication.

V. Conclusion

In this paper, we have discussed the role of key management, which handles secure distribution of group key to all the members. Key management ensures how efficiently a group key can be calculated and also deals with the concept of authentication which is generally taken as pre-requisite in many existing schemes. In addition to efficiency, various different security aspects were discussed for efficient secure communication. Rigorous analysis of few existing schemes of Centralized key Management architecture is done on the basis of performance and security parameters. Our analysis shows that there is no perfect scheme that can cover all such parameters. This leads to need of such schemes that covers all areas like reduced computation cost of key server and user, minimized communicational cost and optimized storage cost for server and user both. We have also presented diverse research directions in this field and what is in demand nowadays. These discussed current trends can be further explored by other researchers as this topic demands for rigorous research and in depth understanding for better schemes.

References

- [1] Sandro Rafaei and David Hutchison, "A Survey of Key Management for Secure Group Communication", ACM Computing Surveys, Vol. 35, No. 3, September 2003, pp. 309–329
- [2] YacineChallal, HamidaSeba, "Group Key Management Protocols: A Novel Taxonomy", World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol. 2, No. 10, 200
- [3] YacineChallal, AbdelmadjidBouabdallah, HamidaSeba, "A Taxonomy of Group Key Management Protocols:Issues and Solutions", World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:1, No:6, 2007
- [4] R. Seetha, R. Saravanan, "A Survey on Group Key Management Schemes", Cybernetics and Information Technologies, Vol. 15, No. 3
- [5] Guang-HueiChiou and Wen-Tsuen Chen, "Secure Broadcasting Using the Secure Lock", IEEE Transactions on Software Engineering Vol. 15 No. 8 August 1989
- [6] Harney, H. and Muckenhirn, C. 1997b, "Group Key Management Protocol (GKMP) Architecture", RFC 2094.
- [7] Wong, C. K., Gouda, M. G., And Lam, S. S. 2000, "Secure group communications using key graphs", IEEE/ACM Trans. Netw. 8, 1 (Feb.), 16–30.
- [8] Wallner, D., Harder, E., and Agee, R. 1999, "Key Management for Multicast: Issues and Architectures", RFC 2627.
- [9] McGrew, D. A. and Sherman, A. T. 1998, "Key establishment in large dynamic groups using one way function trees", Tech. Rep. No. 0755 (May), TIS Labs at Network Associates, Inc., Glenwood, Md.

- [10] Waldvogel, M., Caronni, G., Sun, D., Weiler, N., and Plattner, B. 1999, "The VersaKey framework: Versatile group key management", IEEE J. Sel. Areas Commun. (Special Issue on Middleware) 17, 9 (Aug.), 1614–1631.
- [11] Xianliang Zheng, Chin-Tser Huang, Manton Matthews, "Chinese Remainder Theorem Based Group Key Distribution", ACMSE 2007, March 23-24, 2007, Winston-Salem, North Carolina, USA.
- [12] Iuon-Chang Lin, Shih-Shan Tan, Chung-Ming Wang, "Multicast Key Distribution without Rekeying Processes", The Computer Journal, Vol. 53 No. 7, 2010 (Oxford University Journal).
- [13] J.A.M. Naranjo, N. Antequera, L.G. Casado, J.A. López-Ramos, "A suite of algorithms for key distribution and authentication in centralized secure multicast environments", Journal of Computational and Applied Mathematics 236 (2012) 3042–3051.
- [14] P. Vijayakumar, Sudan Bose, Arputharaj Kannan, "Centralized key distribution protocol using the greatest common divisor method", Computers and Mathematics with Applications 65 (2013) 1360–1368.
- [15] P. Vijayakumar, Sudan Bose, Arputharaj Kannan, "Chinese remainder Theorem based centralized group key management for secure multicast communication", IET Information Security, 2014, Vol. 8, Issue 3, pp. 179–187
- [16] R. Siva Ranjani, Dr.D.LalithaBhaskari, Dr.P.S.Avadhani, "Current Trends in Group Key Management", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 11, Nov 2011



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

PairTester: Pairwise Generation of Test Cases in presence of constraints

Reetika Gupta¹ and Neha Bajpai²

School of Information and Technology (SoIT)

Centre for Development of Advanced Computing (C-DAC)

NOIDA, U.P, INDIA

Abstract: Pairwise testing is an effectual combinative test case generation method that can generate comparatively minimum number of test suite to cover at least all pairs of parameters values once. For an application under testing with m parameters, where every parameter has different values, combinative testing leads to expose the faults provoked by interacting parameters. It has been observed that the difficulty level of finding interaction between the parameters used in a program that need to be tested is high. In this paper a new tool have been introduced called PairTester using Modified In-Parameter-Order Algorithm, it emphasizes on pairing up those parameters values which depends upon each other and leads to a program to act in different way. It also generates the optimum number of test cases. Identification of interacting parameters is important as it leads to a system to behave in different manner. The interaction in the application can be identified with the conditional statements present in the program. The main focus of this approach is to find out the interaction between the parameters and to generate the minimum number of test cases.

Keywords: Combinatorial Testing, Pairwise Testing, Modified In Parameter Order Algorithm, PairTester

I. Introduction

Software testing is a process of exercising a program or application with the purpose of finding software errors. Software testing can also be defined as the procedure of validating and verifying that a software program/application/ product: Satisfies the user requirements and technical requirements that are mentioned in its design phase.

Testing is a process used to identify the correctness, completeness and quality of developed computer software. There are many types of testing techniques available now days. This project focuses on a type of combinatorial/combinative testing named as pairwise testing. It is from one of the black box testing that focuses on the different outputs and errors generated by a set of inputs and execution of the system. It is also called functional testing which take less time as compared to all the other techniques. Practical system testing often has constraints/conditions on the combination of parameter values. With the help of PairTester there is no need of defining constraints manually like in previous approaches. It extracts those parameters automatically which are dependent on each other. In Modified In Parameter Order Algorithm, there is a slight change from basic algorithm in order of taking inputs. Basic algorithm chooses initial two parameters randomly, but PairTester starts with those two parameters having more dependency.

The rest of the paper is organized as follows; Section2 Literature Review. Section3 Proposed Approach. Section4 Result. Section5 Conclusion & Future Work.

II. Literature Review

Various algorithms have been used in the sector of pairwise testing like Search Based Algorithms, Meta-Heuristic Algorithm, Orthogonal Array and In Parameter Order etc. This section briefly describes and compares the techniques used by different researchers with their conclusion. Some of them are described in this section.

Yu Lei & K.C Tai uses basic In Parameter Order Algorithm for testing an Online Registration form written in Thai language to generate every possible combination of input parameters. Secondly they have used Equivalence Partitioning to reduce the values of parameters. They have compared their result in terms of number of test cases, time and percentage of combination coverage. Their result successfully shows that pairwise testing is more efficient in terms of test cases and time but it covers only 80% of combinations. [1]

Shiwei Gao, Binglei Du ,Yaruo Jiang Jianghua Lv and Shilong Ma have worked on Modified Version of In Parameter Order algorithm, they have considered the constraints defined by the tester . Now, the tester converts them into logical expressions manually. After that, test cases are generated in the presence of constraints with

the help of a tool called SAT Solver. They have compared their approach with replace method and the result shows that this proposed approach is better than replace method in terms of test cases. [2]

Abdullah B. Naseer, AbdulRahman A. Alsewari & Kamal Z. Zamli shows a comparison of search based algorithms like Hill Climbing, Simulated Annealing, Genetic Algorithm, Ant Colony algorithm (ACO), Particle Swarm Optimization (PSO) & Harmony Search for generating test suites pairwise. Author's analysis shows all positive and negative points of every algorithm. According to them initialising parameters may lead to optimum solution via any search based algorithm. [3]

Hasneeza L. Zakaria & Kamal Z. Zamli have proposed two approaches; the first is the use of basic Migration Bird Optimization (MBO) algorithm, called Pairwise MBO (PMBO) Strategy and the second one is improved Pairwise Migration Bird Optimization algorithm (PMBO). The improved PMBO enhances the PMBO as it considers the multiple neighbour structure and elitism. According to their conclusion improved PMBO gives optimum solution when the test size generated is relatively small. [4].

Rongzhi Qi, Zhijian Wang and Ping Ping & Shuiyan Li proposed a hybrid approach for generating test suites. In this paper to enhance the capability of Genetic algorithm they have augmented genetic algorithm and two-stage hill climbing. In this the initial stage it improves every individual after applying genetic operations & in other stage improves the best solution of current generation at the end. [5].

A. Comparative Analysis

Title	Work Detail	Analysis
Paper 1 Pair-Wise Testing Applied with Online Registration. [1]	Compares number of test cases generated via manually & pairwise testing.	By using equivalence partitioning and IPO it becomes possible to minimize the number of test cases for a web page of online registration.
Paper 2 An Efficient Algorithm for Pairwise Test Case Generation in Presence of Constraints. [2]	In this paper, modified IPO algorithm is used for handling constraints.	Constraints/Conditions need to be specified by tester before test case generation, constraints are those values combinations that are not allowed to be appear in test suite.
Paper 3 Adopting Search-Based Algorithm for Pairwise Testing. [3]	Comparison of search based algorithm for generating the pairwise test suite.	Search Based algorithms have been used in this paper with two different types of solutions: single and population based. Population based solution is more expensive as compared to single based solution. Single based solutions stuck in local optimum solution whereas population based solution take more time.
Paper 4 Migrating Birds Optimization Strategies for Pairwise Testing. [4]	MBO approach is used which is motivated by the flying pattern of bird.	PMBO works better than any heuristic algorithm Improved PMBO works well when number of parameters are less
Paper 5 A Hybrid Optimization Algorithm for Pairwise Test Suite Generation. [5]	An hybrid approach obtained to improve the genetic algorithm.	The advantages of hill climbing and genetic algorithm are mixed together, to generate pairwise test suite.

Table1: Comparative Analysis

III. Proposed Approach

Pairwise testing is more useful and beneficial where the number of combinations is more and the output of a system differs when two or more fields interact with each other. For these two points a new approach has been introduced in this paper because the survey has identified some problems in the previous approaches.

Those problems are given below:-

- 1) There were lots of manual work in terms of finding parameters and their values from an application.
- 2) To use the existing tool you have to enter manually all those parameters and their values which was very time consuming and difficult if number of parameters are more.
- 3) There were no criteria of automatically detecting the constraints present in applications.
- 4) Previous technique was only used on the design of application. They hadn't considered any programming code for testing.

Following new proposed approach overcome all those above mentioned problems:

Begin

{take a java source code as input}

If (conditional statements: if-else/if/for/while/do-while present)

then,

{Identify the parameters q1,q2,..., qm,}

else

exit;

If (number of parameters ≥ 2) then

{Identify the dependent variables}

else

exit;

for(parameters q1,q2,...,qm)

begin

{Modified In Parameter Order Algorithm}

end

end

Algorithm: PairTester

Pair Tester tool is used to test the java applications in short time, check out the dependencies between the parameters.

To illustrate the functioning of above approach we take an example of Registration form applying for passport if the fields are correctly filled.

The screenshot shows a web form for passport registration. It has two main sections. The left section contains input fields for 'Passport Office' (with a dropdown), 'Given Name', 'Surname', 'Date of Birth' (DD/MM/YYYY), 'E-mail Id', 'Login Id', 'Password', 'Confirm Password', 'Hint Question', 'Hint Answer', and a CAPTCHA. The right section contains radio buttons for 'CPV Delhi' and 'Passport Office', a 'Select the CPV Delhi' option, a 'Tibetan Refugees' note, a 'First Name + Middle Name' field, and a 'Check Availability' button. At the bottom, there is a 'Register' button and a 'Clear' button.

Figure1: Registration form applying for passport

Step1) The first step of this approach is to take a java application source code as input; PairTester will take source code of the Online Registration form applying for passport showed above.

Step2) Now to check whether pair wise testing can be applied on a given source next step will be executed i.e, check for conditional statements. Considering this step is important to avoid the question of why pair wise testing

only. So the answer is, the programs where conditional statements are present pairwise testing are more efficient to use in terms of cost and time. For this example we have conditional statements so we can proceed further.

Step3) Now this step will search for the parameters, the type of values they can accept (data types) and their values if defined in program.

Step4) This step will find out the parameters which are dependent on each other by any means. For example “Passport office field” selection is dependent upon the selection of “Passport Office radio button” selection. Second dependency is between the selection of “yes” or “no” parameter & “login id”.

Step5) So the pairing of parameters starts with Passport Office field and Passport Office drop down menu then, “yes” or “no” parameter & “login id”. Afterwards rest of the parameters are paired up according to the algorithm.

