

**Proceedings**  
**of**  
**FIRST INTERNATIONAL CONFERENCE ON RECENT  
TRENDS IN COMPUTING**

August 30-31, 2016



***Patron***

***Dr. Jaspal Singh***  
***(Vice Chancellor)***

***Chairperson***

***Dr. Lakhwinder Kaur***

***Conference Chair***

***Dr. Raman Maini***

***Conference Co-Chair***

***Dr. Amardeep Singh***  
***Dr. Himanshu Aggarwal***

***Organizing Secretary***

***Dr. Jaswinder Singh***

***Er. Jagroop Kaur***

***Er. Harpreet Kaur***

***Proceeding Editors***

***Er. Amrit Kaur***  
***Er. Abhinav Bhandari***  
***Er. Navdeep Kanwal***

*Organized by*

***Department of Computer Engineering, Punjabi University,  
Patiala, Punjab, India – 147002***

International Conference on  
**Recent Trends in Computing**  
**(ICRTC-2016)**

**Editors**

**Er. Amrit Kaur**

**Er. Supreet Kaur**

**Dr. Williamjeet Singh**

**Er. Abhinav Bhandari**

**Er. Navdeep Kanwal**

**Er. Gurpreet Singh**



Department of Computer Engineering,

Punjabi University, Patiala – 147002

*Web:* <http://www.csepup.ac.in/icrtc/2/>

*Email:* icrtc16@gmail.com

## **Patron Message**

It is a matter of great pleasure to see the Department of Computer Engineering organizing its first International conference on RECENT TRENDS IN COMPUTING (ICRTC- 2016) during 30<sup>th</sup> -31<sup>st</sup> August 2016. I congratulate the Department of Computer Engineering for organizing this conference. I could see the amount of efforts put in by faculty in organizing this conference in this university. It is very much heartening to see the immense response received by the conference from the research community for its very first edition. A good number of distinguished professors and researchers have also agreed to deliver keynote addresses/ invited talks in the conference. Young scholars participating in the conference will immensely benefit from these.



I heartily welcome all the distinguished speakers, scholars presenting papers and the participants to this first international conference organized by the Department of Computer Engineering.

**Dr. Jaspal Singh**

*(Vice Chancellor)*

## **Chairperson Message**

The International Conference ‘RECENT TRENDS IN COMPUTING’ will unveil the latest Technology Trends in Computing, where the research is still going on or the significant research is done by the researchers, scientists, engineers, and scholars from the various fields.



The ICRTC 2016 Conference will provide a wonderful forum for you to refresh your knowledge base and explore the innovations in latest trends in computing. The Conference will strive to offer plenty of networking opportunities, providing you with the opportunity to meet and interact with the leading scientists and researchers, friends and colleagues as well as sponsors and exhibitors.

We are thankful to the sponsoring agencies for their generous support. Thus ICRTC-2016 has planned well to roll out a good set of keynote speeches, invited talks and presentations by contributed authors. We earnestly request all the participants to make use of this effort to derive maximum benefit of the event. I wish the conference a real success.

**Dr. Lakhwinder Kaur**

*(Professor & Head, Department of Computer Engineering)*

## **Conference Chair Message**

Welcome to the proceedings of the First CONFERENCE ON RECENT TRENDS IN COMPUTING (ICRTC- 2016)" at Department of Computer Engineering in the Punjabi University, Patiala on 30th-31<sup>st</sup> August 2016. This is the first time this Conference is being held in Department of Computer Engineering, Punjabi University Patiala, the royal city of Punjab, India. The objective of this event is to promote research in the field of image processing, computer networks, network security, data mining and many more.



In this inaugural year, a total of 167 technical papers were received by us. After a rigorous review process, the Program Committee selected 106 papers for oral presentation and 61 papers for poster presentation. We would like to thank the authors of all the papers for submitting their quality research work to the conference. Special thanks go to the Program Committee members and the reviewers for sparing their time in carrying out the review process meticulously.

We would like to thank "International Journal of Engineering Sciences" kindly permitting us to publish the proceedings of the conference online on their website. We were really fortunate to have a number of very eminent experts delivering keynote and invited talks at the conference. In spite of its relatively short time, it was really heartening to see very good attendance by the delegates which resulted in very fruitful deliberations.

We would like to thank the Conference Patron, Dr. Jaspal Singh, Vice Chancellor of the University, the Conference Chair person, Dr. Lakhwinder Kaur, for putting in a lot of effort and supporting us with positive feedback all throughout the process of organizing.

**Dr. Raman Maini**

*(Professor, Department of Computer Engineering)*

### **Advisory Committee**

Dr. Vijay K. Arora, Professor, Wilkies University, USA

Dr. S. C. Saxena, Ex- director IIT Roorkee

Dr. Vinod Kumar, Ex. Dy. Director, IIT Roorkee

Dr. Ejaz Ahmed, Professor, NIT, Sri Nagar, INDIA

Dr. Savita Gupta, Professor UIET, PU Chd.

Dr. Sukhwinder Singh, Professor UIET, PU Chd.

Dr. Deepak Garg, Professor & Head CSE, TIET, Patiala

Dr. Rinkle Rani, Asso. Professor, TIET, Patiala

Dr. Yogesh Chhaba, Professor, GJU, HISAR

Dr. Ragina Bhandari, SMO, PUP

Dr. Manjeet Singh Patterh, Professor, PUP

Dr. Gurmeet Kaur, Professor & Head ECE, PUP

Dr. Manjit Singh Bamrah, Professor PUP

Dr. Hardeep singh, Professor, GNDU

Dr. Kanwaljeet Singh, Director Computer Centre, PUP

Dr. G.S. Lahal, Prof. DCS and Dean, PUP

Dr. Vishal Goyal, Associate Prof, DCS

Dr. Gurpreet Singh Joshan, AP DCS, PUP

Dr. Jaimal Singh Khamba, Professor ME, PUP

Dr. I.P.S. Ahuja, Professor ME, PUP

Dr. Vinay Gupta, Professor ME, PUP

Dr. Sanjeev Puri, Professor Physics, PUP

Dr. Damanpreet Singh, Asso. Prof., SLIET, Longowal (Pb.)

Dr. Major Singh, Asso. Professor, SLIET, Longowal (Pb.)

## **Technical Committee**

Er. Kanwal Preet Singh	Er. Gaurav Deep
Er. Harmandeep Singh	Dr. Gaurav Gupta
Er. Madan Lal	Er. Ram Singh
Er. Jasvir Singh	Er. Sumandeep Kaur
Er. Nirvair Neeru	Dr. Williamjeet Singh
Er. Rakesh Singh	Dr. Dhavleesh Rattan
Er. Bramhaleen Kaur	Er. Navroz Kaur
Er. Gurjot Singh	Er. Priyanka Jarial
Er. Anantdeep	Er. Supreet Kaur
Er. Amrit Kaur	Er. Karandeep Singh
Er. Lal Chand	Er. Charanjiv Singh
Er. Navdeep Kanwal	Er. Navjot Kaur
Er. Navdeep Singh	Er. Gurpreet Singh
Er. Sikander Singh	Er. Navneet Kaur
Er. Abhinav Bhandari	Er. Neelofer Sohi

## Organizing Committees

<b>Registration Committee</b>	Er. Madan Lal, AP - Coordinator Er. Anantdeep, AP – Co-coordinator Er. Amrit Kaur, AP Er. Gurjit Singh Bhathal, AP Er. Charanjiv singh, AP Er. Navneet Kaur, AP Er. Supreet Kaur, AP Mrs. Parveen Kumari & Mr. Rajesh Kumar (Office)
<b>Stage Arrangement</b>	& Er. Jagroop Kaur, AP - Coordinator
<b>Reception Committee</b>	Er. Harpreet Kaur, AP – Co-coordinator Er. Brahmaleen Kaur Sidhu, AP Er. Navroz Kahlon Er. Nilofar Sohi, AP Er. Charanjeev Sarao, AP Mr. Arun Kumar (LA) & Mr. Kashmir Singh (SA)
<b>Hospitality, Transportation Committee, Boarding and Lodging</b>	Er. Sikander Singh Cheema, AP - Coordinator Er. Lal Chand, AP – Co-coordinator Er. Sumandeep Kaur, AP Er. Gauravdeep, AP Er. Neelofar Sohi, AP Er. Navneet Kaur, AP Mr. Arun Kumar & MR. Paramjit Singh (LA)
<b>Printing Committee</b>	Er. Kanwalpreet Singh Atwal, AP- Coordinator Dr. Gaurav Gupta, AP – Co-coordinator Mr. Jashanpreet Singh, Assistant Programmer Mr. Hardeep Singh Ahuja, LA
<b>T.A./D.A. Committee</b>	Er. Kanwalpreet Singh Atwal, AP- Coordinator Er. Priyanka Jariyal, AP – Co-coordinator Er. Ram Singh, AP Er. Karan Singh, AP Mr. Nirbhay Singh (LA), Mrs. Lakhwinder Kaur & Mr. Jaswinder Singh(office)
<b>Website</b>	Er. Navdeep Kanwal ,AP – Coordinator

	Er. Abhinav Bhandari ,AP
	Er. Charanjiv singh, AP – Co-coordinator
	Mr.Jashanpreet Singh, Assistant Programmer
<b>Technical Event Committee</b>	Dr. Jaswinder Singh, AP (Coordinator)
	Er. Brahmaleen Kaur Sidhu, AP (Co-coordinator)
	Er. Rakesh Kumar, AP
	Er. Abhinav Bhandari, AP
	Er. Navdeep Singh, AP
	Er. Gauravdeep, AP
	Er. Ram Singh, AP
	Er. Sumandeep Kaur, AP
	Dr. Williamjeet Singh, AP
	Dr. Dhavleesh Rattan, AP
	Er. Supreet Kaur, AP
	Er. Neelofar Sohi, AP
	Er. Navneet Kaur, AP
	Er. Navjot Kaur, AP
	Er. Gurpreet Singh, AP
	Ph.D. Research Scholars
	All CE Staff Members
<b>Conference Committee</b>	Er. Abhinav Bhandari, AP - Coordinator
	Er. Navdeep Kanwal, AP
	Dr. Williamjit Singh -- Co-coordinator
	Er. Gurpreet Singh, AP
	Er. Amrit Kaur, Ph.D. Research Scholar
	Er. Sukhpal Kaur, Ph.D. Research Scholar
	Er. Kuldeep Singh, Ph.D. Research Scholar
<b>Discipline Coordinators Committee</b>	Er. Harpreet Kaur, AP
	Er. Gurjit Singh Bhathal, AP
	Er. Sikander Singh Cheema, AP
	Er. Sumandeep Kaur
	Er. Karan Singh
	Mr. Randip singh (LA), Mrs. Surinderjit Kaur
	Mr. Kashmir singh (JTA)

<b>Purchase Committee</b>	Er. Jaswinder Singh, AP Er. Harmandeep Singh, AP - Coordinator Er. Madan Lal, AP Er. Gaurav Gupta, AP - Secretary Mrs. Satinderpal Kaur Supdt. Maintenance
<b>Finance Committee</b>	Er. Harmandeep Singh, AP - Coordinator Er. Gaurav Gupta, AP – Secretary
<b>Editing Committee</b>	Er. Amrit kaur, AP - Coordinator Er. Supreet Kaur, AP Er. Amrit Kaur, Ph.D. Research Scholar Er. Sukhpal Kaur, Ph.D. Research Scholar Er. Kuldeep Singh, Ph.D. Research Scholar Mrs. Navneet Kaur (office)
<b>Sponsorship Collection Co-coordinators</b>	Er. Sikandar Singh, AP Er. Navdeep Kanwal, AP Er. Lal Chand, AP Er. Gauravdeep, AP Er. Dhavleesh Rattan, AP
<b>Publicity Committee</b>	Er. Gurjit Singh Bhathal, AP – Coordinator Er. Rakesh Singh, AP Er. Dhavleesh Rattan, AP Er. Gurpreet Singh Er. Sumandeep Kaur

## Research Papers

# BAD SMELL DETECTION AND REFACTORING TO IMPROVE CODE QUALITY

Sukhdeep Kaur<sup>1</sup>, Dr. Raman Maini<sup>2</sup>

*Department of Computer Engineering, Punjabi University, Patiala-147002 (India)*

<sup>1</sup>sukhdeep\_kaur01@yahoo.com

<sup>2</sup>research\_raman@yahoo.com

**Abstract-** The term refactoring is a crucial part of software engineering that changes the internal structure of source code with less cost of software maintenance to improve overall design of software. Design or code problem in software indicates the bad smell. Basically, bad smell is a sign of badly written code by programmer that makes the code more difficult to change and maintain in future. Bad smells prioritizing the code refactoring to improve software readability, modularity and reusability. In this paper, before applying refactoring techniques, bad smells are detected using some object-oriented metrics such as source lines of code, cyclomatic complexity, Average lines of code and number of parameters. Metrics are used to evaluate the various aspects of complexity of software code. The graphical user interface (GUI) window based applications are developed to detect bad smells in c# code using visual studio ultimate tool 2012. From simulation of refactoring techniques, it has been observed that extract method refactoring technique is easier and best way to remove long method bad smell for reducing complexity and duplicate code.