	PASSPORT_OFFICE	PASSPORT_OFFICE_LIST	GIVEN_NAME	SURNAME	EMAIL_ID	LOGIN_ID	PASSWORD	YES	NO	CONFIRM_PASSWORD	DATE_OF_BIRTH
1	unchecked	delhi	123	tyz	ree@gmail.com	ree@gmail.com	76e54	unchecked	check	3456766#5	31/9/89
2	unchecked	chandigarh	123	-	trrer	ree@gmail.com	76e54	check	unchecked	3456766#5	31/9/89
3	unchecked	-	abc	tyz	trrer	ree@gmail.com	3456766#5	check	check	3456766#5	24/02/92
4	check	-	abc	tyz	ree@gmail.com	trt@gmail.com	3456766#5	check	unchecked	3456766#5	31/9/89
5	check	delhi	abc	-	trrer	trt@gmail.com	3456766#5	unchecked	check	764654	24/02/92
6	check	chandigarh	abc	tyz	ree@gmail.com	trt@gmail.com	3456766#5	check	unchecked	764654	24/02/92
7	check	-	123	-	ree@gmail.com	trt@gmail.com	76e54	unchecked	unchecked	764654	31/9/89
8	check	delhi	123	-	ree@gmail.com	ree@gmail.com	3456766#5	check	unchecked	764654	24/02/92
9	unchecked	chandigarh	abc	tyz	ree@gmail.com	trt@gmail.com	76e54	unchecked	check	764654	24/02/92

Figure2: Test Cases

A modified version of in parameter order algorithm has been introduced in this paper. So this basic algorithm can be rewritten as:

Modified IPO

Assume there is a system Q having various parameters. Arrange them according to their dependencies from q1, q2..., qm,

$m \geq 2$. Below given is the framework of IPO algorithm for producing a pairwise test set suite M for Q

Algorithm Modified In Parameter Order

begin

{for the initial 2 parameters q1 and q2}

$M = \{(r1, r2), \text{where } r1 \text{ and } r2 \text{ are values of } q1 \text{ and } q2 \text{ respectively}\}$

If $m=2$ then exit;

{for rest of the parameters}

for parameters $q_i, i=1,2,3,\dots,m$ do

begin

{Horizontal growth}

For every test $(r1, r2, \dots, r_{i-1})$ in M do

Change it with $(r1, r2, \dots, r_{i-1}, r_i)$,

Where r_i is a value of q_i ;

{Vertical growth}

While M does not cover all pairs between q_i

and each of $q1, q2, \dots, q_{i-1}$ do

add a new test for q_1, q_2, \dots, q_i to M ;

end

end

The point need to be noted here is that we have only changed in the criteria of choosing the initial parameters, the rest of two algorithms horizontal growth and vertical growth works as same as in the earlier approaches .

Step6) The final step of the approach is to compare the result given by the new approach with the existing one.

As we have done with the help of ACTS tool which uses the basic In Parameter Algorithm.

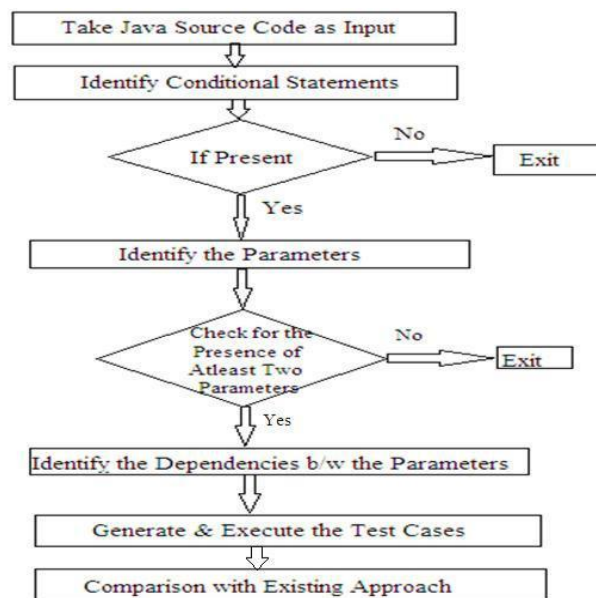


Figure 3: Flow graph of PairTester

IV. Result

In this paper we have successfully tested 20 java applications with optimum number of test cases and compared with the few existing tools available as open source. Main points under consideration are the number of test cases, work nature of PairTester, characteristic of application & dependent parameters. PairTester shows us that it gives less number of test cases in most of the applications. In the above example of applying for passport, test cases are reduced from 16 to 9. Sometimes equal number of test cases is generated by PairTester and Existing tool where there is no possibility of reduction in number of test cases. The second aspect is work nature of PairTester, it has made the testing automated. It saves time in terms of choosing whether the application is suitable for pairwise testing or not. Third aspect of PairTester is it possible to test whether an application can be tested with this approach or not. The fourth and very important feature of PairTester is dependency finder which finds out the dependent parameters. It gives priority to dependent parameters to pair up first with all their possible values. This criterion is more helpful in pairing those interacting parameters due to which system performs in different way.

V. Conclusion & Future Work

In this paper, an existing IPO algorithm is modified and introduces a new tool (PairTester) used to test the java application in short time & check out their dependencies. As dependencies are the major concern of this proposed approach, hence the parameters having interactions in the scale of maximum to minimum is identified and paired up to obtain the desired results. 20 java applications have been tested with PairTester and their results

shows that, this new tool is giving less number of test cases with maximum code coverage. For the future purpose, 3 or more pair testing can be used for better results. With some modification according to the need of

pairwise testing some new algorithms like firefly and cuckoo algorithm can be used for pairwise testing for further exploration. With some modifications, the applications written in different programming languages like C, C++, and ASP.Net etc can be tested. There are again chances of modification in this algorithm where you can define the type of dependency automatically without defining them manually.

References

- [1] Wasan Uthailang, Supaporn Kiattisin and Adisorn Leelasantham "Pair –Wise testing Applied with Online Registration" The 4th Joint International Conference on Information and Communication Technology, Electronic and Electrical Engineering , 2014 IEEE.
- [2] Shiwevi Gao,Binglei Du,Yaruo Jiang,Jianghua Lv and Shilong Ma, "An efficient algorithm for pairwise test case generation in presence of constraints". 2nd International conference on System and Information , pp.406-410, 2014 IEEE.
- [3] Abdullah B.Naseer, AbdulRahman A.Alsewari & Kamal Z.Zamli, "Adopting Search-Based Algorithm for Pairwise Testing". 4th International Conference on Software Engineering and Computer Systems,2015 IEEE.
- [4] Hasneeza L. Zakaria , Kamal Z. Zamli , " Migrating Birds Optimization based Strategies for Pairwise Testing " ,9th Malaysian Software Engineering Conference, Dec. 2015.
- [5] Rongzhi Qi, Zhijian Wang and Ping Ping, shuiyan li, "A Hybrid Optimization Algorithm for Pairwise Test Suite Generation", International Conference on Information and Automation Lijiang, China, August 2015 IEEE.
- [6] Wu H. Y., Nie C. H., Hareton L.,Colbourn C. J. "A Discrete Particle Swarm Optimization for Covering Array Generation". Evolutionary Computation, IEEE Transactions on, 2014, PP(99).
- [7] B. S. Ahmed, K. Z. Zamli, and C. P. Lim, "Constructing at-way interaction test suite using the Particle Swann Optimization approach," International Journal of Innovative Computing, information and Control, vol. 8, pp. 431-451,2012.
- [8] Linbin Yu, Yu Lei, Mehra Nourozborazjany, Raghu N. Kacker, D. Richard Kuhn,"An Efficient Algorithm for Constraint andling in Combinatorial Test Generation". IEEE Sixth International Conference on Software Testing,Verification and Validation, 2013,35: 242-251.
- [9] Tai, Kuo-Chung, Lei, Yu. "A test generation strategy forpairwise testing". IEEE Trans. on Software Engineering,2002 28(1): 109-111.
- [10] C. Nie and H. Leung, "A survey of combinatorial testing,"ACM Computing Surveys (CSUR)", vol. 43, p. 11, 2011
- [11] M. Grindal, J. Offutt, S. F. Andler, "Combination Testing Strategies: A Survey. Softw. Test. Verif.", 2005,15(3),167-199.
- [12] D.R. Kuhn, D.R. Wallace, Jr. A.M. Gallo, "Software fault interactions and implications for software testing". IEEE Transactions on Software Engineering, 2004, 30 (6): 418-421.
- [13] Bansal P., Sabharwal S., Malik S., Arora V.,Kumar V. "An Approach to Test Set Generation for Pair-Wise Testing Using Genetic Algorithms, in Search Based Software Engineering", Ruhe G.,Zhang Y., Ruhe G.,Zhang Y. Editors. 2013, Springer, Berlin Heidelberg, p. 294-299.
- [14] F. Arito, F. Chicano, and E. Alba, "On the application of sat solvers to the test suite minimization problem," in *SSBSE*, ser. Lecture Notes in Computer Science, G. Fraser and J. T. de Souza, Eds., vol. 7515. Springer, 2012, pp. 45–59.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

EFSM Slicer: Generation of Amorphous Slices based on Method Dependencies using EFSM

Aakash Gautam¹ and Neha Bajpai²

School of Information and Technology

Centre for Development of Advanced Computing

NOIDA, U.P, INDIA

Abstract: Program slicing is one of the oldest forms of testing where we divide the whole program into number of executables slices with respect to slicing criteria. However there is little work done on amorphous slicing and further less on amorphous slicing of extended finite state machine (EFSM). Amorphous slicing is transforming the whole program by applying any transformation algorithm. This paper introduces EFSM slicing algorithm using the method based dependence slicing and an accompanying approach. It has been used to conduct experiments on different systems including industrial EFSM and have has shown better results by generating smaller executables slices and automated generated EFSM depicting the dynamic behaviour of system. However the results have shown larger EFSM by 16.6% to 30% but this is more reliable than previous approach as methods works as the states for the EFSM which is identified automatically. This algorithm is successful is generating the non-terminating and non-deterministic slicing, one of the rare approach to do so as it is not possible by traditional syntax-based slicing. The paper also presents an inductive proof and illustrates the application of the algorithm with a detailed example for EFSM slicing.

Keywords: Amorphous slicing, FSM, EFSM, Non-terminating, Non-deterministic.

1. Introduction

Software testing ensures the reliability of the particular software, the system is being test with intend of finding out the errors. One the oldest technique used in the field of software testing is program slicing which in simple terms can be said as dividing the whole of the program into number of parts. It is ensured that each slice is executable which is produced according to some slicing criteria. Program slicing is widely used since its introduction three decades ago but there is little work done on amorphous slicing and further less in amorphous slicing of extended finite state machine. In the conventional finite state machine, the transition is associated with the set of input Boolean condition and set of output Boolean functions where as in extended finite state machine the transition is associated with “if condition” and have stores. If the particular condition is true the transition is triggered moving the machine to the next state. EFSM have memory to store where it stores timer variable and other variable condition, it working is not just restricted to the Boolean functions of true and false. Thus the EFSM can be defined as the FSM with capability of store and holding conditions and not merely the Boolean condition.

Amorphous slicing is completely transforming the whole program into the new one and not just deleting statements like in traditional syntax based program slicing. Any transformation may be applied in amorphous slicing. It does not preserve the syntactic structure of the original program but only the semantic. By transforming the program, smaller slices are made in comparison to the old syntax preserving slices which are desirable to find out source of error in debugging more efficiently. Amorphous slicing is an automated source code extraction technique with applications in many areas of software engineering, including comprehension, reuse, testing and reverse engineering. By extended finite state machine the dynamic behaviour of the system can be easily depicted.

The only previous algorithm of amorphous slicing using extended finite state machine where the tool have been used to generated the required EFSM and then the sliced EFSM is produced. But the states for the same have to be identified manually, so it could be quite difficult especially for the large programs. They have worked on three different control dependencies the Non- Termination Sensitive Control Dependence (NTSCD) (i.e. slicing retains non-terminating subprograms), the Non-Termination Insensitive Control Dependence (NTICD) (i.e. slicing removes non-terminating subprograms) and the Unfair Non-Termination insensitive control dependence (UNTICD) (CD in control sinks is not identified). This paper introduces an approach where the states detection, EFSM generation and slicing are all automated. This paper introduces EFSM slicing algorithm using the method based dependence slicing and an accompanying tool named as the EFSM slicer. EFSM slicer has been used to conduct experiments on 24 systems which generates non-terminating the non-deterministic which includes

industrial EFSM and have shown better results by generating smaller executables slices. This paper also contains the algorithm used and thoroughly explained with help of taking an example. This will show that the algorithm is suitable for producing non-terminating and non-deterministic slices. This algorithm is also compared with the only previous approach amorphous slicing of extended finite state machine

The rest of the paper is organized as follows: Section II describes the related work in the area proposed by the author, III describes the approach, IV contains result and V Describes Conclusion & Future Work.

II. Literature review

Various algorithms have been introduced in the field of amorphous slicing since its introduction including Search Based Algorithms, Loop squashing algorithm, Amorphous slicing of extended finite state machine, GUSTT etc. This section briefly describes and compares the techniques used by different researchers with their conclusion. Some of them are described in this section.

1. Mark Harman, Lin Hu, Malcolm Munro and Xingyuan Zhang have given an algorithm named GUSTT which mixes the reduction transformation with other transformation and produce an algorithm for amorphous slicing. It is one of the early works done in the field of amorphous slicing which provided base for the other algorithm to groom. The algorithm uses the simple approach of domain reduction which uses only the parent domain to test all other which is inherited or can be checked indirectly by checking with only the parent domain. Again it is not a general purpose algorithm and cannot be automated. Also, the sizes of the slices were not reduced greatly but however the algorithm was able to reduce the size of the slices in comparison of the syntax based slicing.
2. Deji Fatiregun, Mark Harman and Robert M. Hierons have worked on different search based algorithms which includes Genetic algorithm, a Hill climb algorithm, Systematic search algorithm and Random search algorithm. It has shown that how amorphous slicing can be generated using different search techniques. The algorithm presents result from a set of program and each of them is applied and explored the applications. Each algorithm is applied accordingly to explore the space possible for transformation sequences. It has been shown that algorithms had given surprisingly good results as it was automated search with no bias. They automatically locate the search space for the solutions which best fit the human assumptions in the fitness function. This is one of the central strengths of the approach. It has provided base for the loop squashing algorithm to evolve. The results have show that this algorithm had given better result from dedicated analytic amorphous slicing system.
3. Lin Hu, Mark Harman, Robert M. Hierons and David Binkley have introduced loop squashing algorithm which transform the loop with the conditional assignment thus reducing the size of the slice. The algorithm can detect when the particular loop must execute at least once. Basically the algorithm is divided into two parts first they syntactically identify the loop induction variable and then it is transformed accordingly. The algorithm was successful when applied to nested loops, it process the inner most first and it work its way out until encountered by a loop that cannot be squashed. The five basic step of the algorithm includes identification of induction variable, normalization, pattern matching of the loop, iteration computation and finally computation. However the algorithm was successful in transforming the program and in producing smaller slices but only which involves reduce able loops. The algorithm is not general purpose transformation algorithm.
4. Kelly Androutsopoulos, David Clark, Mark Harman, Robert M. Hieons, Zhengli and Laurence Tratt have introduced Amorphous slicing of extended finite state machine algorithm. The algorithm uses the extended finite state machine to transform the whole of the program which is more general purpose transformation algorithm in comparison of the above two. EFSMs can hold conditions and have a store which gives its edge over finite state machine which can only hold Boolean condition true or false. The dynamic behavior of the system is depicted through the required EFSM which easily checks the dependencies and test accordingly. Input is given as a text file and then EFSM is generated and sliced according to the user input of the destination node. It is the first approach to produce non-terminating and non-deterministic slices, first algorithm to do so. EFSM generated is based on transitions, if condition is true the machine moves to next state and so on. They have worked on three types of control dependencies- Non- Termination Sensitive Control Dependence (NTSCD) Non-Termination Insensitive Control Dependence (NTICD) and Unfair Non-Termination insensitive control dependence (UNTICD). NTSCD slicing retains the non-terminating subprograms that's why called as the sensitive slicing, the second one is NTICD for which slicing removes non-terminating subprograms which makes its insensitive as the non-terminating part is fully removed and hence smaller slice is produced and the

third one is UNTICD for which the control sinks is not identified in this type of control dependency. The algorithm is highly successful in insensitive slicing as the slice is highly reduced. The algorithm have shown great results in decreasing the size of the slices produced through extended finite state machine as their average slice is smaller 40 percent of the time and larger only 10 percent of the time, with an average slice size of 35 percent for termination for insensitive slicing.