**Keywords:** *software refactoring, bad smell, object-oriented metrics, refactoring techniques.*

## 1. INTRODUCTION

Software engineering is a process of development, design and maintenance of software. In software engineering, software systems are revised time to time due some reasons such as user requirements, advance technologies and cost benefits. Without notable little changes in program's code are done by developers that degrades the internal code quality and design structure of software. To enhance the maintenance of code, term restructuring or refactoring is used. The restructuring is transformation from one framework to another at the identical comparable abstraction level to maintain the existing state. Restructuring creates new version of existing code in traditional way to improve the software maintenance. Refactoring is an object-oriented variant of restructuring. Basically, refactoring is a technique used to modify the existing program code structure of software without occur any changes in the external behavior. The main focus of refactoring is to maintain and reusable code in future. Without refactoring the structure of code will decay, so it helps to preserve the shape of code. Before apply the refactoring technique to source code, it is essential to check when and where code requires the refactoring.

## 2. BAD SMELL

Bad smell is symptom of deep-rooted design problem in code. Bad smells does not provides any effect on external behavior of software but presence of bad smells in code, makes it hard to modify. It does not include any error at the execution time. Basically, bad smell is a sign of badly written code by programmer that makes the code more difficult to change and maintain in future. It shows the complexity in code. In a code, bad smells are detected using different types of object-oriented metrics. After detection of bad smells, code requires refactoring to enhance the software features.

## 2.1. Different Types of Bad Smells

### a) Long Method

More lines of source code in a particular method consider a long method bad smell. It is hard to change and reduce the reusability of code. Long methods are extracted into small methods because small methods are easier to troubleshoot, read and understand.

### b) Refused Bequest

When all data and methods inherit from super-class into sub-class, but sub-class does not fully supports the functionality inherit from its super-class, it indicates the refused bequest bad smell. To keep code clarity and software maintainable, it is necessary to remove refused bequest bad smell from code.

### c) Lazy Class

Additional class's increase the complexity and takes the more time to complete execution process of project. When class has does not enough to earn functionality of code, it is necessary to merge small code functionality class with similar class functionality and after that it is removed from project.

### f) Long Parameter List

More than five parameters are passing into one method signature indicates the long parameter list bad smell. Parameters that are not necessary for method functionality but declare in the parameter list make the more complex, inconsistent and less reusable to code.

### g) Message Chain

Message Chain means one class method calls to second class method which in turn that class calls to third class method and so on. Sequence of calling methods more than two indicates the message chain bad smell. For reducing chain dependency between classes and code bulk, it is necessary to remove message chain bad smell.

### h) Comments

More lines of comments in code reduce the code clarity. Comments are not necessarily a bad smell, but they can be misused to poorly structured code.

## 3. REFACTORING

Refactoring is a familiar procedure used to improve the internal code quality, readability, extendibility and modularity of software by disposing the bad smells from source code. It helps to diminish the debugging time and also provides the fast execution process of code. In software applications, refactoring methods are acquired for different artifacts such as source code, UML, database and models.

## 3.1. Refactoring Techniques

The refactoring techniques are reorganizing for variables, methods and classes beyond the class hierarchy to facilitate future transformations. Different types of refactoring techniques are defined as following:

### a) Extract Method

Extract method is the most simple and best way to refactor the code. Extract method is used to restructure the selection of prior lines of code into new method. New methods can be parameterized or non-parameterized based on the selection of existing code. Extract method provides the more readable and less duplicate code. It also helps to boost the modularity of code.

*b) Encapsulate Field Method*

Encapsulate field method is used to protect class field from direct access by outside world. In the code, if public field is present then encapsulate field method make it private field and create access methods for it. This refactoring method is used for security purpose.

*c) Rename Method*

Rename refactoring method is used to rename the identifier for code symbols such as variables, fields, namespaces, methods, etc. to reveal out their functionality. This refactoring method provides more efficient and readable code.

*d) Remove Parameters*

All parameters are proceeding into particular method is hard. Remove parameter refactoring method is an easy way to remove number of parameters from methods and constructors. These refactoring methods extract new method and pass it as an object reference in parameter list.

### 3.2. Refactoring Process

Various steps are used to refactor the source code are defined as following:

1. Analyze the source code.
2. Identify the bad smell behavior in source code using software metrics rules.
3. Determine how to simplify bad smells.
4. Select and apply suitable refactoring techniques to remove the bad smell.
5. Assess the effect of refactoring on quality characteristics of software or process.
6. Repeat steps until the smell is gone from code.

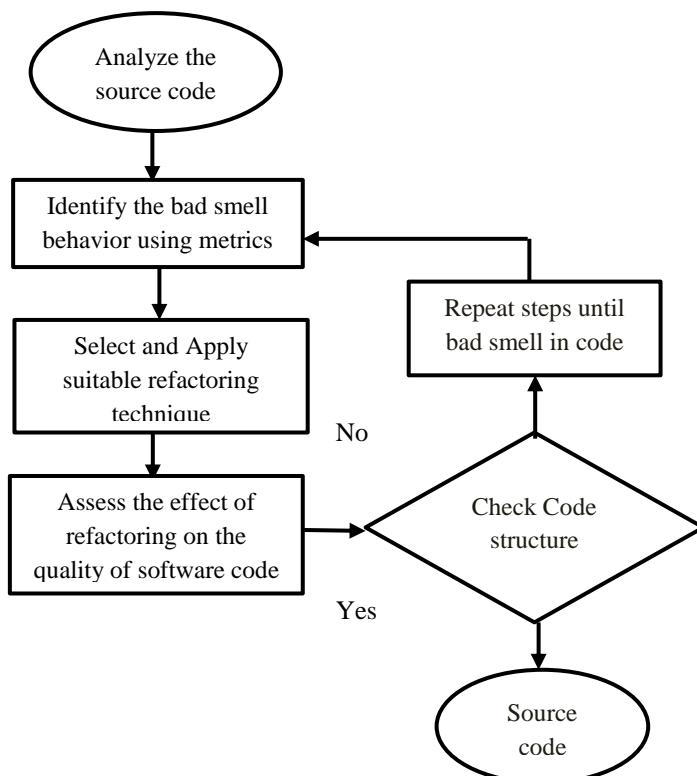


Figure1. Steps for refactoring of code

#### 4. SIMULATION

The visual studio ultimate tool 2012 with framework 4.5 is used to create GUI window based applications. The GUI window forms are developing to detect bad smells in c# code. C# is part of .NET language that supports the multiple inbuilt languages. Following object-oriented metrics are used to detect three types of bad smells:

- a) *Source Lines of Code*: It counts only logical lines of source code, not includes the comment and blank lines.
- b) *Cyclomatic Complexity*: It measures independent paths in class methods through control flow graph structure.
- c) *Average Lines of Code*: Logical lines of code are divided by 3 contains the values of average lines of code.
- d) *Number of Parameters*: It counts the number of parameters are present in method or constructor signature.

##### 4.1. Bad Smell Detection and Refactoring

Three types of bad smells such as long method, comment lines in method and long parameter list are detected using object-oriented code metrics in project medicine inventory c# code. Medicine inventory project contains 30 classes.

###### A. Long Method Bad Smell Detection and Refactoring

1. Analyze the source code.
2. Apply metric rules to detect long method bad smell
  - a) Number of Source line of code (SLOC) > 50
  - b) Cyclomatic complexity > 10
3. If any above rules is/ are true, Long method bad smell is detected.
4. Apply Extract Method refactoring technique that is suitable to remove long method bad smell.

After uploading classes, select any one class to view class information, class all methods and check long methods in a selected class. In Fig 2. two long method bad smells are present in selected login class according to the code metric rules. Total long method bad smells in all uploaded classes are 16.

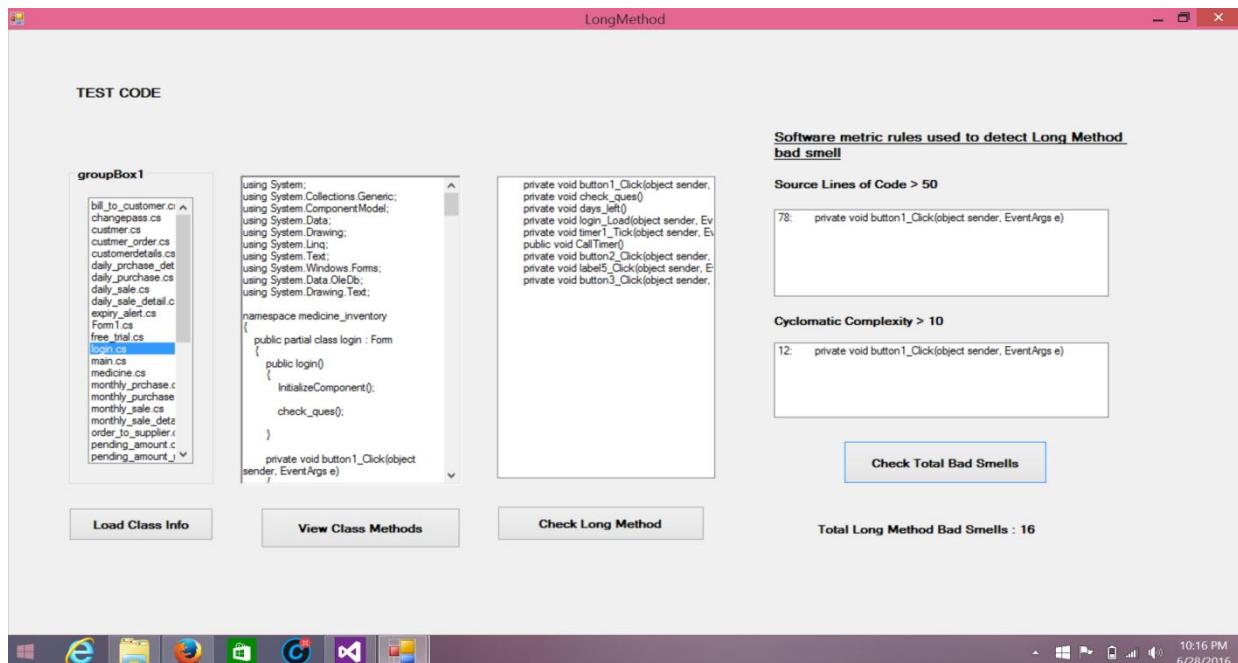


Figure 2. Long method bad smell detection

Selected login class that has two long method bad smells requires the refactoring to improve code quality. So, select the part of code where to apply the refactoring technique. Right click on selected part of code, choose the Extract method refactoring technique to refactor the code is shown in Fig 3. Long method is extracted into small new created method because small methods are more valuable as compared to long methods.