All the three approaches have readily contributed to the evolution of the amorphous slicing. The transformation of the whole program is difficult and finding out the general purpose transformation algorithm is even more difficult. The search based algorithm has although given surprising results in some cases but it was not one of the reliable form and transformation for the amorphous slicing. Same with the loop squashing algorithm where it has given good results where the loop is present but its efficiency reduce readily as soon as program with few or no loops is tested through the approach.

However the amorphous slicing of extended finite state machine is successful in producing the non-terminating and non-deterministic slices which is the first algorithm to do so but not fully automated. By working on the limitation of the previous algorithm we have developed our approach accordingly which is fully automated, described in the next section. The comparative analysis of all the above discussed approach is given in the Table-1 for much clearer view to differentiate.

A. Comparative Analysis

Title	Work Detail	Analysis
Paper 1 GUSTT: An Amorphous Slicing System which Combines Slicing and Transformation. [5]	Uses the dependence reduction transformation with the other pre-processing and the traditional syntax based slicing to achieve the amorphous slicing.	It is the one of the early work in the field of amorphous slicing. Algorithm adds up reduction transformation with the syntax slicing, where it uses the single domain to test all other domains which come under it. Manually done, provided base for upcoming techniques.
Paper 2 Loop Squashing Transformations for Amorphous Slicing. [8]	Here they have reduced the loop in the program to the one conditional statement by applying their loop squashing algorithm. The algorithm is divided into two part first part of induction variable identification and second is loop transformation.	Although the large program can be transform by this approach but it is not effective in the program where the loop is not present. It cannot be said as the general purpose transformation algorithm.
Paper 3 Search-Based Amorphous Slicing. [9]	In this approach the various searched based algorithm like genetic algorithm, a hill climb algorithm, systematic search algorithm, random search algorithm and deduced results.	The surprising result shown as the search for transformation is made auto and not human based. The results with random search algorithm were comparatively better than the other search algorithm.
Paper 4 Amorphous slicing of Extended Finite State Machine [11]	A set of dependence-based EFSM slicing algorithm and a accompanying tool for amorphous slicing. The system is transformed to extended finite state machine and sliced EFSM is produced according the user input of destination node.	The algorithm used here is capable of slicing non-deterministic and non-terminating EFSM which was not possible by earlier used algorithm. The slices produced were much smaller in comparison with the traditional syntax based slicing.

Table-1 Comparative Analysis

III. Proposed approach

This approach has been work out by keeping the problems of the previous approaches in mind. Many problems have been identified while research and development of this algorithm. Every possible dimension has been kept in mind to develop an approach which is automated and sidelining the problems of the previous approaches. Some of the problems of the previous approaches have been given below which are worked out in this approach.

- 1) The problem in finding out general purpose transformation algorithm. Loop squashing, GUSST were all program specific problem transformation.[5]
- 2) Carrying out auto search space for transformation in case of amorphous slicing by search based algorithm often gave surprising results but it is not one of the dependable technique, size of slices were not reduced greatly.[9]
- 3) The states and the transition associated were identified manually, which involves a lot of manual work in amorphous slicing of extended finite state machine.[11]
- 4) To produce the EFSM by using the existing tool one has to manually identify dependencies, which is quite a complex work to carry out.
- 5) The tool involve is not user friendly as the lot of manual work has to be done.
- 6) Much of the work was successful in insensitive slicing that produced smaller slices but it have arisen the problem of the identification of source of error.

Working on the above specified problems, this approach has been developed. Efforts have been made to overcome the limitations of the previous approaches and deliver a more reliable approach. Each step of the proposed approach has been explained briefly here. The proposed approach contains the following steps:-

- 1) It will take the executable Java program as an input and then only carry out further steps of the approach.
- 2) The different methods will be identified in this step. Each identified method will act as a state for the system's EFSM.
- 3) The dependencies among the different identified methods of input program which will act as a state for EFSM. Each state will be holding some conditions of variables known as transition; if true the next state will be triggered.
- 4) On the basis of extracted methods and identified dependencies the EFSM of the whole system will be generated depicting the dynamic view of the system. The whole of the program flow and specified transition will be shown through the generated EFSM.
- 5) After generating the system EFSM, amorphous slices of the input program are generated. Input the destination node in order to produce the required sliced EFSM. It will backtrack from the destination node and mark the traverse nodes. After it will delete the unmark nodes and copies states which results in generating the required sliced EFSM.

The above proposed algorithm allows the system to be fully automated. The user has to only input the required program and the EFSM is produced. This will further produce the slice EFSM according to the user input. The dynamic view of the system is depicted through the EFSM produced the one has not to go through the entire program in order to understand it dynamically and understand its dependencies. It has many applications in the field of the software engineering including debugging, finding out source of error, re-engineering, code optimization, maintenance etc. It is useful for reverse engineering as smaller slices are highly recommended for it. Moreover it is only concentrated on the semantic structure of the program and not the syntactic one. Following is the graphical representation of the approach for better understanding.

This approach is used to test the java program dynamic behaviour in short time, check out the dependencies among methods and produce the following EFSM. Amorphous slicing is done by first transforming the program in form of extended finite state machine and then slicing it according to the user input of destination node. Dependencies are the major concern of the proposed approach so that every method dependency is shown to depict the behaviour of program and slicing it accordingly.

To illustrate the functioning of above approach we take an example of an elevator. Elevator is a non-terminating system and contains the non-deterministic node. Non-terminating is a system which goes on until infinity and does not contain the exit node. Non-deterministic means the sibling transition with same trigger event, the system will behave differently for the same input.

Step1) The first step of this approach is to take a java program as input, so we took the source code of a elevator as to produce non-terminating and non-deterministic slices.

Step2) The EFSM slicer tool will identify the all methods defined in the elevator program. The identified method will now act as the states for EFSM and condition will be hold by the transition.

Step3) The EFSM slicer will now identify the given dependencies of the states. States will be holding some conditions or timer stores depend on which the machine will move to the next state if the condition hold is true. For each triggered method there is a transition that that will takes place and will draw the dependencies for the EFSM. All these will be included in the system EFSM.

Step4) Now according to the dependencies drawn from the above step the system EFSM is generated. The system EFSM will contain the dynamic view of the system and will contain all the states from which the system has to travel. The transitions will also be shown on the basis of which next transition will be triggered. The snapshot for the same elevator program is given below.

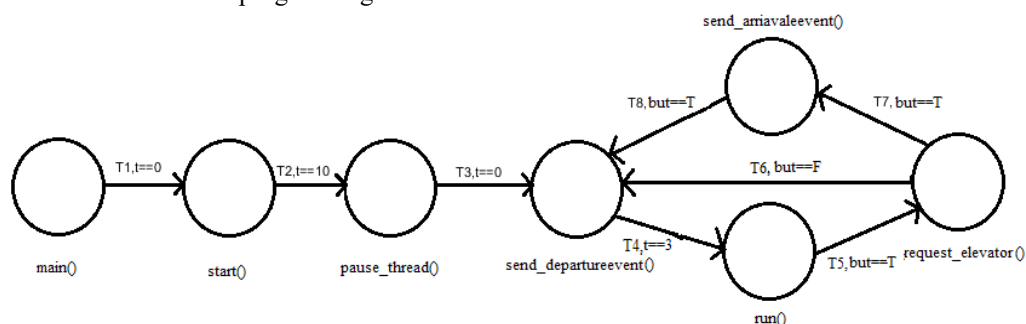


Fig 1: System EFSM

Step5) After generating the required EFSM for the elevator now the slicing part comes. The sliced EFSM will be generated according to the destination node and its variable. Here in this example the EFSM is sliced for the destination node request_elevator. An image for the sliced EFSM with slicing criterion of destination node request_elevator and its variable transition T5 is given below for reference.

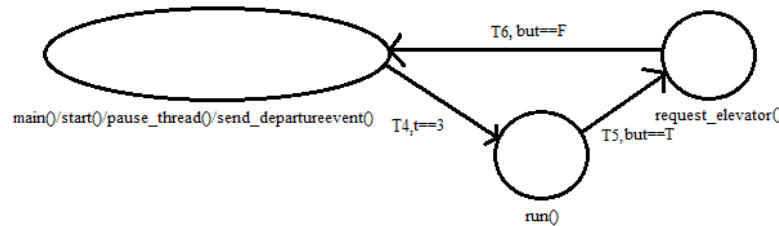


Fig. 2: Sliced EFSM

Here the example of an elevator program has been taken to show the working of this approach. The EFSM slicer tool takes input as a java program. It first checks its working it will only be run if the program is in run able state or not. Then it will pop up dialogue box showing the hierarchy of the methods in the program. At this point in the given elevator it has identified seven methods as states and built its EFSM according to its dependencies as shown in the snapshot Fig. 1.

To slice the given EFSM the destination node is entered, in this case the destination node entered is request_elevator. According to the algorithm it has back track the EFSM till it reaches the starting node that is main(). All the other nodes are deleted which has not been traverse during the back tracking. Also the equivalent nodes main, start, and pause thread are made to merge with send_departure making them as a single node. This has made the slice further smaller and we get a non-terminating slice which is not be possible with the traditional syntax based slicing.

IV. Results

The results of the empirical study on elevator program indicate that this algorithm can significantly reduce the size of EFSM slices in comparison to the traditional syntax-based slicing. EFSM slicer has successfully tested 22 systems and produced slices which are non-terminating and non-deterministic. However the EFSM produced was larger from about 16.6% to 30% but it has made the testing automated and much reliable. The EFSM slicer tool is one the rare tool to produce the non terminating slices of the program. Manual work tends to small mistakes which often lead to disaster in the field to software testing that all have seen in the past. The tool allows the states recognition, dependencies, EFSM generation and slicing to be fully automated. The work has shown it is more reliable and produces smaller slices which is the key thing for the reverse engineering.

V. Conclusion and Future Work

This paper introduced a slicing algorithm based on method dependence EFSM and an approach for EFSM slicing. The algorithm is capable of slicing non-deterministic, non-terminating EFSMs. The paper has shown the suitability of properties of the algorithm for amorphous slicing and presented a detailed empirical study on elevator EFSM model, using standard benchmarks. The study compared the algorithm to only previous algorithm of EFSM slicing. The approach is used to test the java program in short time, check out the dependencies between the methods. The EFSM generated finely shows the dynamic view of the system. The smaller slices produced has many applications in field of software engineering including re-engineering, reverse engineering, code optimization, locating source of error, maintenance and etc.

For the future purpose we can have algorithm that can work on more than one class of the system. We may also consider some much complex dependencies arise with the sub classes and inherited methods. With some modifications we may be able to test the systems programmed in different programming languages like C, C++, ASP.Net etc. For the future concern one may again modify this algorithm to work on multiple class and concept of inheritance.

VI. References

- [1] H. Agrawal, "On Slicing Programs with Jump Statements," Proc. ACM SIGPLAN Conf. Programming Language Design and Implementation, vol. 29, no. 6, pp. 302-312, June 1994.
- [2] D. Binkley and K.B. Gallagher, "Program Slicing," Advances in Computing, M. Zelkowitz ed., vol. 43, pp. 1-50, Academic Press, 1996.
- [3] C. Bourhfir, R. Dssouli, E. Aboulhamid, and N. Rico, "Automatic Executable Test Case Generation for Extended Finite State Machine Protocols," Proc. Int'l Conf. Testing of Communicating Systems, pp. 75-90, 1997.
- [4] C. Calude, E. Calude, and B. Khossainov, "Finite Nondeterministic Automata: Simulation and Minimality," Theoretical Computer Science, vol. 242, nos. 1/2, pp. 219-235, 2000.
- [5] M. Harman, L. Hu, X. Zhang, and M. Munro, "GUSTT: An amorphous slicing system which combines slicing and transformation" 1st Workshop on Analysis, Slicing, and Transformation, IEEE, 2001.

- [6] D. Binkley and M. Harman, "An Empirical Study of Predicate Dependence Levels and Trends," Proc. 25th IEEE Int'l Conf. Software Eng., pp. 330-339, May 2003.
- [7] M.H. Albert and S. Linton, "A Practical Algorithm for Reducing Non-Deterministic Finite State Automata," Technical Report OUCS-2004-11, Univ. of Otago, 2004.
- [8] Lin Hu, Mark Harman, Robert M. Hierons and David Binkley, "Loop Squashing Transformation for amorphous slicing" 11th Working Conference on Reverse Engineering, IEEE 2004
- [9] Denim Fatiregum, Mark Harman and Robert M.Hierons, "Search-Based Amorphous Slicing", 12th Working Conference on Reverse Engineering, IEEE 2005
- [10] K. Androutsopoulos, D. Clark, M. Harman, Z. Li, and L. Tratt, "Control Dependence for Extended Finite State Machines," Proc. Fundamental Approaches to Software Eng., vol. 5503, pp. 216-230, Mar. 2009.
- [11] Kelly Androutsopoulos, David Clark, Mark Harman, Member, IEEE, Robert M.Hierons, Senior Member, IEEE and Laurence Tratt, "Amorphous Slicing of Extended Finite State Machines", 13th International Working Conference on Reverse Engineering, IEEE 2013



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

XSS-VDetector: XSS VULNERABILITY DETECTION IN WEB APPLICATIONS

Tannu Rohela¹ and Neha Bajpai²
School of Information and Technology
Centre for Development of Advance Computing
Noida, UP, INDIA

Abstract: The ease of availability of web applications has led to the development of different types of threats and these threats causes security issues. Web applications are used as a channel for communication, for obtaining information, for transaction and other similar related activities. Due to these threats web applications fail to perform their function. One of the major threats these days is cross-site scripting (XSS) attack which is ranked number 2 among the vulnerabilities by OWASP. XSS allows attackers to inject vulnerable content into the web pages leading to security risks. In this paper, an approach has been developed to identify vulnerable points in a web page. This approach uses an encoder function which identifies vulnerabilities and classifies them into three different categories: Tag vulnerable, Attribute vulnerable and Content vulnerable. This classification helps the developer to identify vulnerable points in the code and changes can be made accordingly.

Keywords: web security, Cross-site scripting, OWASP, vulnerability.

I. Introduction

A web application provides a subtle channel of communication between web service providers and end-users. Therefore safeguarding the security of these applications is compulsory. The most frequent security delicacy is the negligence of properly validating the input from the user. This ignorance causes the most common vulnerability in web application, injection of malevolent content known as Cross-site scripting (XSS). Websites and services that have large number of input fields are potentially vulnerable. JavaScript plays a pivotal role in XSS. It can redesign the page making the page to appear different and act in a contrary fashion. The most important part of XSS comes when `<script>` tag is used. The `<script>` tag is invisible to the user and tells that anything between the `<script>` tag is JavaScript. There are many different ways by which XSS can be instigated. One way is to embed the malevolent code in the link which gets implemented when consumer visits the page. The URL is encoded in such a way that it seems authentic to the inexperienced user. Stealing user's credentials and posting ads pop-up are some of the ways by which the attacker scrambles malevolent code to look bona fide to the user. XSS impacts can be classified as: Technical impact and business impact. Technical impact is when the malevolent link gets executed on the users browser for stealing the credentials. Business impact is when the system and the data gets affected due to the XSS exploit. XSS is of three different types: DOM based, Stored XSS and Reflected XSS. For checking the vulnerability towards Reflected XSS the script used is: `<script>alert("XSS")</script>`.

To tackle this delicacy, this paper presents an approach which automatically detects XSS vulnerabilities. This approach is built on the analysis of vulnerable tags, attributes and content.

The pattern of the paper is as follows. Section 2 narrates the work done in this area. Section 3 delineates the proposed approach. Section 4 sketches the result. Section 5 outlines the conclusion and the future work.

II. Literature Review

A lot of work has been done in this field. This part of the paper describes the various techniques used by some of the researchers.

Mohit Dayal et al have outlined different impressions of XSS, kinds of XSS, scanned the site for checking if the site is vulnerable to XSS or not, scrutinized various XSS tools and summarizes the preventive measures against XSS. They have sketched different XSS examples such as attacks via email, stealing user's cookies, sending an unauthorized request and XSS attacks in comment fields. They have talked about the technical and business impacts of XSS. They have briefly explained different types of XSS. For their implementation they have used 'XAMPP' web server. They have displayed the results of each modification done using the `<Script>` tag. They have also mentioned rules for preventing XSS.[1]

Kanpata Sudhakara Rao et al, proposed an approach in which they have split each HTML request parameter into HTML and JavaScript context and stock them distinctly. This approach works against all XSS attack vectors including attribute injection, partial script injection and HTML injection. This approach also works clickjacking

attacks. They have used threshold value and concept of regular expression for detecting and encoding malevolent context.[2]

Thiago S. Rocha and Eduardo Souto have proposed an approach which identifies and analysis all entry points of the application and creates a particular code insertion test for each of the entries. Their results have proved that the correct filling of the input fields with only valid information assures a better efficiency of the tests, increasing the rate of detection of XSS attacks.[3]

Bill chu et al presents an approach that extracts the encoding functions which is used in a web application for sanitizing untrusted input and then evaluates the effectiveness by generating the XSS attack strings automatically. Their evaluation shows this technique can detect 0-day XSS vulnerabilities.[4]

Jose Fonseca et al have proposed an approach for evaluating and comparing web application vulnerability scanners. Approach is based on the injection of software faults in web application so that efficiency of various tools in detecting the possible vulnerabilities can be compared. The result shows that various scanners display different results and all of them leaving a considerable amount of vulnerabilities undetected. The percentage of false positive is very high, varying from 20% to 77%. Results show that this approach allows easy comparison of coverage and false positive. As this approach shows the limitations of the scanners used, it can be used for the improvement of the scanners. [5]

Danny Alvarez et al have focused on three main hacking techniques in which XSS is also included. Their goal is to scrutinize how Colombian companies and organizations give relevance to security. They have first sketched about the attack techniques working and its preventive techniques. Then they have talked about their searching process of the vulnerable sites and their way of testing these vulnerabilities sites. They have found that almost eighty percent of the web sites tested outlined at least one basic vulnerability. They have also found that education sector is being affected the most. [6]

Comparative Analysis

Title	Work Details	Analysis
A Comprehensive Inspection Of Cross Site Scripting Attack.	They have explained XSS, its types and scanned the sites for vulnerabilities.	From their work, it can be said that JavaScript plays an important role in embedding the malevolent code into the page.
Two for the price of one: A combined browser defense against XSS and clickjacking.	In their work they have split each HTML parameter into HTML and JavaScript context.	They have used threshold concept which limits the usage of this approach.
ETSSDetector: a tool to automatically detect Cross-Site Scripting.	In their approach, they have identified all the data entry points and have generated code insertion tests for each one.	From their result it can be stated that if input fields are filled correctly there are less chances of XSS.
Automatic Web Security Unit Testing: XSS Vulnerability Detection.	In their approach, they have extracted encoding functions used for sanitizing untrusted input.	In their approach, they have used unit test extraction. They have divided each block into further blocks such that each block contains only one attack vector.
Testing and comparing web vulnerability scanning tools for SQL injection and XSS.	They have compared different vulnerability scanners. They have done this comparing using software fault injection techniques. The results are compared by analyzing coverage of vulnerability detection and false positives.	Their implementation result shows that various scanners display different results. This approach permits simple comparison of coverage and false positives. The percentage of false positive varies from 20% to 77%. This approach can be used for the implementation of various scanners as they display their limitations.
An Analysis of XSS, CSRF and SQL Injection In Colombian Software And Web Site Development.	They have focused on three main hijacking attacks. They have scrutinized how Colombian Companies give relevance to security.	According to their result, from the total web sites tested, 80% of the web sites have at least one basic vulnerability.

Table 1: Comparative Analysis

III. Proposed Approach

The work done so far has not focused on the vulnerable entry points such as tags, attributes and content. These entry points can easily allow the attacker to perform XSS. So to overcome this we have designed an approach. Our main goal is to design an approach that detects XSS vulnerabilities automatically. This section describes the approach proposed for the vulnerabilities detection. Firstly the source code of the web page to be tested is extracted. After the extraction of the source code, encoder function is executed. Encoder function generates an output which gives the vulnerable tags. With this output vulnerable blocks are classified. These vulnerable blocks are then classified into attribute vulnerable block and content vulnerable block. The main advantage of this approach is that it deals with all the vulnerable tags. This helps to identify all the vulnerable entry points by reading each line and comparing it with dataset. With these vulnerable tags, vulnerable attributes and vulnerable content are also identified helping to identify loopholes hidden deep inside.

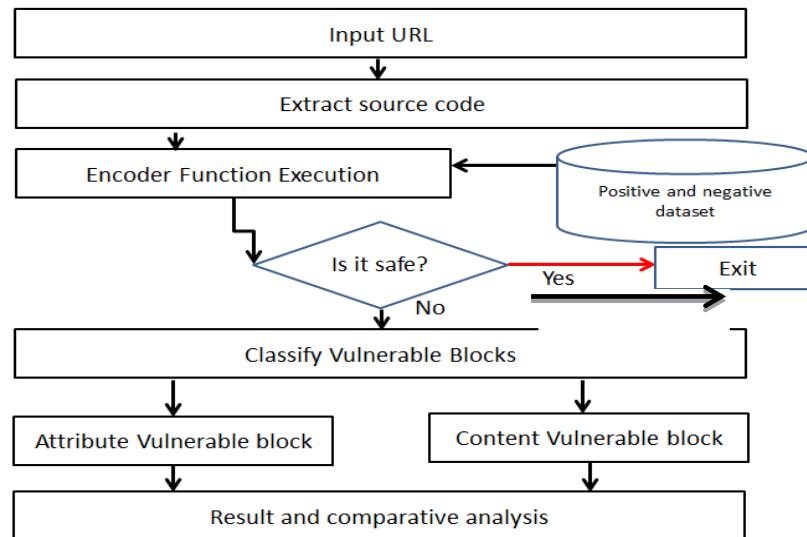


Fig 1 Proposed Approach

- A. **Input URL:** The approach starts by selecting a web page for testing. As the web page is selected the URL of the same page is selected for further processing. URL is needed for the source code extraction.
- B. **Extract Source Code:** The source code of the web page to be tested is extracted automatically. The source code is extracted so that all the tags in the source code can be identified. This step is important for the identification of vulnerabilities in the source code.
- C. **Positive and Negative dataset:** This dataset is the collection of almost all the vulnerable tags and non-vulnerable tags. This dataset has two datasets, one for non-vulnerable tags and another for vulnerable tags. Dataset of vulnerable tags is negative dataset and dataset for non-vulnerable tags is positive dataset. This data set is used for the identification of safe list and unsafe list. The negative dataset (partial) is shown in fig 2 and positive dataset (partial) is shown in fig3.

area	form	article	P
base	iframe	aside	meta
body	img	audio	noscript
blockquote	label	b	P
datalist	legend	basefont	style
embed	map	bdi	table
fieldset	menu	big	tbody
figcaption	menuitem	br	tfoot
figure	noframe	button	title
form	textarea	caption	track
iframe	output	center	ul
img	link	cite	var
label	iframe	code	video
legend	eval	col	P
		data	svg
		dd	polygon
		del	g
		details	path
		div	symbol
		i	style
		em	circle
		font	polyline
		h1	li
		h2	sup
		h3	span
		h4	b
		h5	tr
		h6	td
		html	ul
		head	small
		header	th
		P	table
		meta	tbody
		noscript	br
		P	b

Fig: 2 Negative dataset (Partial) Fig: 3 Positive dataset (partial)

- D. **Encoder Function Execution:** Encoder function is a function which takes source code and dataset as an input and generates an output on the basis of the input given. The output is in the form of 'safe' or 'unsafe' lists. The safe part occurs when a vulnerable tag is identified by the encoder function in the source code and unsafe part occurs when a non-vulnerable tag is identified in the source code by the encoder function. The encoder function reads the source code line by line and checks for the tags. The tags identified are then matched with the positive and negative dataset simultaneously. This matching results in the generation of vulnerable and non-vulnerable tags. Collection of non-vulnerable tags is called the safe list and the collection

of vulnerable tags is called the unsafe list. The further processing is done on the unsafe list. If the tag is identified as safe then it exits. If the tag is unsafe then it is moved to the next step.

- E.** Classify vulnerable blocks: A vulnerable block is a block with some loophole. Classification of the vulnerable block is done on the basis of attribute and content. Attribute vulnerable block is a block which contains vulnerable attributes and content vulnerable block is the block that contains vulnerable content.
- F.** Attribute vulnerable block: Attribute vulnerable block is a block which contains vulnerable attributes. Here in this research we are considering four vulnerable attributes: href, url, onclick and src. These attributes have been found vulnerable because they can redirect the page, can change the look and feel of the page, can change how the page behaves and such similar reasons. These attributes can be increased with further research. This section checks on the identified vulnerable tags if they contain any vulnerable attribute by matching the attributes with the above mentioned attributes. If so then again a 'safe' and 'unsafe' list is generated. This tells the number of yes and no. On the basis of yes and no a bar graph is generated as a result.
- G.** Content vulnerable block: Content vulnerable block is a block which contains vulnerable content. The content is the part that starts just after the closing of the tag, that is, the tag should not contain any attribute and just after the tag name closing of the angular bracket should come. In this research we have considered only few such contents. Vulnerable content is taken from the OWASP cheat sheet which proves the content considered in this research is vulnerable. This vulnerable content is used as a dataset. This dataset is used for matching the vulnerable content in the web page.

IV. Implementation and Result

Implementation

After the execution of encoder function the first result is to display safe and unsafe list. The safe list is the one with 'no' and remaining is the unsafe list. This is shown in fig 2. The further work is done on the unsafe list. The bar graph of this result is shown in fig 3.