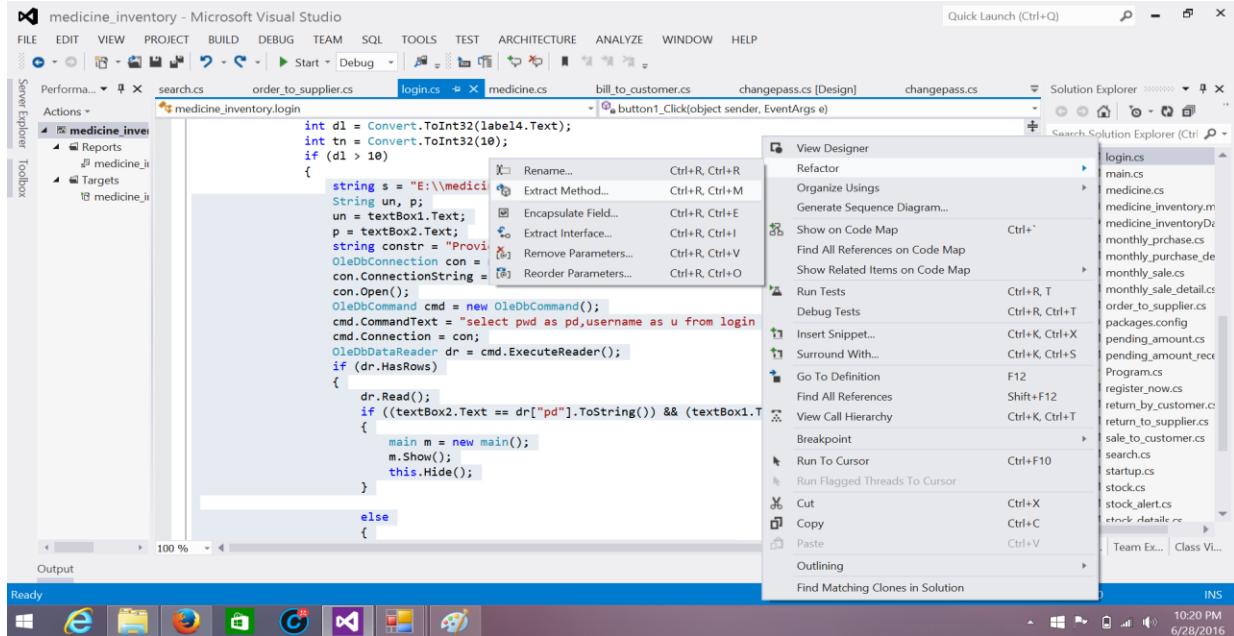


Figure 3. Select Extract method refactoring technique

Type the name of new created method according to the working of selected code. Fig 4. shows the name of new created method as loginPwd.

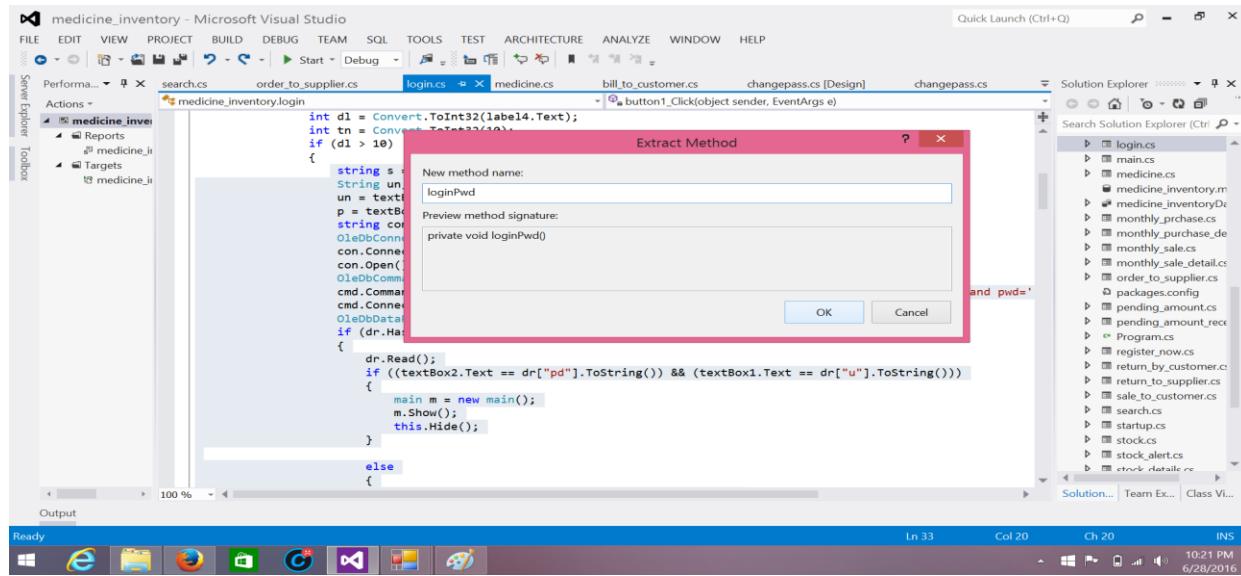


Figure 4. Type the name of new created method

Fig 5. shows the code after refactoring that removes duplicate code from long method by extracting new method like loginPwd.

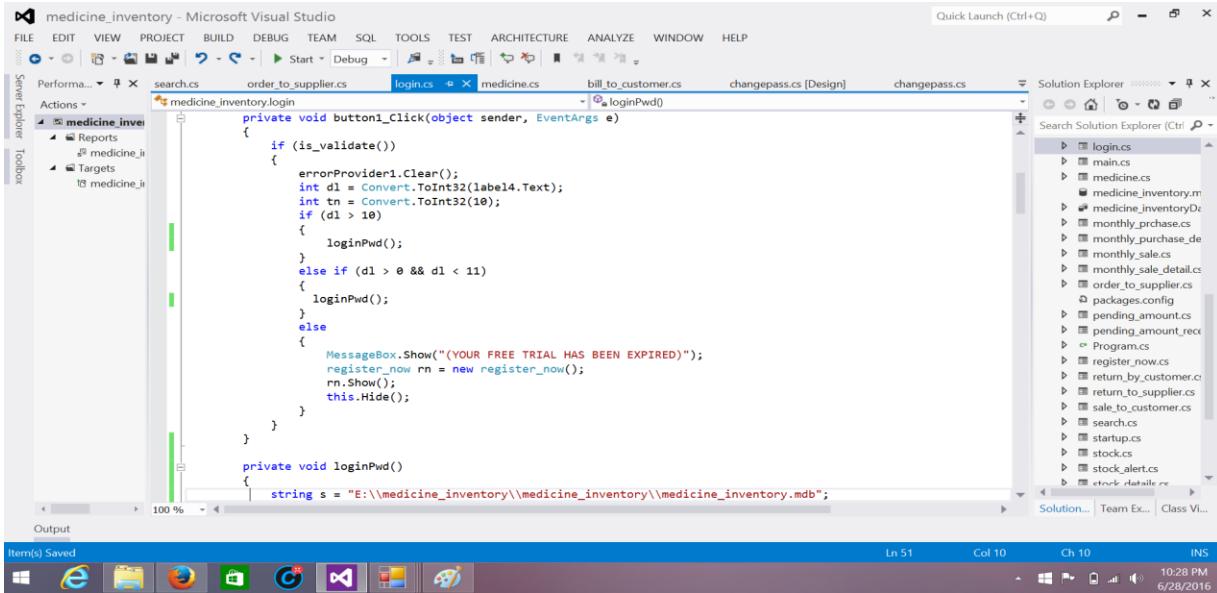


Figure 5. Source code after refactoring

#### B. Comment Lines Bad Smell Detection and Refactoring

1. Analyze the source code.
2. Apply metric rules to detect comment lines bad smell
  - a) Comment lines > Average lines of code
3. If above rule is a true, Comment lines in method bad smell is detected.
4. Code is refactored by simply removing the comment lines bad smell from method.

After uploading classes, select any one class to view class information, class all methods and check comment lines bad smells in a selected class. In Fig 6. two comment lines bad smells are present in selected customer order class according to the code metric rule. Total comment lines bad smells in all uploaded classes are 38.

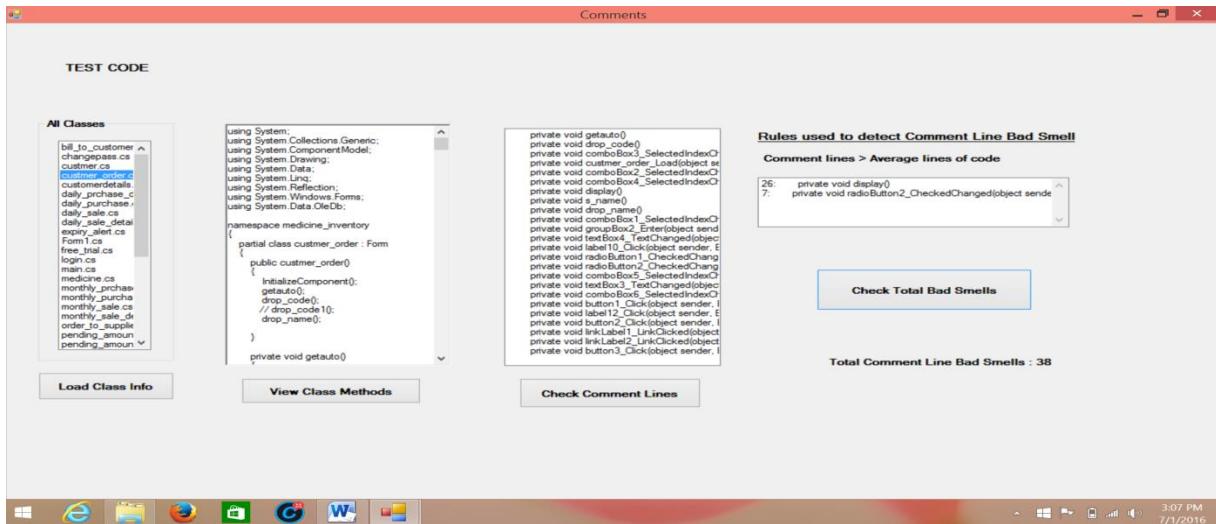


Figure 6. Comment lines in method bad smell detection

After detecting comment lines bad smell, it requires the refactoring for code to manually remove comment lines by programmer is shown in Fig 7.

```

// int sc = Convert.ToInt32(comboBox5.SelectedValue.ToString());
// string s = "E:\medicine_inventory\medicine_inventory\medicine_inventory.mdb";
// string constr = "Provider=Microsoft.Jet.OLEDB.4.0; Data Source=" + s + ";";
// OleDbConnection con = new OleDbConnection();
// con.ConnectionString = (constr);
// con.Open();
// OleDbCommand cmd = new OleDbCommand();
// cmd.CommandText = "SELECT m_id as i,m_name as mn,u_price as up,doe as d, record as r FROM medicine_w";
// cmd.Connection = con;
// OleDbDataReader dr;
// dr = cmd.ExecuteReader();
// if (dr.HasRows)
// {
//     dr.Read();
//     textBox3.Text = dr["i"].ToString();
//     textBox4.Text = dr["mn"].ToString();
//     textBox5.Text = dr["up"].ToString();
//     textBox6.Text = dr["r"].ToString();
//     label11.Text = dr["d"].ToString();
// }
// if (radioButton2.Checked == true)
// [
string n = (comboBox6.SelectedValue.ToString());
string s = "E:\medicine_inventory\medicine_inventory.mdb";

```

Figure 7.Simply removing Comment lines

### C. Long Parameter List Bad Smell Detection and Refactoring

- Analyze the source code
- Apply metric rules to detect long parameter list bad smell
  - Number of Source line of code (SLOC) > 50
- If above rule is true, long parameter list bad smell is detected.
- Apply Remove parameter refactoring technique that is suitable to remove long parameter list bad smell and extract new method for code reusability.

After uploading classes, select any one class to view class information, class all methods and check long parameter list in selected class. In Fig 8. two long parameter list bad smells are present in selected customerdetails class according to the code metric rule. Total long parameter list bad smells in all uploaded classes are 2.

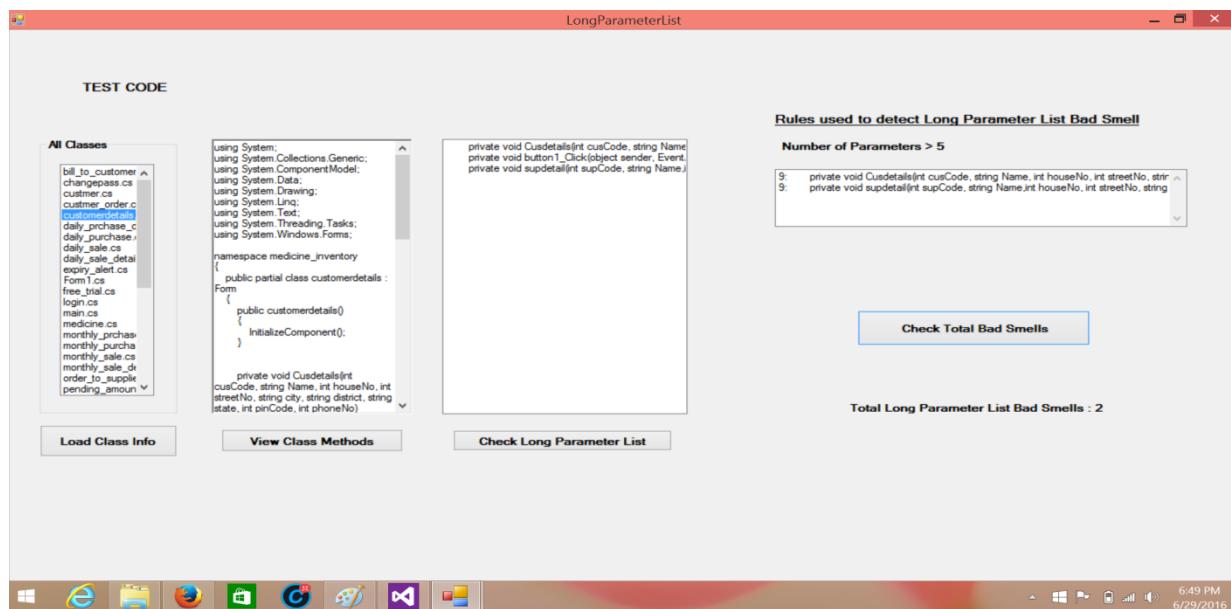


Figure 8. Long parameter list Bad Smell detection

Selected customerdetails class that has two long parameter list bad smells requires the refactoring. Right click on selected part of code, choose the remove parameter refactoring technique to refactor the code is shown in Fig 9.

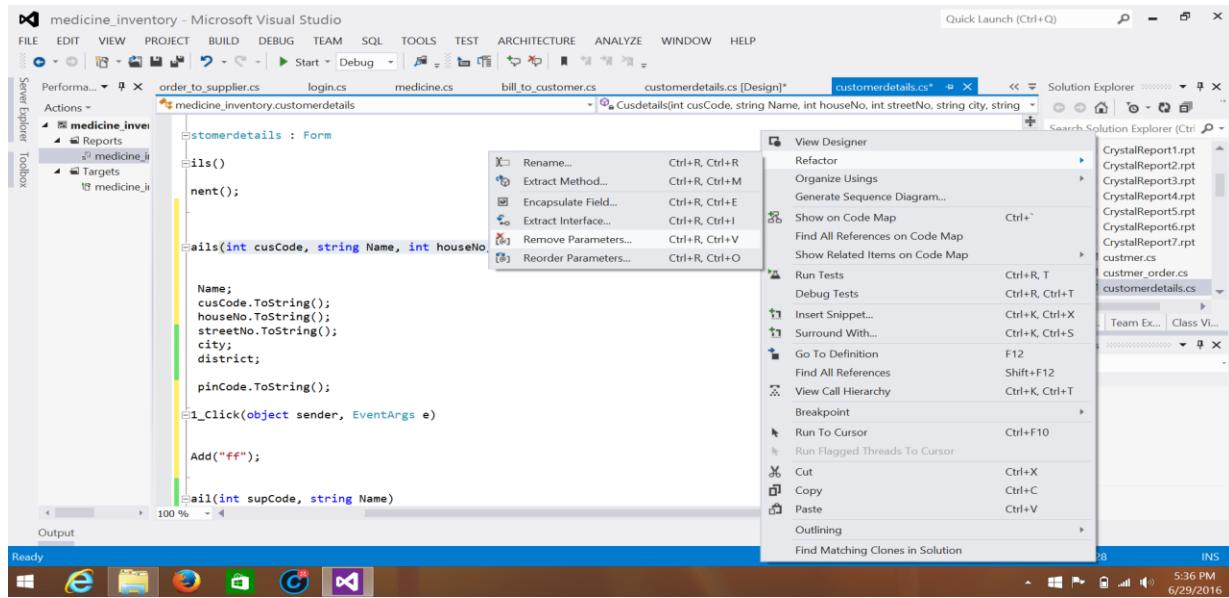


Figure 9. Select remove parameter refactoring technique

Remove parameters from method's parameter list are shows in Fig 10.

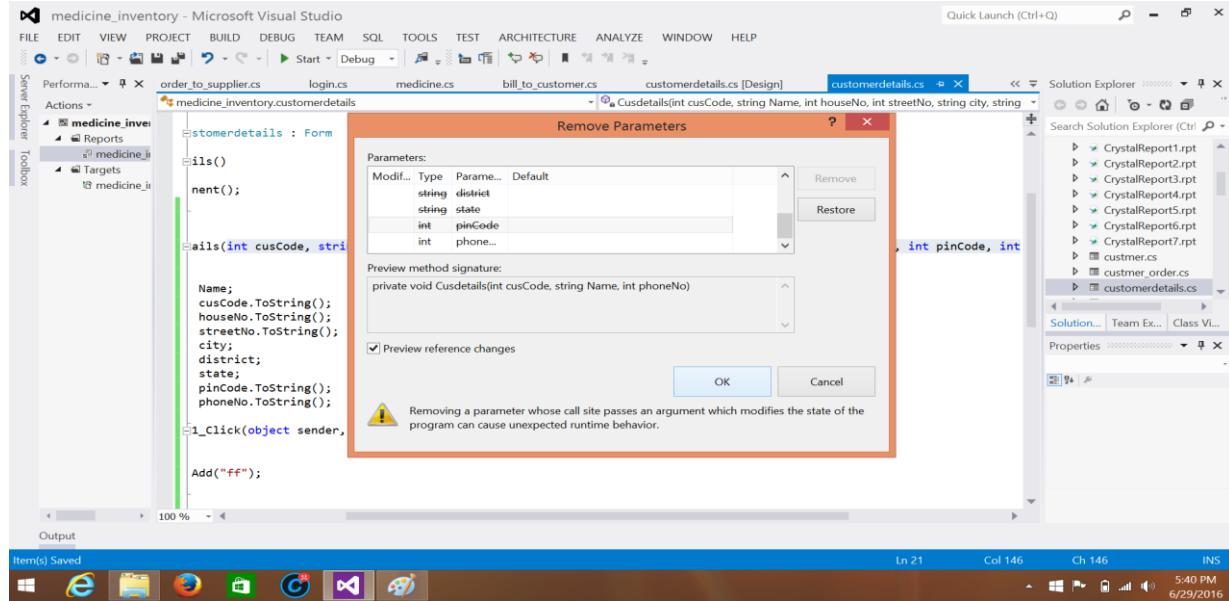


Figure 10. Remove parameters of selected method

Choose the extract method refactoring technique for creating new method from selecting the existing code of method. Parameters that are removed from list are declared in new created method for easy to call in another functionality of method. Gives the name of new created method is shown in Fig 11.

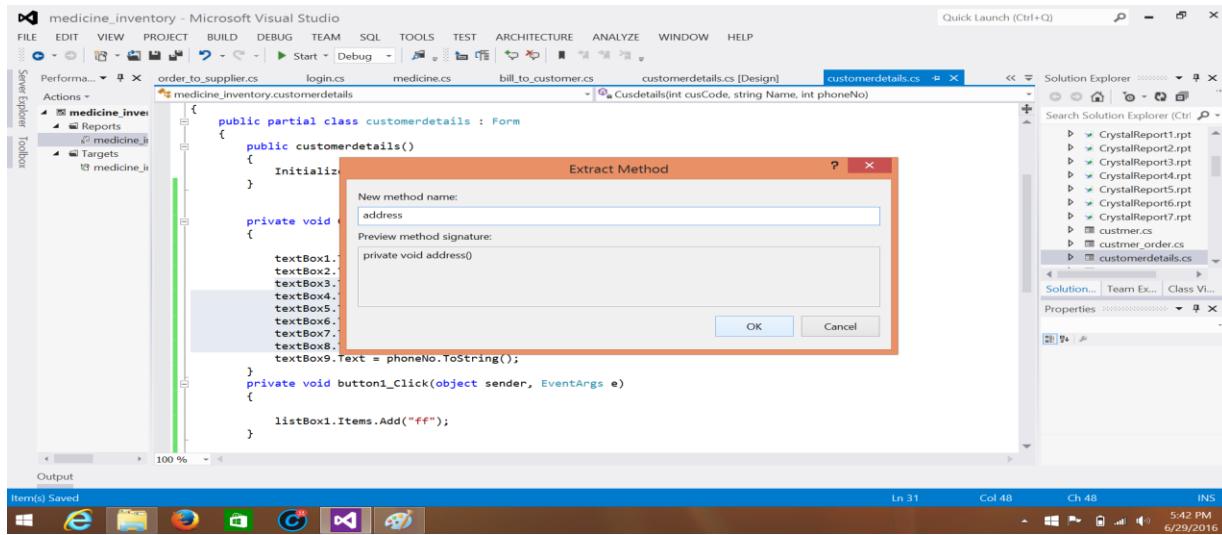


Figure 11. Type the name of new created method

Source code after refactoring with address method that provides the code reusability and easier to understand is shown in Fig 12.