XSS

Entered URL:

ID	TAG	Vulnerable
1	html	No
2	head	No
3	meta	No
4	title	No
5	link	yes
6	link	yes
7	meta	No
8	meta	No
9	meta	No
10	meta	No
11	meta	No
12	meta	No
13	link	yes
14	link	yes

Fig 2: Result for number of tags for a particular URL

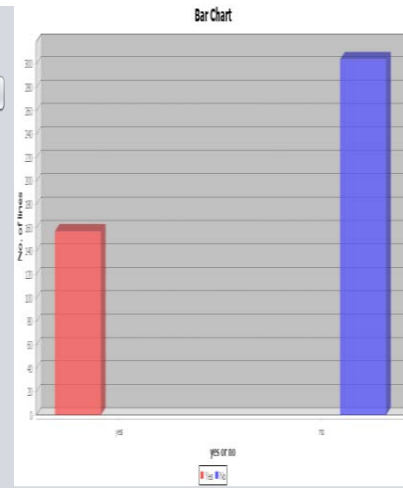


Fig 3: Bar graph for the safe and unsafe list

From these vulnerable tags, vulnerable attributes and vulnerable contents are identified. Result of Vulnerable attributes is shown in Fig 4 and result of vulnerable content is shown in Fig 5.

ID	TAG	Content	Attribute	Vulnerable
0	link	<link rel="profile" href="http://...	href	Yes
1	link	<link rel="pingback" href="htt...	href	Yes
2	link	<link rel="shortcut icon" type=...	href	Yes
3	link	<link rel="apple-touch-icon-p...	href	Yes
4	link	<link rel="apple-touch-icon-p...	href	Yes
5	link	<link rel="apple-touch-icon-p...	href	Yes
6	link	<link rel="icon" sizes="196x1...	href	Yes
7	script	<script>function(f,b,e,v,n,t,s)...	src	No
8	link	<link rel="canonical" href="ht...	href	Yes
9	script	<script type="application/javascript"...	url	No
10	link	<link rel="dns-prefetch" href=...	href	Yes
11	link	<link rel="dns-prefetch" href=...	href	Yes
12	link	<link rel="dns-prefetch" href=...	href	Yes
13	link	<link rel="dns-prefetch" href=...	href	Yes
14	script	<script type="text/javascript">	src	No

Fig: 4 Result of vulnerable attributes

ID	TAG	Content	Vulnerable
21	body	<svg xmlns="http://www.w3.org/2000/...	No
22	nav	<div class="container pos-r"> <div cl...	No
23	a	<svg height="33px" width="220"> <u...	No
24	a	Skip to content	No
25	a	Web Vulnerability Scanner <span cla...	No
26	a	Vulnerability Scanner	No
27	a	Indepth Crawl & Analysis	No
28	a	Highest Detection Rate	No
29	script	function(f,b,e,v,n,t,s){if(f.fbq)return;n=...	yes
30	a	Reporting and Remediation	No
31	a	WordPress Checks	No
32	a	Network Security	No
33	a	Advanced Features	No
34	a	Download Software	No
35	a	Free Online Scanner	No
...

Fig: 5 Result of vulnerable content

Result

Proposed approach has identified that there are total 550 tags in the source code out of which 156 tags are vulnerable tags. From these 156 vulnerable tags 129 are the tags which contain vulnerable attributes. Only 1 content is found as vulnerable.

- Total no. of lines: 550
- Total vulnerable tags : 156 (Red + Green)
- Non vulnerable tags: 394 (Blue)
- Vulnerable attributes: 129 (Red)
- Vulnerable content:1 (green)

V. Conclusion

In this work, we have designed an approach that detects vulnerabilities automatically. From the results this approach seems to be effective as it identifies all the vulnerable points on the basis of tags, attributes and content.

As future work we can add more number of tags in the dataset and more attributes and content can be added for the classification purpose.

VI. References

1. MohitDayal, Nanhay Singh , Ram Shringar Raw, “ **A Comprehensive Inspection Of Cross Site Scripting Attack**”, International Conference on Computing, Communication and Automation (ICCCA2016), IEEE 2016
2. KanpataSudhakara Rao, Naman Jain, Nikhil Limaje, Abhilash Gupta, Mridul Jain, Bernard Menezes , “ **Two for the price of one: A combined browser defense against XSS and clickjacking**” , International Conference on Computing, Networking and Communications, Communications and Information Security IEEE 2016Thiago S. Rocha and Eduardo Souto, “**ETSSDetector: a tool to automatically detect Cross-Site Scripting vulnerabilities**” 13th International Symposium on Network Computing and Applications, IEEE 2014
3. Mahmoud Mohammadi, Bill Chu, Heather Richter Lipford, Emerson Murphy-Hill, “**Automatic Web Security Unit Testing: XSS Vulnerability Detection**” , 11th IEEE/ACM International Workshop in Automation of Software Test , IEEE 2016
4. José Fonseca, Marco Vieira, Henrique Madeira, “**Testing and comparing web vulnerability scanning tools for SQL injection and XSS attacks**” 13th IEEE International Symposium on Pacific Rim Dependable Computing, IEEE 2010
5. Danny Alvarez E, Daniel Correa B , Fernando Arango I, “**An Analysis of XSS, CSRF and SQL Injection In Colombian Software And Web Site Development**”, 8th Euro American Conference on Telematics and Information Systems(EATIS), IEEE 2016
6. Guowei Dong1, YanZhang2,Xin Wang1,Peng Wang2, Liangkun Liu2, “**Detecting Cross Site Scripting Vulnerabilities Introduced by HTML**”,11th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE 2014
7. Dr. G. Shanmugasundaram, S.Ravivarman, P. Thangavellu” **A study on removal techniques of Cross-SiteScripting from web applications**” International conference on computation of power, energy, information and communication, IEEE 2015
8. Piyush A. Sonewar, Nalini A. Mhetre, “**A Novel Approach for Detection of SQL Injection and Cross Site Scripting Attacks**”,International Conference on Pervasive Computing (ICPC), IEEE 2015
9. Chih-Hung Wang, Yi-ShauinZhou,”**A New Cross-site Scripting Detection Mechanism Integrated with HTML5 andCORS Properties by Using Browser Extensions**”, International Computer Symposium,IEEE 2016
10. https://www.w3schools.com/tags/ref_byfunc.asp last accessed on 22/5/2017
11. https://www.owasp.org/index.php/Top_10_2017-Top_10 last accessed on 22/5/2017
12. https://en.wikipedia.org/wiki/Cross-site_scripting last accessed on 22/5/2017
13. Cross-site Scripting (XSS)[https://www.owasp.org/index.php/Cross-site_Scripting_\(XSS\)](https://www.owasp.org/index.php/Cross-site_Scripting_(XSS))last accessed on 22/5/2017
14. https://www.owasp.org/index.php/XSS_Filter_Evasion_Cheat_Sheetlast accessed on 22/5/2017
15. Cross Site Scripting Attacks: Xss Exploits and Defenseby Seth Fogie, Jeremiah Grossman
16. IsatouHydara, Abu Bakar Md. Sultan, HazuraZulzalil, and NoviaAdmodisastro. Current state of research on cross-site scripting a systematic literature review. Information and Software Technology, 58(0):170 – 186, 2015

VII Acknowledgments

I would like to show my gratitude to Miss Neha Bajpai for guiding me and sharing the pearls of her wisdom with me. Without her help this research would have been incomplete. I would also like to thank my panel members for their impeccable guidance. Last but not the least I would like to thank CDAC Noida for giving me such a knowledgeable faculty which helped me in overcoming all the obstacles.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Web Application Security: An Integral part of Web Application Development Life Cycle

Pragya Sharma¹, Priyesh Ranjan², Praveen Kumar Srivastava³

Health Informatics Division

Centre for Development of Advanced Computing

Anusandhan Bhawan, C-56/1, Institutional Area, Sector-62, Noida, India

Abstract: With the advancement in technologies and emerging use of business and delivery services over the internet, the use of websites and web services has increased enormously to attract mass users. Along with the increment in users and profit, the exposure of sensitive data to suspicious users and cyber criminals has emerged as a growing aspect of concern. As a result, organizations need to pay increased attention to the security of web applications in addition to the security of underlying networks and operating systems. However, ensuring that all critical applications are assessed for security before and during production is a complex and challenging task. Security is often overlooked during design and development phases when the key objective is rapid time-to-market. Implementing security protocols in later stages of the Software Development Life Cycle (SDLC) results in an increase in delivery time frame and budget. In this paper, we outline an approach along with its advantages which will help designers and developers in ensuring security of web application at the early stages of software development in a cost effective and timely manner. Designing an application, keeping in view the aspect of addressing possible security vulnerabilities, that may come up with new technologies or when applications migrate to cloud has also been taken care of in this approach.

Keywords: web application; web security; vulnerabilities; web application development life cycle;

I. Introduction

Web security vulnerability is a weakness which allows an attacker to reduce the application's and system's information assurance. Vulnerability is the intersection of three elements:

- An Application's flaw,
- Attacker access to the flaw,
- Attacker capability to exploit the flaw

Open Web Application Security Project (OWASP) is a worldwide non-profit organization with an objective to reach various communities to standardize and benchmark security issues. OWASP conducts security surveys every year to identify the vulnerabilities that have caused attacks [1]. The list of top ten web application vulnerabilities defined by OWASP classified on the basis of risk factors is shown in below Table I [2].

Table I presents a summary of the 2013 Top 10 Application Security Risks.

Risk	Threat Agents	Attack Vectors	Security Weakness (Prevalence)	Security Weakness (Detectability)	Technical Impacts	Business Impacts
A1-Injection	App Specific	EASY	COMMON	AVERAGE	SEVERE	App Specific
A2-Authentication	App Specific	AVERAGE	WIDESPREAD	AVERAGE	SEVERE	App Specific
A3-XSS	App Specific	AVERAGE	VERY WIDESPREAD	EASY	MODERATE	App Specific
A4-Insecure DOR	App Specific	EASY	COMMON	EASY	MODERATE	App Specific
A5-Misconfig	App Specific	EASY	COMMON	EASY	MODERATE	App Specific
A6-Sens. Data	App Specific	DIFFICULT	UNCOMMON	AVERAGE	SEVERE	App Specific
A7-Function Acc.	App Specific	EASY	COMMON	AVERAGE	MODERATE	App Specific
A8-CSRF	App Specific	AVERAGE	COMMON	EASY	MODERATE	App Specific
A9-vuln. Components	App Specific	AVERAGE	WIDESPREAD	DIFFICULT	MODERATE	App Specific
A10-unval. Redirects	App Specific	AVERAGE	UNCOMMON	EASY	MODERATE	App Specific

The awareness on the importance of addressing security flaws in a web application is very important and should be a mandatory part of induction programs and trainings [3][4]. In scenarios where the team believed it could pile on the features and then clean up the security issues during security scan sometimes result in a situation where it will be too expensive or infeasible to fix the vulnerabilities as the application is in a very advanced stage of its life cycle. This may result in applications with major security vulnerabilities on deployment. The major challenge in such scenarios is not just implementing the appropriate vulnerabilities' fixes but also to decide the most appropriate stage of software development where it should be applied, so that the development cost would remain tractable, while future vulnerabilities could be addressed without many changes and the application can be delivered in the stipulated timeframe [5][6].

The process of making a secure web application starts with listing out the security goals of the applications and vulnerabilities required to be addressed in achieving these goals. The security of web applications and authenticity should be taken into account from the design phase and it must be integrated as early as possible in the software life cycle [9]. Security goals can be brainstormed and specified during the system specification phase and requirement elicitation stage of the project. This will help the team to ensure that the security at various stages of the SDLC can be defined in conformance with security rules, coding guidelines, security tools and test cases, deployment strategy etc. There are tested solutions, guidelines, tools and techniques available on the internet through open source communities such as (W3C) [2], (OWASP) [1] etc. which could be referred to while selecting the solutions or implementing a custom solution.

II. Methodology

To overcome Web Application Security issues, various approaches have been classified and categorized by different organizations, communities and researchers. The aim of all these approaches is to ensure the security of the application [12]. However, since multiple applications have different capabilities and consequently the impact at different stages of the SDLC to security vulnerabilities varies, it results in different costing and delivery effects. Every Web Applications goes through the following stages during its SDLC:

- Requirement Elicitation
- Design
- Development
- Testing
- Deployment
- Maintenance

The stage to implement solutions for Security Vulnerabilities in Web Applications depends on the certain factors:

- Awareness & knowledge level of the stakeholders involved in building applications,
- Profile of the applications being developed, whether it is a low risk or high risk application.
- Time constraints and budget.

Based on the experience of development of multiple web applications at Centre for Development of Advanced Computing (C-DAC), Noida and analysis on a number of Internal and External Security Audit Reports [7] of our applications, we have tried to broadly categorize the approaches we used as follows:

A. Address security issues during implementation phase in Web Applications

This approach was followed for the applications that were developed a decade ago. These applications were designed, developed and tested based on functional requirements only. Security was not taken into consideration during development and several factors contributed to this decision at that time:

- Less awareness of the team about the security vulnerabilities.
- Mindset of the developers or designers which was set to protect all such vulnerabilities during deployment/implementation phase.
- Timelines of the projects were short.
- Limited Budget.
- Applications needed to be hosted on Virtual Private Networks at the time of requirement elicitation.

To keep up with the challenges of the dynamic business world, there were requirement specifications to relocate the applications on cloud. As per the security policies, every cloud application needed to clear the requisite security certifications mostly the state-to-host certificate before hosting the application on a data centres. The terms and conditions for the certification generally depends on the individual data centres guidelines [13] [14]. As a result, Security Audit/Scan of web applications was done by authorized external agencies like AKS IT Services, Cyber Security and Cyber Forensics- C-DAC, Tata Communications etc. During this process, a list of vulnerabilities found in the applications were identified which needed to be resolved. Extensive analysis of the issues and their existing solutions were done to assess the severity and impact. Result of the analysis required changes in design, coding, server configuration etc. A lot of rework had to be done in all previous phases of the application and the security implications were re-assessed, multiple cycles of this scan and fixes were followed to resolve the security issues.