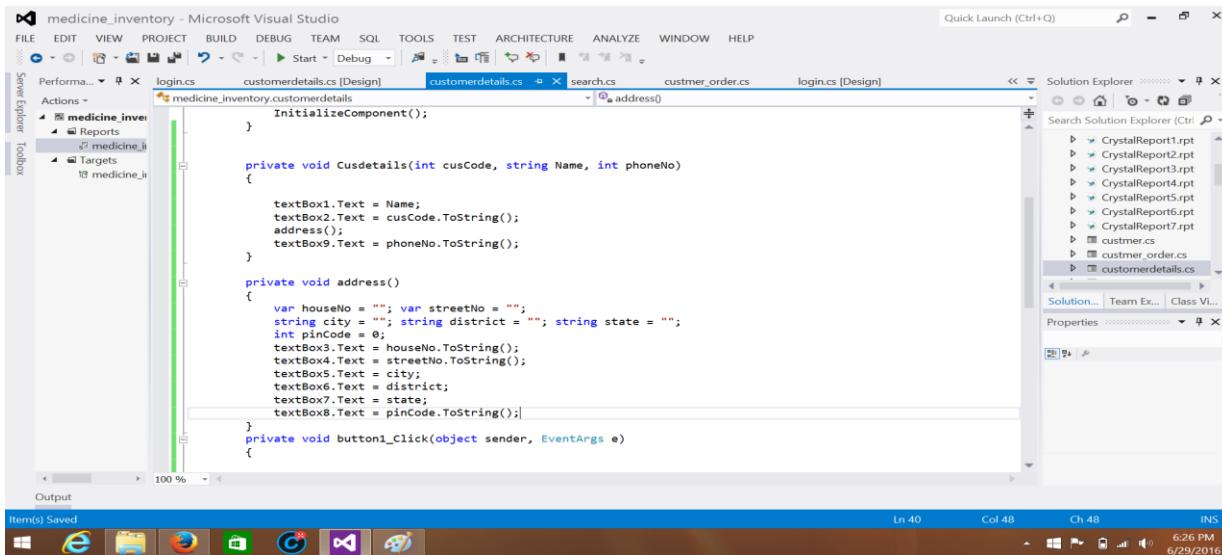


Figure 12. Source code after refactoring

## 5. RESULTS AND DISCUSSIONS

Various software metrics are used to detect bad smells to refactor the source code. Different types of bad smells are detected using visual studio ultimate tool 2012 with framework 4.5 that supports the .NET language. The .NET has many inbuilt languages like c++, VB, C# etc. In this work C# language is used to detect bad smells from the source code of program. The GUI window based forms are created to detect three types of bad smells such as long method, comment lines in method and long parameter list using software metrics. It only supports the .cs extension forms or classes from window applications and web applications.

Fig 2. shows the long method bad smell detection process using two types of software code metrics like Source lines of code (SLOC) and Cyclomatic complexity (CC). After detecting long method bad smell, Fig 3. Shows the

extract method refactoring technique is applied to remove the long method bad smell from source code and its result provides the more readable and less duplicate code. It helps in increasing the modularity and reusability of code. Extract method refactoring technique is easier and best way to refactor the source code that is used by software developers to achieve good maintainability of software code.

Fig 6. shows the comment lines bad smell detection process using two types of software code metrics like Average lines of code and Comment lines. After detecting comment lines bad smell, Fig 7. shows the code that is refactored by simply removing the comment lines and its result reduce the complexity from source code. To remove comment lines bad smell, code is manually refactored by programmers.

Fig 8. shows the long parameter list bad smell detection process using software code metric like Number of parameters (NP). After detecting long parameter list bad smell, Fig 9. Shows the remove parameter refactoring technique is applied to remove the long parameter list bad smell and its result provides the more consistency of code. Extract method refactoring technique is also applied to extract new method for more usability of same code structure and easier to understand a code.

### 5.1. Quantitative Analysis

Total 16 long methods, 38 comment lines and 2 long parameter list bad smells are detected in project medicine inventory (c#) code using different types of software metrics are shown in Table 1.

TABLE 1. RESULT OF TOTAL DETECTING BAD SMELLS

Object-oriented metrics	Bad smell name	No. of bad smells	Refactoring method to remove bad smell
Source lines of code and Cyclomatic Complexity	Long Method	16	Extract method
Average lines of code and Comment line	Comment Lines	38	Simply removed by selecting the bad smell part of code
Number of parameters	Long Parameter List	2	Combination of remove parameter and extract method

## 6. CONCLUSION AND FUTURE SCOPE

Day by day demand of refactoring for software artifacts such as models, UML, database and source code structure are increasing to obtain good maintenance of software. Refactoring is an essential part of development life cycle of software for producing the better software design and code structure. The tools for refactoring are faster used to diminish the occurrence of bugs from source code. In this work software metrics are used to detect bad smells. After detecting bad smells, Refactoring techniques are applied to source code according to behavior of bad smells. Simulation of refactoring techniques and detection of bad smells like long method, comment lines and long parameter list are done using visual studio ultimate tool 2012 for c# source code. From simulation it has been observed that extract method refactoring technique is easier and best way to remove long method bad smell for reducing complexity and duplicate code from source code.

For better software design or code structure refactoring is necessary to fix bug problems and refactor the code. The main aim of refactoring is to maintain code structure in future. The future work for refactoring is calculates the dynamic metric values to detect bad smells and apply refactoring techniques according to bad smell behavior. Compare the performance of refactoring tools for refactoring techniques.

#### REFERENCES

- [1] T. Mens and T. Tourwe “*A Survey of Software Refactoring*”, IEEE Transactions on Software Engineering, vol. 20, no. 150, pp. 1-12, 2004.
- [2] M. Lakshmanan and S. Manikandan “*Multi-Step Automated Refactoring For Code Smell*”, International Journal of Research in Engineering and Technology, vol. 03, no. 03, pp. 278-282, 2014.
- [3] Piyush Chandi “*Work Paper : Code Optimization using Refactoring*”, International Journal of Scientific & Engineering Research, ISSN 2229-5518, vol. 4, no. 8, pp. 414-421, 2013 .
- [4] M. Fowler, K. Beck, “*Refactoring: Improving the Design of Existing Code*”, Addison Wesley, 2002.
- [5] Phongphan Danphitsanuphan and Thanitta Suwantada “*Code Smell Detecting Tool and Code Smell-Structure Bug Relationship*” IEEE, Conference Paper, DOI:10.1109/SCET, 2012.
- [6] B Ramalkshmi1and D. Gayathri Devi, “*An Efficient Sdmpc Metric Based Approach For Refactoring Software Code*”, International Journal Of Engineering And Computer Science, vol. 4, no. 5, pp. 11733-11742, 2015.
- [7] J. Fields, S. Harvie, M. Fowler, K. Beck; “*Refactoring in Ruby*”, Addison Wesley, 2009.
- [8] Ganesh B. Regulwar and Raju M. Tugnayat “*Bad Smelling Concept in Software Refactoring*”, vol.45, no.012, pp. 56-61, 2012.
- [9] Karnam Sreenu 1and D. B. Jagannadha Rao, “*Performance - Detection of Bad Smells In Code for Refactoring Methods*”, International Journal of Modern Engineering Research (IJMER), vol. 2, no. 5, pp. 3727-3729, 2012.
- [10] Jiang Dexun, Ma Peijun “*Detection and Refactoring of Bad Smell Caused by Large Scale*”, International Journal of Software Engineering & Applications (IJSEA), vol.4, no.5, pp. 1-13, 2013.
- [11] Anshu Rani, Harpreet Kaur “*Detection of Bad Smells in Source Code According to their Object Oriented Metrics*”, International Journal for Technological Research in Engineering, vol. 1, no. 10, pp. 1211-1215, 2014.
- [12] Sandeep Kaur and Harpreet Kaur, “*Identification and Refactoring of Bad Smells to Improve Code Quality*”, International Journal of Scientific Engineering and Research (IJSER), vol.3, no. 8, pp. 99-102, 2014.
- [13] Manik Sharma and Dr. Gurdev Singh “*Analysis of Static and Dynamic Metrics for Productivity and Time Complexity*” International Journal of Computer Applications vol. 30, no.1, pp. 7-13, 2011.
- [14] Manik Sharma1, Gurdev Singh “*A Comparative Study of Static Object Oriented Metrics*” International Journal of Advancements in Technology, vol. 3, no. 1, pp. 25-34, 2012.
- [15] Ankush Vesra and Rahul “*A Study of Various Static and Dynamic Metrics for Open Source Software*” International Journal of Computer Applications vol. 122, no.10, pp. 17-20, 2015.

# Cost Based Energy Efficient Routing Protocol for Wireless Body Area Networks

**Ramneet Kaur, Er.Supreet Kaur**

Student Department of Computer Engineering, Punjabi University, Patiala, India  
ramneetgill65@gmail.com

Assistant Professor, Department of Computer Engineering, Punjabi University, Patiala, India  
supreetgill13@gmail.com

**Abstract:-**In this paper, the cost based energy efficient routing protocol is proposed for the Wireless Body Area Networks (WBANs). The major aim is to improve the lifetime network of the sink node. We propose a high throughput, path loss, residual energy, data send to sink and data received at sink routing protocol for (WBANs).The Cost function defined in this approach which is responsible to decided which node forward the data to the sink node. Cost function selects the node which has minimum threshold energy node which sends the data. Simulation outcomes show that proposed protocol raises the network stability amount and keep alive for longer amount.

**Keywords:** WBANs (Wireless Body Area Networks), Threshold, Cost Function, Path Loss, No of Dead nodes.

## I. INTRODUCTION

Wireless Sensor Networks is used in many applications like dweller monitoring [1] cultivation, enforcement monitoring, health monitoring, military fight, smart homes and weather checking etc. In wireless sensor networking the number of related sensor are connected with each other and performing the same functions. These sensors are used for checking the environmental factors. Sensor is a device which changes the physical quantity into a signal, which can be read by equipment. The sensor working is classified into four modules:-

- A. *Computing module:* - The sensor can be control by computing module and execution of communication protocols.
- B. *Communication module:* - It is responsible for radio communication between outside world and neighboring nodes.
- C. *Power supply module:* - It is responsible to give the supply power of the nodes.
- D. *Sensing module:*-It consists of group of sensors and links node to outside the world.

Now WBAN is the new sub-field of WSN .Wireless Body Area Network is mainly used in the health monitoring like check the fitness of sportsman ,medicine etc. In wireless body area network the nodes are of patient. These nodes are placed on the body of patient in star and multi-hop topology. The nodes monitor the different parameters of the patient like temperature, glucose, heartbeat, blood pressure etc. On the body there are eight nodes. One node are used mostly on body sink node which sends all the data of patient to the medical server and other node collect

the data from the patient body and send all the data to the sink node .These nodes required to use minimum energy to send data from nodes to sink node. In WBAN are used to solve problem is that to batteries need to be charged again and again. Much energy efficient routing protocols is used solves the problem of recharging batteries. The energy efficient routing protocols are proposed in wireless body area network [2], [3], [4].

The sink node is placed at waist of patients and the nodes for glucose and EGC are placed near the sink .The sensor nodes on the patient body at placed fixedly .Other nodes transmit the data to sink node.

Use of WBAN technology to monitor health parameters reduces cost of patient in hospital. With the help of WBAN technology, patients are controlled at home for longer period. Most importantly, while terminals in wireless networks such as cellular or Wi-Fi usually have abundant memory and enough power, sensors in WBANs are limited in both size and energy. As sensors are attached to the patient body, WBANs are under severe energy and power constraints. Long system lifetimes and low emitting power are desired. Medium Access Control (MAC) protocol design is therefore one of the most important issues in WBAN development.