B. Focusing on Security Vulnerabilities in Web Applications from Early Stage of their life cycle

We followed this approach for applications that were developed recently or are in the pipeline for development. For these applications security issues were considered from the very initial stages of the application development. During requirement analysis and planning phases, security plan of the application was prepared. Design of the application was done by taking into account the non-functional security requirements of the application e.g. Configuration Management, Authentication, Authorization, User and Session Management, Data Validation, Error Handling, Data Protection and Logging. Coding guidelines were also set to adhere to the security requirements and development was done in accordance and compliance with the guidelines.

Testing of the applications was also done from each and every aspect of the identified security vulnerabilities. As a result, when the application was scanned for security vulnerabilities, very few issues were identified. So now the application could safely be hosted on web. However, dynamically changing business requirements led to constantly changing the code-base of the application. As a result, there were chances of introduction of new vulnerabilities in the application. So during maintenance phase, periodic assessment and testing of the application security was planned to keep the application up to date in terms of protection from security vulnerabilities.

Different security testing techniques (both manually and using tools) were employed to unearth application security vulnerabilities, weaknesses and concerns in the following aspects: Input Validation, Authentication and Session Management, Access Control, Error Handling, Data Protection, Denial of Service, File Extensions Handling and Web Application Finger Print etc. Few of the strong recommendations for Deployment Server and Software infrastructure safeguarding are:

- The production server should have operating system and web server hardening done using advanced and updated security measures in place. For e.g., Entire website should enforce protocols such as SSL.
- The servers should be physically protected from unauthorized access.
- Write permission to be given to only selected folders and sub folders whose transaction processes have passed the vulnerability scan test.

III. Inferences from Our Approach

There was a lack of awareness and knowledge about security vulnerabilities and their impact in our initial projects. We had not considered security goals in the earlier phases of the application development life cycle. When we got awareness of Web Security Vulnerabilities, we followed the approach (A) as discussed in the previous section (II). The problem with this approach was that applying and fixing security vulnerabilities after application was developed resulted in compromising the cost and expected delivery time frame of application life cycle. Extra manpower and cost had to be involved for the rework and the deadline for the project had to be reworked which caused delay. Some of the bugs found during security scan were still hard to fix completely because it could invalidate a few of the crucial application requirements and were expensive to fix at this stage. As a result, they are partially patched up, making the product vulnerable to attack in some scenarios. Due to these findings, we decided not to continue the approach (A) in successive projects.

We followed the approach (B) as mentioned in the previous section and received several advantages. We have suggested ensuring certain guidelines in each phase of the SDLC so that security aspects are identified and addressed during product development and not as an isolated process or approach. In the paragraphs below, we explain the stage-wise additions, drawback if not followed with a practical example we faced and the advantages obtained (if followed):

A. Requirement Elicitation

1) Additional Changes:

- Elicitation of Application Security Parameters such as
 - Nature and Size of Target Audience
 - Hosting/Production Environment of Application
 - Level of Exposure to the different categories of users and corresponding privileges/access rights
 - Interface points to Internal and External Components keeping in mind the security exposure/impact
- Listing of Security Vulnerabilities required to focus based on Security Parameters of Application
- Analysis the impact of each Vulnerability on application
- Listing of the available security testing tools
- Setup the Security Goals in project management plan

2) Drawback(s) if not followed

- If vulnerabilities and their impact are not listed at this stage, designers and developers may waste their efforts on the non-significant and low impact vulnerabilities for the application.
- Vulnerabilities with critical risk factors for the application environment may not be considered seriously.

3) Advantage(s) if followed

- Re-work on analysis of vulnerabilities finding were not left for designers and developers, so resulted in time and efforts savings

- Security scope of the application was clearly identified and defined.

B. Design:

1) Additional Changes:

- Designing secured Authentication and Authorization of Process
- Selection of Best suited Solutions against Vulnerabilities
- Segregated List of Sensitive Data and Normal Data
- Listing of Secure Coding Practices
- Listing of Configuration aspects for Deployment Environment

2) Drawback(s) if not followed

- May result in extensive wastage of efforts at development time
e.g. Setting 'Content Type' rule is not set in 'Coding Practices' then, it may result extensive rework after application has been developed.

3) Advantage(s) if followed

- Ensuring structured development environment taking into consideration security vulnerabilities.
- Coding rework is reduced

C. Development:

1) Additional Changes:

- Development of Security Solutions
- Coding as per Secure Coding Practices proposed in design
- Development Environment configurations should be same as Production
e.g. Purge confidential and sensitive information from exception code block is important, if your code is trying to read an underlying configuration files and that file is not located it would give java.io.FileNotFoundException containing the file path and display the call back path in error block which would eventually explains the layout of entire file system to the hackers.

2) Drawback(s) if not followed

- Rework efforts for security are increased.

3) Advantage(s) if followed

- Less Security Bugs

D. Testing:

1) Additional Changes:

- Testing web security with tools stated in requirement elicitation
- Testing Application on multiple Security Parameters
e.g. Penetration test or popularly known as pen test by targeting system which are showing random test reports for a particular vulnerability while scanning through tools (like invalidating previous session ids on application login and logout).

2) Drawback(s) if not followed

- May result in selection of wrong tool for security testing.
- Manual testing may result in vulnerable application.

3) Advantage(s) if followed

- Secured Tested Application in all identified aspects

E. Deployment

1) Additional Changes:

- Secured Configuration(set at design) of Deployment Environment

2) Drawback(s) if not followed

- If Configuration is not set at design time, new configurable settings at deployment time may result in functionality impact on Application.
e.g. While fixing vulnerabilities related to adding Http Only setting in cookies, before enabling the HttpOnly flag in deployment server configuration developer should ensure that the relative settings in deployment descriptor files during coding. If code and server configuration are not in sync that may result in deployment failures.

3) Advantage(s) if followed

- Secure environment setups in one go based on predefined guidelines.

F. Maintenance:

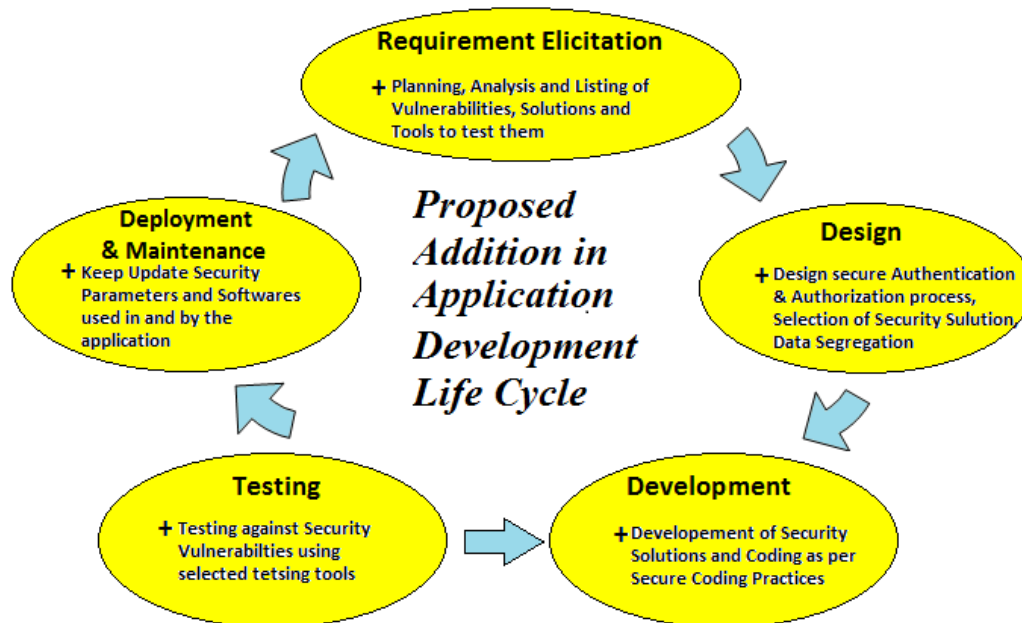
1) Additional Changes:

- Security ensuring in deployment environment time –to-time
- Regular update of Softwares/Components used
- Regular Assessment of the security measures

2) Drawback(s) if not followed

- In future, application may be exposed to new vulnerabilities
- 3) *Advantage(s) if followed*
- Ease in Updating of Security Features

Figure 1 shows Proposed Addition to Application Development Life Cycle



Security assurance is believed to be challenging when software is developed in successive iterations. There are currently three changes that impact the security of software: security requirements changes, code changes, and security mechanism changes. Changes that impact software security are rather not frequent but consume time to identify their impact on the security. This issue is challenging considering that agile development teams do not maintain up-to-date architecture diagrams for their software, which makes threat modeling using such diagrams inconsequential[8]. Therefore, conscious attention should be given while choosing the appropriate development model for the project in the project initiation phase [10][11].

IV. Conclusion

Security of the applications should not be compromised by providing their solutions at a later stage as it is difficult to implement, as well as involves more cost and time. For business critical web applications, security flaws, if left, may result in compromising data integrity, media attack and loss of client confidence. So security features should be weaved into each and every phase of a conventional SDLC. This will make the application secure from scratch as well as help in easily adoption of the solutions for newly introduced vulnerabilities, especially in evolving scenarios with advancements in technologies and increased techniques of cyber crimes. A regular assessment of the security measures of web applications should also be done in order to keep the application security updated.

V. References

- [1] Open Web Application Security Project (OWASP);Top Ten project 2013: <https://www.owasp.org/index.php/Category:>
- [2] World Wide Web Consortium (W3C): <http://www.w3.org/>.
- [3] Kaur, Daljit, and Parminder Kaur. "Empirical Analysis of Web Attacks." *Procedia Computer Science* 78 (2016): 298-306.
- [4] Kanniah, Sri Lakshmi, and Mohd Naz'ri Mahrin. "A Review on Factors Influencing Implementation of Secure Software Development Practices." *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering* 10.8 (2016): 2935-2942.
- [5] Gupta, Shivangi, and Saru Dhir. "Issues, Challenges and Estimation Process for Secure Web Application Development." *Computational Intelligence & Communication Technology (CICT), 2016 Second International Conference on*. IEEE, 2016.
- [6] Subedi, B., et al. "Secure paradigm for web application development." *RoEduNet Conference: Networking in Education and Research, 2016 15th*. IEEE, 2016.
- [7] CDAC Internal and External Security vulnerability developers' reports.

- [8] Othmane, Lotfi Ben, and Azmat Ali. "Towards Effective Security Assurance for Incremental Software Development the Case of Zen Cart Application." Availability, Reliability and Security (ARES), 2016 11th International Conference on. IEEE, 2016.
- [9] Hakim, Hela, Asma Sellami, and Hanene Ben Abdallah. "Evaluating Security in Web Application Design Using Functional and Structural Size Measurements." Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA), 2016 Joint Conference of the International Workshop on. IEEE, 2016.
- [10] Ríos, Jimmy Rolando Molina, et al. "Analysis Methodologies Web Application Development." International Journal of Applied Engineering Research 11.16 (2016): 9070-9078.
- [11] Kazim, Ali. "A Study of Software Development Life Cycle Process Models." International Journal of Advanced Research in Computer Science 8.1 (2017).
- [12] Sharma, Anuradha, and Praveen Kumar Misra. "Aspects of Enhancing Security in Software Development Life Cycle." Advances in Computational Sciences and Technology 10.2 (2017): 203-210.
- [13] Guidelines of Hosting Website at ERNET India | ERNET: http://www.ernet.in/services/guidlines_hosting.html
- [14] Guidelines for Indian Government Websites: <http://guidelines.gov.in/>

VI. Acknowledgments

The authors would like to thank Mr. Sumit Soman, Senior Technical Officer, Centre of Development for Advanced Computing, Noida, India for his evaluation and valuable inputs to the paper.



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

BIG DATA: A Survey paper on Recommendation System

Alok Barddhan¹ and Nidhi Jain²

School of Information Technology (So IT)

Centre for Development of Advanced Computing (C-DAC), Noida

B-30, Sector 62 Noida, Uttar Pradesh, INDIA

Abstract: *The recommended system helps users find items of interest. It can be defined as a subclass of information filtering system that displays a list of items based on user interests. Big data is about dealing with large amounts of data. In modern times, the number of customers, services and online information is increasing rapidly, so large data analysis generate a problem for the service recommendation system. With the increase in alternative services, it is recommended the user-preferred service has become an important research topic. Service recommendation system already reveals as an expensive tool to help users handle service overload and provide them with the appropriate advice. In this article investigates content-based, collaborative, hybrid methods, and related recommendations describe the study of the recommended system.*

Keywords: -BIG DATA, Recommendation System, Content Based filtering, Collaborative Filtering and Hybrid Approach.

I. Introduction

The technology is grown significantly. New technology, equipment and communication methods like social networking sites, the amount of human production is growing day by day. Big data means that the real big data, it is a collection of large data sets that can not be processed using our old computing technology. Big data is not just data, but rather a theme that involves a variety of tools, techniques and frameworks. Big data is vital to our lives. It has become one of the world's best technology. Several other authors often refer to three large Versus of Big Data: volume, variety and velocity. Volume refers to the actual size of the analysis data set. Variety refers to the various types of data sets (structures, semi-structures, unstructured) may combine to generate new insights and Velocity to the frequency at which data is recorded and/or analyzed. In order to do and try to understand the concept of large data, "Mapreduce" and "Hadoop" the word is inevitable.

There are some benefits to the conceptual data, which is familiar to us. Using information in social media, such as consumer preferences and product ideas, these companies are planning their goals based on their stores. Use a previous patient history in the hospital to provide better service. Large-scale parallel processing of the database system and MapReduce, for the analysis of complex analysis may provide most or all of the data analysis. The accuracy of large data can lead to more confident decisions, better decision-making can bring good operational efficiency, reduce costs and reduce risk.