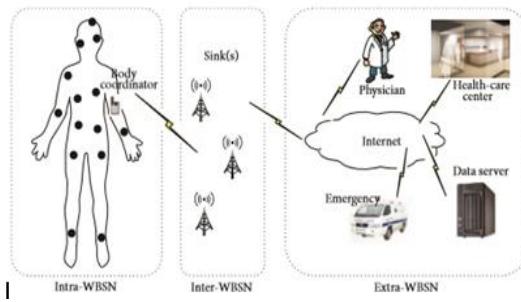


Figure 1. Architecture of wireless body area network [6]

## II. LITERATURE SURVEY

**W. Joseph et al** .had proposed scheme of Energy Efficient Topologies for Wireless On-Body Channel [1]. The author terminate that the propagation analysis is used to examine to determine the perfect network topology in terms of power efficiency for look at subjects. It had shown the single-hop communication in terms of energy consuming and multi-hop communication choice the path. A collective approach can update energy consumption; this distributes the transmission load on the network. This algorithms can choose which node sends data and that allow a node to communicate whether it is capable of cooperating or not.

**Jocelyn Elias et al** .had proposed Energy-aware Topology scheme for WBAN [2]. In this paper author describes topology method issue for Wireless Body Area Networks, propose the active model based on mathematical programming (1) the optimal number and opening of relay nodes, (2) the best appointment of sensors to relays (3) the best traffic routing, taking proper account of both the total network cost and energy consuming. This model can used to decrease the total energy consuming and the network cost of sensors.

**Elisabeth Reusens et al.** had proposed pretending of On-Body Communication Channel and Energy Efficient Topology scheme for WBAN [3]. In this paper physical propagation is used to examine the application on protocol level to complete the most optimal network topology in expressions of energy efficiency. This Paper shows the relay equipment, or a more cooperative way, can search energy consuming. this present the transmission load on the network. This review must be seen the first step approaching a highly energy efficient WBAN. Based on the results new communication protocols can be refined or actual ones can be adapted.

**J. I. Bangash et al.** had proposed Reliability Aware Routing for Intra-Wireless Body Sensor Networks [4]. The author exposed averagel power consuming and average temperature of RAR which related to TMQoS because RAR choose most wanted next hop locates on path loss and reliability.

**G. R. Tsouriet et al.** had created on Increasing Network Lifetime in BAN Using Global Routing with Energy Consuming Balancing [5]. This paper proposed global routing which allows WBANs to complete efficiently for longer periods of time, but previously recharging of batteries is needed. Due to the balancing of energy use in the network, equipment would deplete their energy sources at almost the same time. This is good because all equipment can be recharged simultaneous. Since NL is heightened as well, depletion of batteries is lower frequent, decreasing carrying requirements even further.

### III. COST BASED ENERGY EFFICIENT ROUTING PROTOCOL

#### A. Steps of network Working

The working of network is discussed as follow:-

*a. Starting Phase*:-Base node transmission its place through data packet, and then Sensor node store the place of base node. Each sensor transmits data packet to base which contains node ID, its residual energy and place. Base node transmits data to all sensors nodes.

*b. Selection of Sending Node*:-Smallest amount cost function value used to select most correct data sending which contributes toward network high transmission. A node with high residual energy and less distance to base has smallest amount cost function.

$$\text{Cost Function (i)} = \text{distance (i) /Residual Energy (i)}$$

*c. TDMA Scheduling*:-Sending node assigns TDMA schedule to its children node, TDMA establish minimum collision. Children nodes transmit their data in given time slot. The energy of sensor nodes saves by TDMA scheduling. TDMA is used a data aggregation. Duplicity can be removed by data aggregation in the network, data high throughput, more residual energy.

$$\text{DCT} : -X_k = \frac{1}{2}(x_o + (-1)^k x_{N-1}) + \sum_{n=1}^{N-2} x_n \cos \left[ \frac{\pi}{N-1} nk \right] \quad k = 0, \dots, N-1.$$

#### IV. METHODOLOGY

WBSN with 8 nodes will be implemented to carry out extensive simulations. The field's dimension will be 0.6 x 1.6 meters. Then the transmission power, receiving power and electronics power used as per given radio model.

$$ETx(k, d) = ETx - elec(k) + ETx - amp(k, d) \dots 1$$

$$ETx(k, d) = ETx - elec * k + Eamp * k * d^2 \dots 2$$

$$ERx(k) = ERx - elec(k) ERx(k) = ERx - elec * k \dots 3$$

where  $ET_x$  is the energy consumed in transmission,  $ER_x$  is the energy consumed by receiver,  $ET_x$ -elec and  $ER_x$ -elec are the energies required to run the electronic circuit of transmitter and receiver, respectively.  $E_{amp}$  is the energy required for amplifier circuit, while  $k$  is the packet size.

The various parameters are used in our research which is shown in tabular form as below:-

PARAMETERS	VALUES
No of nodes	8
Field dimensions	0.6 x 1.6 meters (human body)
No of rounds	8000
Initial energy ( $E_{int}$ )	0.5 volt
Data aggregation DCT	$5*0.00000001$

TABLE 4.1:- Various Parameters used in Implementation

#### V. SIMULATION RESULTS AND ANALYSIS

To classify the proposed protocol, we have operated an expanded set of observations using MATLAB 2013. It improves the performance of SIMPLE protocol and comparability with the existing protocol M-ATTEMPT.

*A. Number of dead nodes*:-Figure 2 show the dead node in the Wireless body area networks. In the old Attempt protocol 3 nodes properly dead at 3000 rounds. Now in the cost based routing protocol 3 nodes are dead at 7000 rounds. In the new cost based routing protocols the networks works for long time and performance of the network is high, rather than the attempt protocol. Through the CBRP data properly send on the server and data integration is also decrease. Te stability of the network is high in CBRP.

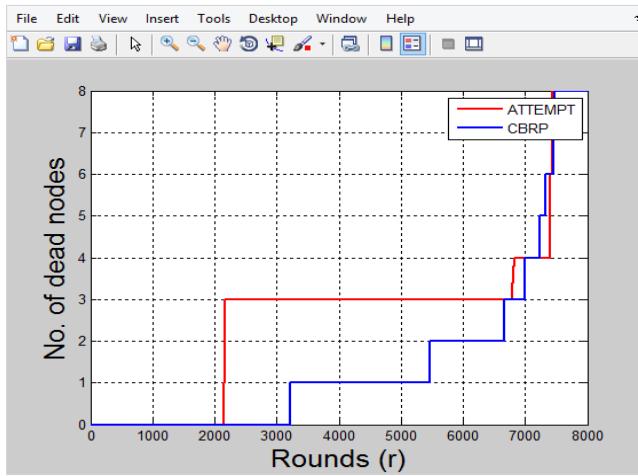


Figure .2: Number of dead nodes

*B. Data sent to sink:* - In figure 3 the data send to the sink node has shown in the new cost based routing protocol. In the attempt protocol not shows the data send to the sink node in the WBAN. In the attempt protocol  $0.5 \times 10^4$  data send at 1000 rounds ,In the cost based protocol first the data is send to sink node is low and when the no of rounds is increase, then sending of data is also increases . At 8000 rounds the data is send to sink node is  $3.8 \times 10^4$  in CBRP .In attempt protocol  $3 \times 10^4$  data send to sink at 8000 .In the CBRP has minimum packet drop. The stability of the network is higher than the attempt protocol.

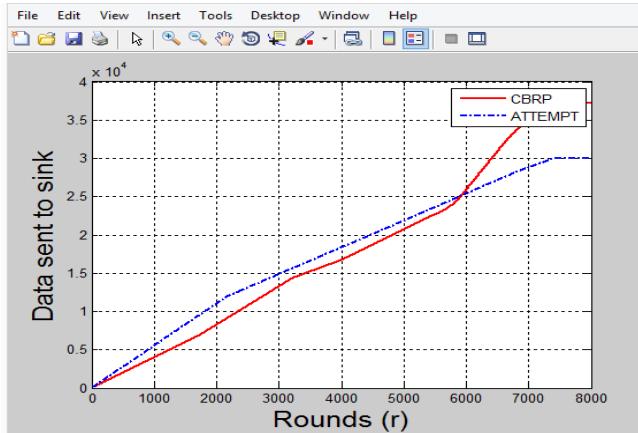


Figure.3: Data sent to sink

*C.Path Dropped:-* In Figure.4 has shown the path loss in the WBAN. The proposed new CBRP results in less path loss and use less distance method for sending .The proposed CBRP is better then ATTEMPT protocol. It shows its characteristics diatance and ferquency. It is measured from its distance to sink node with constant frequency 2.4 ghz. The CBRP protocol reduces the path loss shown in the figure 4 . Intially in the attempt protocol at 3000 round 275 data dropped decrease due to no of dead nodes in the wireless body area networks. In the CBRP the data drop is minimum, rather then the attempt protocols. The stability of the network is high. If the data drop in the network is less the lifetime of the sink node is more and the data is properly sent to the server.

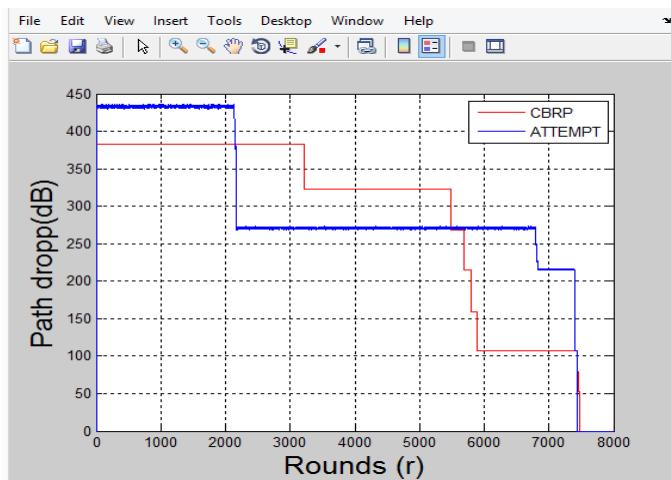


Figure.4: Path dropped

*D. Delay:*-The delay of the CBRP shown in Figure.5. The delay shown by cost based routing protocol is minimum than the attempt protocol. It increase the output of the wireless body area network.If the delay of network is less ,then the work for long time.The stability of the network is high.

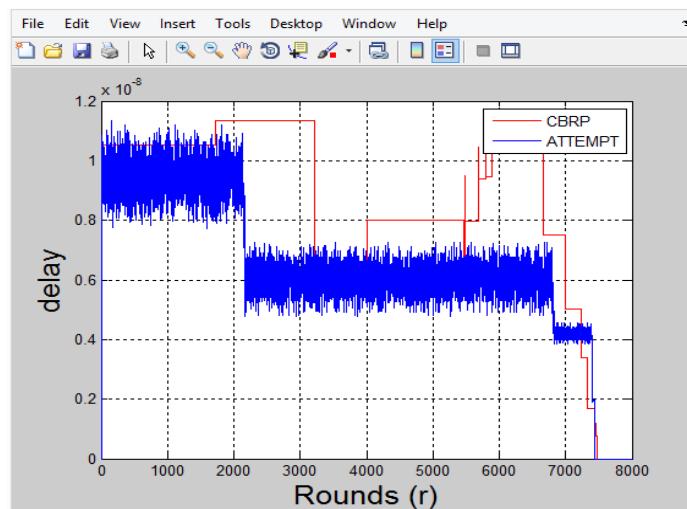


Figure.5: Delay

*E. Residual Energy:*In Figure 6, the residual energy exposed by the CBRP in shown in figure 6. In the attempt protocol the residual energy is less at 1000 rounds then the CBRP .If the residual energy is greater the stability of network is high and it improves the throughput of the wireless body area network. Due to the residual energy the data transmission is more and packet drop or path loss is less. The residual enegy of network is increase and the lifetime of the sink node is also increase and it works for long time.

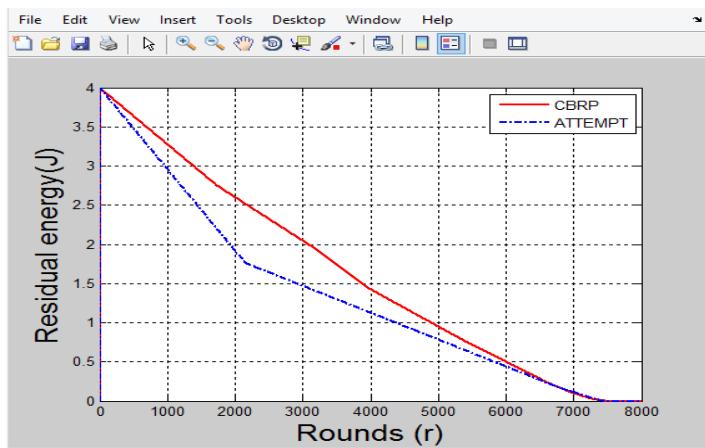


Figure.6: Residual energy

## VI. CONCLUSION

In our proposed approach a new cost function is proposed, which is depending on vulnerability and energy factor. The energy factor computes the critical paths which are not capable of sending requested data. The critical paths are dropped as some request with lower load can be fulfilled using that path. This approach saves the energy and makes transmission successful too. In vulnerability factor the distance, RSSI considered along with incorporating the probability of movement of patient which is neglected in previous approach. Received signal strength indication (RSSI) is basically a measure of power of node. Higher RSSI signify more transmission range. The probability of movement is randomly considered as performance of patient is unpredictable. So our proposed cost function is based on both these factors also considering the other odd factors, to address the performance of routing in best manner. Our new cost based routing protocol improve the performance of the wireless body area network with the various characteristics like residual energy ,path loss, data sent to sink , no of dead nodes and delay etc. It improves the lifetime of the sink node due to these characteristics.

## REFERENCES

- [1] W. Joseph, B. Braem, E. Reusens, B. Latre, L. Martens, I. Moerman, and C. Blondia, "Design of Energy Efficient Topologies for Wireless On-Body Channel," *Wirel. Conf. 2011-Sustainable Wirel. Technol. (European Wireless), 11th Eur.*, vol. 5, no. 4, pp. 1–7, 2011.
- [2] J. Elias and A. Mehaoua, "Energy-aware topology design for wireless body area networks," in *2012 IEEE International Conference on Communications (ICC)*, 2012, pp. 3409–3410.
- [3] E. Reusens, W. Joseph, B. Latré, B. Braem, G. Vermeeren, E. Tanghe, L. Martens, I. Moerman, and C. Blondia, "Characterization of on-body communication channel and energy efficient topology design for wireless body area networks.,," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 2, pp. 933–945, 2009.
- [4] J. I. Bangash, A. H. Abdullah, M. A. Razzaque, and A. W. Khan, "Reliability Aware Routing for Intra-Wireless Body Sensor Networks," *Int. J. Distrib. Sensors Networks*, vol. 2014, no. 1, 2014.

- [5] G. R. Tsouri, A. Prieto, and N. Argade, “On Increasing Network Lifetime in Body Area Networks Using Global Routing with Energy Consumption Balancing,” *Sensors*, vol. 12, no. 9, pp. 13088–13108, 2012.
- [6] G. Lo, S. Member, S. Gonz, and V. C. M. Leung, “Wireless Body Area Network Node Localization using Small- Scale Spatial Information,” *IEEE J. Biomed.*, vol. 01, no. 00, 2012.
- [7] D. Zhang, G. Li, K. Zheng, X. Ming, and Z. H. Pan, “An energy-balanced routing method based on forward-aware factor for wireless sensor networks,” *IEEE Trans. Ind. Informatics*, vol. 10, no. 1, pp. 766–773, 2014.
- [8] G. Subramanian, “Efficient and Secure Routing Protocol for Wireless Sensor Networks using Mine detection,” *IEEE Trans. Netw.*, vol. 10, no. 7, pp. 141–145, 2014.
- [9] J. Choi, “Secure Multipath Routing in Wireless Multihop Networks based on Erasure Channel Modeling,” in *IEEE Wireless Advanced*, 2012, pp. 6–10.
- [10] C. S. Raghavendra, S. Lindsey, and S. Lindsey, “PEGASIS : Power-Efficient Gathering in Sensor Information Systems Stephanie Lindsey,” in *Aerospace Conference Proceedings*, 2002, p. 7.
- [11] M. Quwaider and S. Biswas, “Delay Tolerant Routing Protocol Modeling for Low Power Wearable Wireless Sensor Networks,” *Netw. Protoc. Algorithms*, vol. 4, no. 3, pp. 15–34, 2012.
- [12] Q.Nadeem,N.Javaid,S.N.Mohammad,M.Y.Khan,S.Sarfraz,M.Gull. (2013). SIMPLE:Stable Increased-throughput Multi-hop Protocol for Link Efficiency in Wireless Body Area Networks. *Eighth International Conference on Broadcast,Wireless Computing Communication and Applications* (pp. 221-226). IEEE.

# A REVIEW ON IMAGE ENHANCEMENT

Kiranjeet Kaur<sup>1</sup>, Ashok Kumar Bathla<sup>2</sup>

<sup>1,2</sup>Department of Computer Engineering, YCOE, Punjabi University, Patiala, Punjab, India

[jeetkiran020@gmail.com](mailto:jeetkiran020@gmail.com)<sup>1</sup>, [ashokashok81@gmail.com](mailto:ashokashok81@gmail.com)<sup>2</sup>

**Abstract:** Over the many decades, there had been notable capability enhancements in Digital cameras together with resolution and sensitivity. Despite these enhancements, however, trendy digital cameras area unit still restricted in capturing high dynamic vary pictures in low-light conditions. Furthermore, dust storms are additionally climate occasions that are regularly experienced when driving in a few districts. A novel and effective haze removal approach to cure these issues created by confined light sources and color shifts, which subsequently accomplishes better restoration results for single hazy images. Poor visibility degrades quality and performance of computer vision algorithms for smart transportation frameworks, for example, traveling vehicle data recorders and traffic surveillance systems, activity observation frameworks which must operate under a wide range of weather conditions. Movement detection and Darkness are also problems in the captured images. The goal of this work is to enhance the road scene images using different filters and enhancement techniques. The distinctive sorts of parameters are figured that are PSNR, MD, MSE and Processing Speeds.

**Keywords:** PSNR, MD images, HE, BBHE, DPC, RSWHE.

## I.INTRODUCTION

With the quick advancement of the computerized photography innovation, the securing of computerized video is currently a simple assignment. Recording scenes of our everyday life as far as video clasps has turned into a famous way of life. In any case, it is even nontrivial for capable picture takers to catch fantastic recordings in low light-level photographic environment. In this manner, worldwide or nearby underexposed video clasps are definitely made alongside the day by day photography of learner picture takers. Because of the low perceptibility these underexposed recordings as a rule neglect to exhibit outwardly satisfying searching. In the field of open wellbeing, the broadly utilized video observation frameworks likewise create underexposed evening recordings, which truly debilitate the framework in auto collisions examination and wrongdoing crime scene investigation. To enhance the visual appearance of the underexposed recordings, video improvement method develops. Presently, video upgrade is still a dynamic exploration theme in the PC representation and PC vision groups. Underexposed video upgrade is a testing issue. To date, there is no agreement standard for assessing whether an upgraded video is outwardly satisfying. In light of human visual discernment hypothesis, we tailor three fundamental destinations for our video upgrade method: the first concealed points of interest ought to be anything but difficult to recognizing the improved video; the improved video ought to abstain from presenting visual ancient rarities, for example, gleaming and uneven introduction, that initially don't exist in the source underexposed video; the improved video ought to be transiently predictable. We likewise show that our reasonably basic destinations for underexposed video improvement can supplement other important strategies, e.g., video de-hazing. Numerous methodologies have been proposed for upgrading underexposed recordings. These methodologies can be basically arranged into two classifications: setting based methodologies and connection free methodologies. Setting based approaches improve low-quality utilizing so as to even time recordings brilliant daytime recordings.

## II. IMAGE ENHANCEMENT TECHNIQUES

### i) Histogram Feat

Histogram feat may be a technique for adjusting image intensities to reinforce distinction.

Let  $f$  be a given image delineated as a matrix by  $M \times N$  matrix of whole number picture element intensities starting from zero to  $L - 1$ .  $L$  is the variety of potential intensity values, often 256. Let  $p$  denote the normalized bar chart of  $f$  with a bin for every potential intensity. So

$$p_n = \text{variety of pixels with intensity } n \quad n = \text{zero, } 1, \dots, L - 1.$$

Total variety of pixels

### ii) Brightness Conserving Bi-histogram Leveling (BBHE)

This methodology divides the image bar chart into 2 elements. during this methodology, the separation intensity is bestowed by the input mean brightness worth, that is the common intensity of all pixels that construct the input image. Once this separation method, these 2 histograms area unit severally equal. By doing this, the mean brightness of the resultant image can lie between the input mean and the center grey level. The bar chart with vary from zero to  $L-1$  is split into 2 elements, with separating intensity noise. This separation produces 2 histograms. the primary bar chart has the vary of zero to noise , whereas the second bar chart has the vary of  $X_{T+1}$  to  $L-1$ .

### iii) Recursive Mean-separate HE Technique (RMSHE)

Recall that the extensions of the HE technique represented up to now during this section were characterized by moldering the initial image into 2 new sub-images. However, associate degree extended version of the BBHE technique named algorithmic suggests that separate HE (RMSHE), proposes the subsequent. Rather than moldering the image just one occasion, the RMSHE technique proposes to perform image decomposition recursively. After, every one of those sub-images  $I$  is severally increased sing the CHE technique. Note that once  $r = \text{zero}$  (no sub-images are generated) and  $r = \text{one}$ , the RMSHE technique is admire the CHE and BBHE ways, severally. They mathematically showed that the brightness of the output image is higher preserved as  $r$  will increase. Note that, computationally speaking, this technique presents a drawback: the quantity of rotten sub-histograms could be a power of 2.