II. Recommendation System

Recommender System recommends items to the user. These items could be news items, music, movies etc. Now, a question is that how to user interacts with large number of catalog (items)? Answer is user interact with large number of items with the help of Recommended System. Recommendation Systems are the unique type of information filtering technique gives suggestion for item to the uses to user. A suggestion relate to different decisions making processes, such as what online news to read, what item to buy, or what music to listen to. User's profile to some reference characteristics is compared by Recommendation Systems. These characteristics may be form of the Content based approach (as information item) or the collaborative approach (user's social environment). Typically Recommender system produces recommendation in three way: Collaborative Approach, Content based Approach, and Hybrid recommender system. A web site may, for example, present potentially a vast number of alternative items. It is difficult for individual who lack of enough personal experiences to calculate such data. This is where Recommended Systems come into play. In recent years, the importance in recommendation system has increased dramatically, followings fact indicates that the Recommender systems have fun an important role in such a way very much rated Internet sites as amazon.com, youtube, netflix, yahoo, trip advisor, last.fm, and IMDb.

III. Basic Approaches

There are mainly two approaches: Content based Approach and Collaborative Approach. Other approaches (such as hybrid approach) also exist.

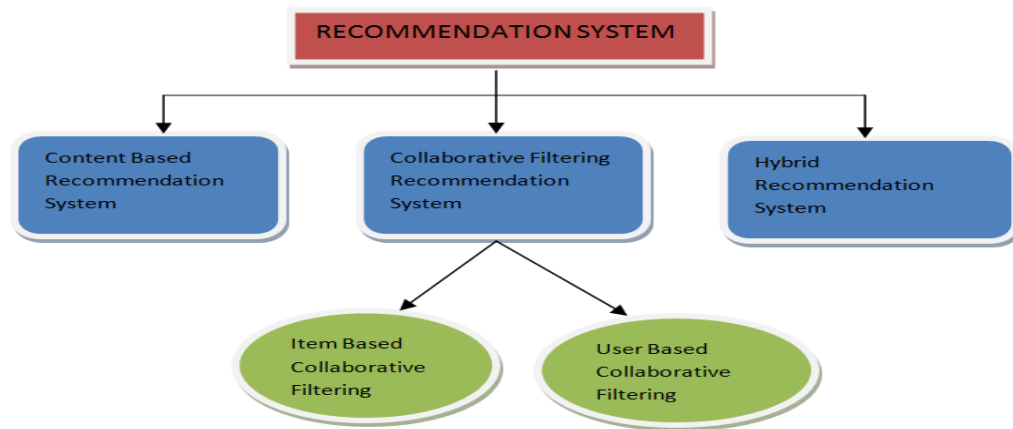


Figure1: Recommendation system

1.) Content Based Recommendation System

Content based filtering system is a system recommending similar items to the user which is like in past. This system focuses on algorithm, which accumulate users' preference into users' profile and item information into items' profile. Then recommend item which is similar to the user by similarity of their profile. A user profiles generally contains keyword (terms, features). Which collect the algorithm from item found interesting by the user. An item profile generally contain keywords are terms and features of the item itself. Actual profiles build the process which is handled by a variety of information retrieval. For example, Item's profile is represented by item which is the most common terms in the document. Different algorithms can be used for content-based filtering. The most commonly used technology is TF IDF, Naive Bayesian classifier etc.

Benefits and Disadvantages of content based recommendation:

ADVANTAGE:

Once there is information about the available items, it can start recommending.

DISADVANTAGE:

1. Items and attributes must be machine-recognizable.
2. We can not filter item on some opinion of quality, viewpoint/style. Because short the other people's experience, so the system can not make any assessment of a quality, viewpoint/style for the item.
3. No serendipitous items.
5. Synonymy. If spelling of two words is different but meaning is same – pure content-based filtering recognized them and will not find similarities.

2.) Collaborative Filtering Recommendation System

This recommendation system recommends user based on similar taste. There are two types of collaborative system: item based recommendation system and user based recommendation system. Item based recommendation system is dependent on same users' rating on other similar items. User based recommendation system is dependent on similar users to the same item rating recommendations.

A. User-user collaborative filtering

This approach was introduced in 1990s by the professor Jonathan L. Herlocker of University of Minnesota. This filtering approach selects a subset of user depended on the similarity with current users. Then predicts the users' rating based on its weighted ratings. Following steps are general:

After that, the weighted combination of its ratings is used to predict the user's rating.

1. Assign weighted similarity to all users based on similarity to current users.
2. Then choose a value k so that find out k similar users.
3. After that prediction calculate as target user is which is depend on the weighted function and same user rating k.

B. Item-item collaborative filtering

This is introduced by the University of Minnesota researchers in 2001. As the system grow then searching complexities of same users increase. So traditional recommendation does not scale good then another technique introduced named as item-item collaborative filtering which is find same items. In this item item similarity between pair of item i and j are calculated as offline using Pearson correlation:

ADVANTAGES:

- a.) The first advantage is that the Collaborative Filtering system can produce personalized suggestions because they believe that other people's experiences and suggestions are based on this experience.
- b.) Another famous advantage is that the Collaborative Filtering recommendation system can suggest contingencies by observing the behavior of similar people.

DISADVANTAGES:

- a.) If there are no ratings available. Then Collaborative Filtering systems cannot produce recommendations.
- b.) When the data on user ratings is small, their accuracy is poor. This and the previous shortcomings are called the cold start problem.
- c.) Collaborative Filtering systems are not content aware meaning that information about items are not considered when they produce recommendations.

2.) Hybrid Recommendation System

Content + collaborative = hybrid approach in a variety of ways. This system is used to progress the effectiveness of the recommender system. Use this method to make more accurate predictions. The many ways to combine the collaboration and the content are as follows:

- a.) Combine both (content and collaborative) separately and then make a prediction
- b.) Combine some content characteristics to collaborative.
- c.) Combine some collaborative characteristics to content.
- d.) construct a model which combine both characteristics of content and collaborative .

Limitations with Hybrid Approach: Hybrid suffers from limitations of both content based or collaborative filtering.

IV. Issues in Recommender Systems

A. Cold start problem

Cold start problem is basic problem in information filtering system as it states that system itself cannot draw any interpretation for items or users about which no sufficient information is present. Cold start problem implies that without gathering required information about the user system cannot make intelligent recommendations in both content based or collaborative filtering. Both these approaches may fail if no information present for user past history or no ratings provided to items previously.

B. Scalability

Scalability is the major problem in recommender systems as amount of data growing enormously leading to big data problem and handling such huge data became difficult for single machine or node .Therefore, for improving scalability of recommender systems many recommender systems started using hadoop map reduce framework (parallel processing paradigm).

C. Sparsity

Some people usually purchase or rate relatively few items compare with the total number of items. That leads to a sparse users-items representation matrix and therefore inability to locate neighbors or derive common behavior patterns, that's why final result is low-quality recommendations. This problem is addressed in latent factor models algorithms, which utilizes dimensionality reduction on items and users resulting in finding common behavior patterns in reduced dimensional space, which is not sparse. Matrix factorization methods proved efficiency of such reductions during the Netflix Prize competition, namely, the MF methods were applied to a 99% sparse matrix with 8.4 billion values missing.

D. Loss of neighbor transitivity

Assume that user A is highly correlated with user B, user B is highly correlated with user C. Possibly, user C is also highly correlated with user A. Such relationships are not captured by recommender systems, but can be captured with knowledge of users from, for instance, ontology. For example, people aged 22-70 are correlated as adults, when people aged 3-5 as children.

E. Recommending the items in long tails

Long tail basically consist of small number of items popular having well known hits and rest located in heavy tail i.e. items that are not being sold so well. So, it's a challenge for many recommendation systems whether they are ready to assist in discovery task by providing recommendation of unpopular or hidden gems in long tail. The problem today is to filter and present right choice to the user as per their preference and even it is difficult

to filter tail product due to data sparsity problem. Retailers such as amazon.com and Netflix are basically working on long-tail phenomenon.

F. Accuracy of prediction

Accuracy metrics in recommendation systems are basically designed to judge the accuracy of prediction of individual items. Recommender systems help to find relevant information in large space as per user preference and improving their accuracy but recommendations that are most appropriate according to standard metrics sometimes may not be useful to users. So improving accuracy as per user preference became major challenge nowadays.

G. Ranking of recommendation

Ranking in most cases leads to better formulation of recommendation problems. Ranking makes crucial difference in importance of personalization as we look forward for ways of optimizing personalized model i.e. apart from considering popularity or rating prediction other features related to user can be thought of for improving user experience. Ranking helps to fetch only relevant results from enormous amount of data. Basic fundamental in ranking is sorting items by popularity which is quite effective. Way of doing so is by adding some sort of personalized feature that helps to produce different ranking for different user.

H. Privacy concerns

Recommendation systems usually deals with privacy concerns as sensitive information of user are being gathered. Building recommendation system using collaborative filtering may become privacy issue as user profile information is being used for making predictions. People became concerned about their privacy as information provided in recommendation systems, basically in social networks are being misused easily. To eliminate privacy risk many recommender systems are making privacy-sensitive data inaccessible by encryption of data which does not allow outsiders to access user-sensitive information. Much research has been conducted to solve this issue.

I. Diversity of recommendation

Another important aspect of recommendation quality is diversity of recommendation. Many times user may be more satisfied with recommendations when diverse recommendation is being made. For example: song recommended by different authors. Many e-commerce sites like amazon.com recommend items based on likings or list of items present in user profile. For example- information retrieved from user profile mentions that user like songs of artist named x so to improve accuracy recommender system recommend other songs of artist x in spite of considering diverse user preferences i.e. user may also like songs of artist p, q, r so considering diverse user preference is also a major challenge in recommender systems.

V. Literature survey

To perform an analysis of various existing Recommendation approaches which recommends some items to the user in Big Data environment, brief description of the Recommendation is given below

Shunmei Meng, Wanchun Dou, Xuyun Zhang, and Jinjun. [1]et.al 2014 proposed an KASR method which is point out user's preferences and value of candidate service. Generate appropriate recommendations depend on the user's CF algorithm. It is designed to calculate the personal rating of every candidate service of the user then presents a list of personal service recommendation and recommends the more suitable service to him. In order to enhance the scalability of KASR in the big data atmosphere, they implemented this in the Map Reduce structure of the Hadoop platform. Limitation of this survey paper It cannot distinguish between user's optimistic and pessimistic preference from their review which are to build guess more accurate.

Bartosz kupisz, Olgierd Unold. [2]et.al 2015 develop item based Collaborative filtering recommendation algorithm and based on Hadoop and Spark. In this approach, the Mahout library is used to provide an execution of the parallel algorithm in the machine learning of the hadoop environment. It focuses on categorization, combination and collaborative filtering algorithm.

According to the map reduce paradigm, the full execution of the parallel collaborative filtering algorithm is implemented in nine consecutive operations. for the duration of the execution of the program, the specific data from the disk is read nine time and then record. This has a major impact in system's response time.

That's why the assumption that the sparks platform's algorithm will lead to higher efficiency .The Spark environment is the open execution of RDD. The aim of the creators of the Spark environment is used to eliminate the restrictions of the Map Reduce paradigm.

Chunzhi Wang, Zhou zheng, Zhuang Yang. [3]et.al 2014 introduces to improved hybrid recommendation

algorithm and join mapreduce paradigm which is use the Hadoop platform. This paper experimented on eight pc machine to make Hadoop Clusters.

Bo He ,Hongyuan Zhang. [4]et.al 2016 presented library personalized recommendation methods and strategy of big data. a.) customer hierarchical design, and b.) the user recommend strategy and methods. The paper divided the user into advanced users and ordinary users. This paper shows that the accuracy of two recommendation algorithm are 0.67 and 0.86. High accuracy was based on MapReduce and association rules mining of library personalized recommendation method, the method of advanced users, need a more accurate recommendation.

Suhasini Parvatikar, Dr. Bharti Joshi. [5]et.al 2015, in this paper join the collaborative approach and association mining rules to solve data sparsity problem, to achieve better performance. Collaborative approach is used to find the similarities between items which will facilitate the systems to recommend the items and for filling the blank rating, we use association mining where necessary. Then use item based collaborative approach to apply the objective users' to the objective items. Therefore, the use of these two method be able to help control the data sparsity and cold start problem in recommended systems.

VI. Conclusion

The recommendation system opens up new opportunities to retrieve personalized information on the Internet. It also helps to alleviate the problem of information overload, which is a very common phenomenon in the information retrieval system, enabling users to access products and services that are not readily available to users in the system. This article discusses two kinds of traditional recommendation techniques and emphasizes their advantages and challenges to improve their performance across a variety of hybrid strategies.

VII. References

- [1] Shunmei Meng, Wanchun Dou, Xuyun Zhang, and Jinjun Chen, Senior Member, IEEE, "KASR: A Keyword Aware Service Recommendation Method on MapReduce for Big Data Applications", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, DECEMBER 2014.
- [2] Bartosz kupisz, Olgierd Unold , "Collaborative Filtering Recommendation Algorithm based on Hadoop and Spark", International Conference on Industrial technology, IEEE 2015.
- [3] Chunzhi Wang, Zhou zheng, Zhuang Yang , " The Research of Recommendation System Based on Hadoop Cloud Platform", 9th of International Conference on Computer Science & Education (ICCSE 2014), IEEE 2014.
- [4] Bo He, Hongyuan Zhang, " Library Personalized Information Recommendation of Big Data", International Conference of Online Analysis and Computing Science, IEEE 2016.
- [5] Suhasini Parvatikar, Dr. Bharti Joshi, "Online Book Recommendation System by using Collaborating filtering and Association mining", International Computational Intelligence and Computing Research, IEEE 2015.
- [6] <http://recommender.no/info/problems-challenges-recommender-systems/>.
- [7] <http://recommender-systems.org/content-based-filtering/>
- [8] <https://www.coursera.org/specializations/recommender-systems/>
- [9] Poonam Ghuli, Atanu Ghosh, Dr. Rajashree Shettar, "A Collaborative Filtering Recommendation Engine in a Distributed Environment", 2014 International Conference on Contemporary Computing and Informatics (IC3I).
- [10] Santhini M, Dr. Balamurugan M, Govindaraj M, "Collaborative Filtering Approach for Big Data Applications Based on Clustering", International Journal of Recent Research in Mathematics Computer Science and Information Technology.
- [11] Jai Prakash Verma, Bankim Patel, Atul Patel, "Big Data Analysis: Recommendation System with Hadoop Framework", International Conference on Computational Intelligence & Communication Technology, IEEE 2015.
- [12] Shunmei Meng, Xu Tao, Wanchun Dou, "A Preference-Aware Service Recommendation Method on Map-Reduce", 16th International Conference on Computational Science and Engineering, IEEE 2013.

Restoration of Mural Images

Gunjan Mishra¹ and Tushar Patnaik²

School of Information Technology (So IT)

Centre for Development of Advanced Computing (C-DAC), Noida

Uttar Pradesh, India

Abstract: A mural is a wall painting, an artwork that is painted or applied directly on the wall or any other large surface area. Old mural images can deteriorate, get distorted, develop cracks, fade away and may even peel out due to various reasons including social, climatic, environmental, historical factors. An approach to virtually restore these mural images using the Digital Image Processing technology which tries to generate the original image. The suggested approach consists of four major steps which are described further in this paper. An Edge Enhancement process is implemented followed by K means clustering and averaging is performed. The final step is to perform Histogram equalization and its comparative analysis with other enhancement techniques like Sharpening and Adaptive histogram equalization. The results of the experiment are good and are providing improved images based on image restoration parameter Peak signal to noise ratio (PSNR). This implemented approach can also be used in future to restore faded deteriorated mural images.

Keywords: K means clustering, averaging, histogram equalization, adaptive histogram equalization, PSNR.

I. Introduction

Digital image processing has been popular for detection and recognition of different images. The different algorithms used in digital image processing solve the purposes for detecting and cleaning the noises and pixels in various image.

Wall painting is a human creation which depicts a culture expressions which exist from the ancient day till present day. Any kind of deterioration or destruction of these mural paintings can cause significant harm to our cultural heritage. Due to different climatic changes and other external impact of environment the wall paintings sometimes get distorted, the intensity of the colors of painting seems to be faded and since the perception capability of our eye is limited it becomes difficult to identify the details of faded image.. To preserve the history and diverse culture, the mural images are required to be restored as the original.

The mural painting recovery i.e the wall painting restoration to originality is a major challenge. Since lot of algorithms and filtering techniques have been developed for restoration of such type of images different research papers have given different approaches and results on different types of images but the look of images to the originality requires lot of analysis and computations. The degraded image was detected keeping ground truth data in view. The process of reconstructing a blurred, damaged or a noisy image to provide an uncorrupted image is termed as image restoration.



Figure 1. Few Deteriorated Mural Image

Restoration of Mural Images can be used for preservation of historical assets. A system that could restore mural images can become a contribution to heritage conservation societies which is built to conserve the national heritage. Building a system that could restore Mural images using digital image processing which is a technical

domain could build a bridge between the technical and art communities. Mural image preservation is not only a contribution to national heritage but also it can be used to infer historical discoveries. A number of old destroyed images carry building blocks for historical discoveries, a mural restoration system could help in reconstructing such images thereby providing an restored image which is close to original image. The restored image could provide the details that the deteriorated mural image failed to reveal

II. Literature review

Karianakis, N., et al. [2] had worked on an approach which was focused on restoring the missing parts of old wall paintings. They applied an algorithm for seamless image stitching of the missing area. In addition to this they also applied TV inpainting. TV inpainting was applied for extracting area and also for repairing the image.

S. Awate ,et al. introduced an adaptive filter, this filter would restore the images by finding out there statistical information. The filter was found to be capable of restoring different types of images.

Bhabatosh Chandra, et al. [1] proposed a patch matching technique. This technique will use a database which will have clean paintings , wherever there will be a patch that is to be filled that patch will be replaced by another patch which will be the best patch found out of all.

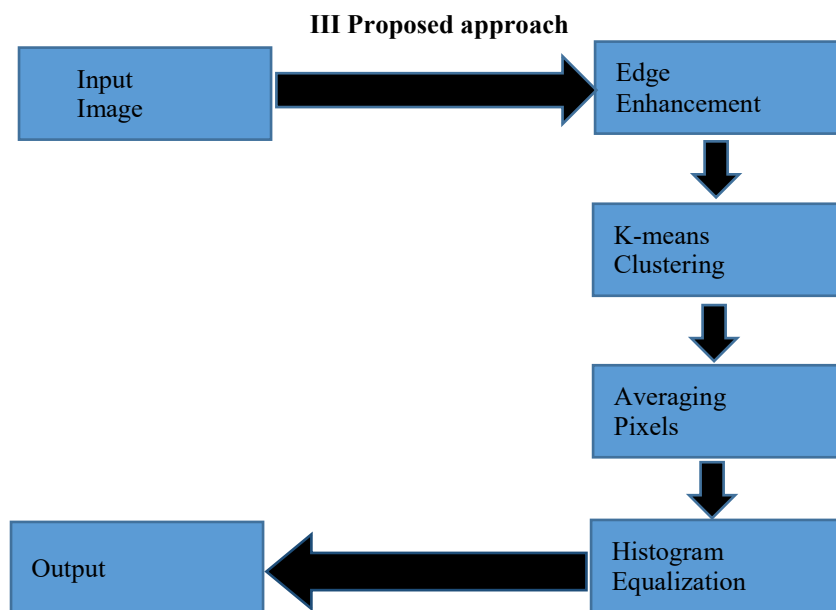


Figure 2. Proposed Approach

A. Input image

Input images are deteriorated mural wall paintings which are to be restored Given the ground truth data and if we have visual comparison with the deteriorated images the wall paintings are totally unclear and extremely deteriorated. The actual content of the image is totally hidden hence the restoration methods should adaptively select the low intensity pixels and enhance them.



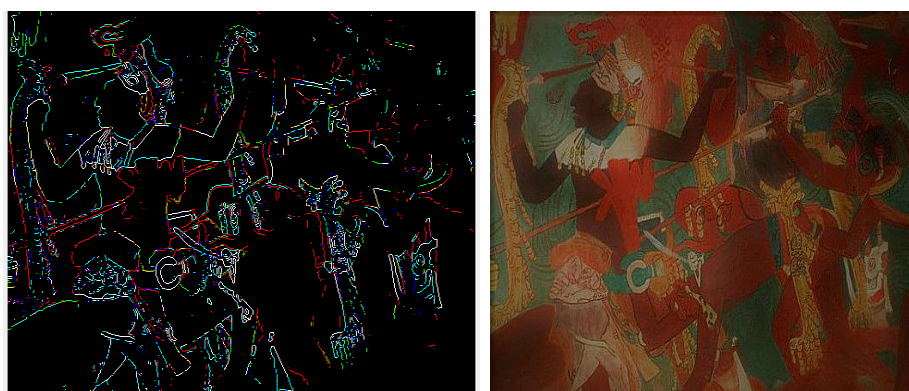
Figure 3. (a) Deteriorated Mural Images



Figure 3.(b) Ground Truth Data

B. Edge Enhancement

Since edges are the boundary of an defined structure colored image the first step to detect an image is the edge enhancement. For Edge enhancement edges are first extracted from the input deteriorated mural image. In this step we need to detect lines, these lines can be edges or other parts of the image. The next step was to use a standard operator sobel to detect and enhance edge. Sobel operator computes the approximation of gradient of the image intensity function. It uses a set (usually two) 3×3 kernels which are then to be convolved with the original image and this is done to calculate the approximations of derivatives.



C. K-means Clustering and Averaging pixels

K-means Clustering subdivides an image into multiple clusters based on the colors. A mean value is calculated for each cluster and averaging is performed by assigning the mean value to all the pixels of the cluster. The smoothening of colors is done by segmenting the image into different clusters using k-means clustering. All of the pixels that are present in each of the cluster are then replaced by the calculated k-mean value of the cluster. At the end of clustering and averaging pixels a smooth image is obtained in which colors are also restored.



D. Histogram Equalization and Adaptive Histogram Equalization

Histogram equalization is applied to enhance and improve mural image for restoring. Histogram equalization is used to enhance mural image for restoring details, colors and contrast further. A sharpening mask can also be applied to the image at the final step but it didn't give good result. The best results were given by adaptive histogram





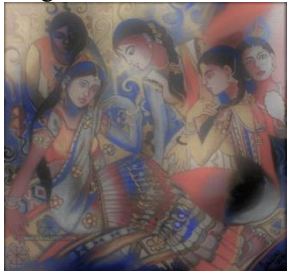







equalization. Histogram equalization is a digital image processing technique that is used for manipulating intensity of the images, it is applied for contrast enhancement. This method is useful in images with backgrounds and foreground that are both bright and dark.

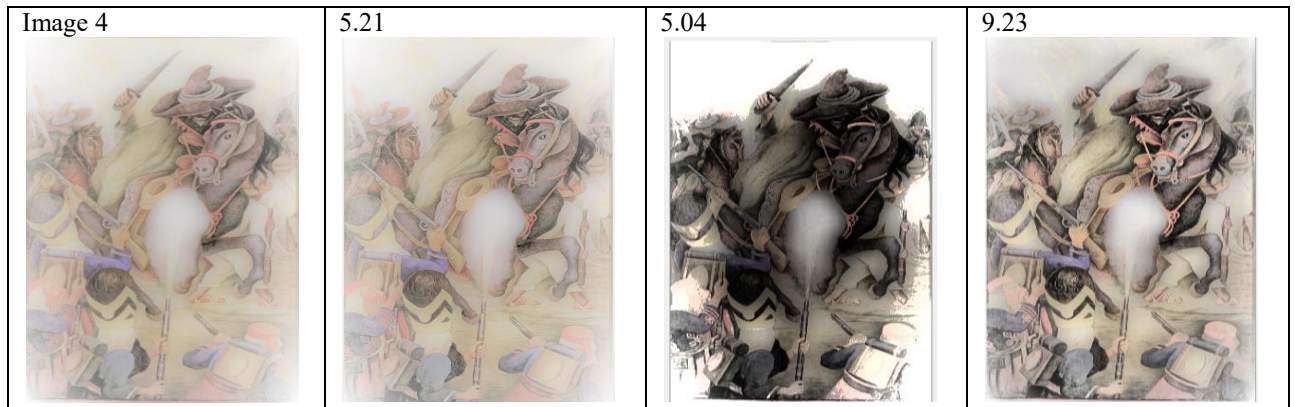
Since the evolutions of the result is measured through PSNR, the increment of PSNR value was not consistent hence Adaptive Histogram Equalization was used. The result turned out to be better i.e histogram equalization increased PSNR for 18 mural images out of total 44 mural images whereas adaptive histogram equalization increased PSNR for 35 mural images out of total 44 mural images.



III. Results

In this paper we have applied methodology to restore deteriorated mural images. The images are successfully restored using the above implemented methodology.

Mural Image	Sharpening Mask	Histogram Equalization	Adaptive Histogram Equalization
Image 1 	18.20 	17.02 	19.20 
Image 2 	15.14 	15.46 	16.21 
Image 3 	15.17 	16.07 	15.5 



IV. Conclusions

The approach improves the quality of the distorted image. The lines, colors, contrast and the structure of the image can be restored by this methodology. The image restoration success rate is concluded by the image restoration parameter PSNR(Peak Signal to Noise Ratio) adaptive histogram equalization is improving PSNR of most of the images. The adaptive histogram equalization algorithm is restoring 80% whereas conventional histogram equalization restored only 40% deteriorated images.

The implemented approach is restoring most of the deteriorated images except the images that are completely washed out, had wide cracks or missing parts.

V. References

- [1] Bhabatosh Chanda, Dhruv Ratna, B.L.S. Mounica, "Virtual Restoration of Old Mural Paintings using Patch Matching Technique", 2012 Third International Conference on Emerging Applications of Information Technology (EAIT), 978-1-4673-1827-3/12, 2012 IEEE.
- [2] Karianakis, N., & Maragos, P., "An integrated system for digital restoration of prehistoric Thera wall paintings" IEEE International Conference on Digital Signal Processing, pp. 1-6, 2013.
- [3] S. Awate and R. Whitaker, "Unsupervised, information-theoretic, adaptive image filtering for image restoration", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.-28, NO.-3, March 2006.
- [4] Rohit T. Pushpalwar, Smriti H. Bhandari, "Image Inpainting Approaches – A Review," Department of Computer Science and Engineering Walchand College of Engineering, Sangli, 2016 IEEE 6th International Conference on Advanced Computing, India, 978-1-4673-8286-1/16, 2016 IEEE..
- [5] S. C. Pei, Y. C. Zeng and C. H. Chang, Virtual Restoration of Ancient Chinese Paintings Using Color Contrast Enhancement and Lacuna Texture Synthesis, IEEE transactions on image processing, Vol. 13, pp.416429, 2004.
- [6] B. Chanda and Pulak Purkait, Digital Restoration of Damaged Mural images, The 8th Indian Conference on Vision, Graphics and Image Processing, 2012.
- [7] A. Buades, B. Coll, and J.-M. Morel, A non-local algorithm for image denoising, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 2, June 2005.
- [8] B. Chanda and D. Dutta Majumder, Digital Image Processing and Analysis, PHI Learning, New Delhi, 2011.