## III. RELATED WORK

Zhang et al. (2015) Underexposed video improvement aims at revealing hidden details that area unit barely noticeable in LDR video frames with noise. Previous work generally depends on one heuristic tone mapping curve to expand the dynamic vary, that inevitably results in uneven exposure and visual artifacts. For Associate in Nursing input underexposed video, we have a tendency to 1st remap every video frame employing a series of tentative tone mapping curves to get Associate in Nursing multi-exposure image sequence that contains totally different exposed versions of the first video frame Besides, we have a tendency to demonstrate applications of our approach to a group of issues together with video de-hazing, video de-noising and HDR video reconstruction. [7] Ms. Pallavi et al. (2015) Over the many decades, there had been notable capability enhancements in Digital cameras together with resolution and sensitivity. Despite these enhancements, however, trendy digital cameras area unit still restricted in capturing high dynamic vary pictures in low-light conditions. These cameras usually believe automatic exposure management to capture image of high dynamic vary, however the longer exposure time usually results motion blur. several approaches area unit developed for enhancing low light-weight video; but most of them think about video from moderately dark conditions. [8] P. Bennett et al. (2014) Enhance underexposed, low dynamic vary videos by adaptively and severally variable the exposure at every photoreceptor in a very post-process. This virtual

exposure may be a dynamic operate of each the special neighborhood and temporal history at every picture element. Temporal integration allows North American nation to expand the image's dynamic vary whereas at the same time reducing noise. Our non-linear exposure variation and de-noising filters swimmingly transition from temporal to special for moving scene components. [9] Rao et al. (2012) Video improvement is one among the foremost vital and troublesome parts in video analysis. The aim of video improvement is to enhance the visual look of the video, or to produce a "better" transform illustration for future machine-controlled video process, like analysis, detection, segmentation, recognition, police investigation, traffic, criminal justice systems. [10] Cheng et al. (2011) had planned a completely unique background subtraction approach so as to accurately notice moving objects. The strategy involves 3 vital planned modules: a block alarm module, a background modeling module, associated an object extraction module. The block alarm module expeditiously checks every block for the presence of either a moving object or background data. This can be accomplished by exploitation temporal differencing pixels of the designer distribution model. The background modeling module is utilized so as to get a high-quality adjectives background model employing a distinctive two-stage coaching procedure and a completely unique mechanism for recognizing changes in illumination. Menotti et al. (2007) Histogram exploit (HE) has well-tried to be a simple and effective image distinction improvement technique. However, it tends to alter the mean brightness of the image to the center level of the gray-level vary, that isn't fascinating within the case of images from consumer physical science merchandise. In the latter case, preserving the input brightness of the image is needed to avoid the generation of non-existing artifacts within the output image. To surmount this downside, Bi-HE ways for brightness protecting and distinction sweetening are projected. Although these strategies preserve the input brightness on the output image with a significant distinction improvement, they may manufacture pictures with do not look as natural because the input ones. In order to overcome this drawback, this work proposes a unique technique known as Multi-HE that consists of rotten the input image into many sub-images, and so applying the classical HE method to everybody. [16] Lee et al. (2007) A gradient domain tone mapping algorithm is planned to show high dynamic vary (HDR) video sequences in low dynamic range (LDR) devices in this work. The proposed formula obtains a pixel wise motion vector field and incorporates the motion information into the Poisson equation. Then, by attenuating large abstraction gradients, the proposed formula will yield a high-quality tone-mapped result without a flicker artifact. Simulation results show that the proposed formula provides a higher performance than the frame based method, which processes every frame severally. [17] Xuan Dong et al. (2010) we describe a novel and effective video sweetening rule for low lighting video. The algorithm works by initial inverting the input low-lighting video and then applying a picture de-haze algorithm on the inverted input. To facilitate faster computation and improve temporal consistency, correlations between temporally neighboring frames square measure used. Simulations mistreatment naive implementations of the algorithmic rule show smart sweetening results and 2x speed-up as compared with frame-wise improvement algorithms, with any enhancements in every quality and speed accomplishable. [18] Jinno et at. (2007) A two layer secret writing formula for high dynamic vary pictures is mentioned. In the First layer, a low dynamic range image is encoded by a conventional codec, and then the residual information that represents the difference between associate original and therefore the decoded pictures within the initial layer is encoded within the second layer, that realizes compatibility with standard image file formats. Our technique utilizes the approximation of associate inverse tone mapping operate that reduces the high dynamic vary to a displayable vary. Our algorithmic rule considerably improves a compression performance, compared to traditional ways. [19]

#### IV. METHODOLOGY

The dissertation is to improvement in Road Scenes Captured by Intelligent Transportation Systems. It is based mostly upon user interface (graphical user interface) in MATLAB. It is an effort to additional grasp the fundamentals of MATLAB and every of them consists of m-file and figure file. The video sweetening continues to be an energetic space of analysis by several consultants. There square measure still several issues of video sweetening, like false background downside, color shift downside etc. Video sweetening is one in every of the foremost vital and tough element of video security closed-circuit television. The increasing use of night operations desires extra details and integrated information from the improved image. However, inferiority video of most police work cameras isn't happy and tough to know as a result of they lack close scene context attributable to poor illumination. Thus one in every of key issues is image/frame fusion downside to confirm higher image reconstruction and color assignment. An outsized variety of techniques are projected to deal with this downside. During this we have a tendency to concentrate on the present techniques of video sweetening, which might be created higher in poor visibility light-weight condition. Desired outcome of the project is to reinforce video. The steps to get desired outcome is as shown in fig.

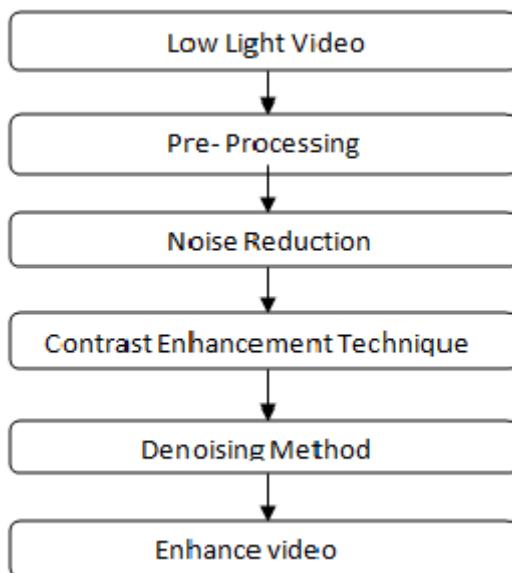


Figure 1. Work flow chart

## V. CONCLUSION

Different enhancement techniques for enhancing images are described in the paper. Image enhancement of general images improves the interpretability or perception in images. Diverse issues emerge when the captured foggy scene images contains restricted light sources or shading movement issues because of dust storm conditions, sunlight, fog or extensive variety of climate conditions. Motion Detection is also among one of the issue areas. Future extension can be to enhance the Road scene images using various methods. The distinctive sorts of parameters are computed that is PSNR, MSE and AMBE and to analyze the results being obtained.

## ACKNOWLEDGEMENT

I am Thankful to my respected guide Er.Ashok Kumar Bathla, Assistant Professor (Computer Engineering), Yadavindra College of Engineering, Talwandi Sabo for his invaluable and enthusiastic guidance, useful suggestions.

## REFERENCES

- [1] Shih-Chia Huang, Bo-Hao Chen, Yi-Jui Cheng, "An Efficient Visibility Enhancement Algorithm for Road Scenes Captured by Intelligent Transportation Systems" IEEE Transactions On Intelligent Transportation Systems, Vol. 15, No. 5, pp.2321-2332, October 2014.
- [2] Ivan et.al, "Multi-Label Classification of Traffic Scenes" Proceedings of the Croatian Computer Vision Workshop, pp. 9-14, September 16, 2014.
- [3] Ajay Raghavan, Robert Price, Juan Liu, "Detection of Scene Obstructions and Persistent View Changes in Transportation Camera Systems" 15th International IEEE Conference on Intelligent Transportation Systems Anchorage, Alaska, USA, pp. 957-962, September 2012.
- [4] Fan-Chieh Cheng, Shanq-Jang Ruan, "Accurate Motion Detection Using a Self-Adaptive Background Matching Framework" IEEE Transactions on Intelligent Transportation Systems, Vol. 13, No. 2, pp. 671-679, June 2012.
- [5] J.-E. Ha, W.-H. Lee, "Foreground objects detection using multiple difference images" Opt. Eng., Vol. 49, no. 4, pp. 047-201, April 2010.
- [6] Giri Nandan, "Image Resolution Enhancement Methods for Different Applications" International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, no.17, pp. 1733-1738, 2014.
- [7] Qing Zhang et.al., (2015), "Underexposed Video Enhancement via Perception-driven Progressive Fusion" IEEE journal of latex class files.
- [8] Ms. Pallavi H. Yawalkar et.al., (2015), "A Review on Low Light Video Enhancement Using Image Processing Technique" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 1.
- [9] Eric P. Bennett et.al. "Video Enhancement Using Per-Pixel Virtual Exposures" Microsoft world publication.
- [10] Yunbo Rao et.al., (2012), "A Survey of Video Enhancement Techniques" Journal of Information Hiding and Multimedia Signal Processing, ISSN 2073-4212 Ubiquitous International Volume 3, Number 1
- [11] C. Stauffer, W.E.L Grimson, "Adaptive background mixture models for real-time tracking" in Proc. IEEE Computing Vision And Pattern Recognition, 1999, Vol. 2, pp. 246–252, June 1999.
- [12] A. Doshi et.al, "Smoothing of optical flow using robustified diffusion kernels" Image Vis. Computing, Vol. 28, no. 12, pp. 1575–1589, Dec. 2010.
- [13] P. Bhat, C. L. Zitnick, M. Cohen, and B. Curless, (2010), "Gradientshop: A gradient-domain optimization framework for image and video filtering," ACM Transactions on Graphics (TOG), vol. 29, no. 2, p. 10.
- [14] Henrik Malm Magnus Oskarsson Eric Warrant, (2007), "Adaptive enhancement and noise reduction in very low light-level video" IEEE 978-1-4244-1631-8/07.
- [15] A. A. Wadud, M. Kabir, M. H. Dewan and M. C. Oksam,(2007), "A dynamic histogram equalization for image contrast enhancement," IEEE Trans. Consumer Electronic, volume 53, number 2, pp. 593-600.
- [16] D. Menotti, L. Najman, J. Facon and Arnaldo de A. Araujo, (2007), "Multi-histogram equalization methods for contrast enhancement and brightness preserving," IEEE Trans. Consumer Electronics, Volume 53, Number 3, pp. 1186-1194.
- [17] C. Lee, C. S. Kim, (2007), "Gradient domain tone mapping of high dynamic range videos," Proceedings of IEEE International Conference on Image Processing, no. 3, pp. 461-464.
- [18] X. Dong, Y. Pang and J.Wen,(2010), "Fast efficient algorithm for enhancement of low lighting video," Proceedings Of the 37th International Conference and Exhibition on Computer Graphics an Interactive Techniques.

- [19] T. Jinno, M. Okuda and N. Adami, (2007), "Acquisition and encoding of high dynamic range images using inverse tone mapping," Proceedings of IEEE International Conference on Image Processing, vol. 4, pp.181-184.

# A Review on Biometric Authentication using Adaptive Iris features

Rajdeep Kaur

Department of Computer Engineering  
YCOE, Talwandi Sabo, Punjab, India  
[Rajdeepbamrah61@gmail.com](mailto:Rajdeepbamrah61@gmail.com)

Er.Rajan Goyal

Assistant Professor Department of Computer  
engineering  
YCOE, Talwandi Sabo, Punjab, India  
[Er.rajangoyal@gmail.com](mailto:Er.rajangoyal@gmail.com)

**ABSTRACT**—There are several biometric identities associated with the human which prominently includes the fingerprint, palm-print, palm-vein, finger-vein, retina and Iris features. The human beings are identified by their biological identities for the attendance systems, authorization systems or other similar applications. The biometric systems have found their way into almost all of the organizations with the medium to large employee base. Many of the small organizations with the adequately higher number of employees are also incorporating the biometric systems. The biometric systems based upon the Iris features are being popular as the standalone or hybrid biometric or with other authentication entities. The Iris recognition requires the accurate localization of the Iris features from the image of eye collected for training or testing purposes. The Iris extraction requires two demarcation circles, where first circle demarcates the outer boundary and second circle demarcates the inner boundary by detecting the outer boundary of the pupil. Also, the angular shift mechanism can be incorporated to study the movement of the Iris in the given image for accurate localization of the region of interest containing the Iris feature. The proposed solution will utilize the probabilistic classification based upon the multi-class SVM to detect the Iris features with or without contact lenses. The proposed solution aims at improving the existing model for the robust performance.

**KEYWORDS**—IRIS recognition, moved feature localization, probabilistic classification, authorization systems.

## I. INTRODUCTION

### A. ANGULAR SHIFT DETECTION

In general, most up-to-date victorious object recognition algorithms involve a) identifying features of an object that are invariant to transformation, and b) seeking near-matches of these features in potential candidate pictures. Such searches also performed by reducing object pictures to a collection of interest points, using Lowe's Distinction of Gaussian (DoG) detector [6] or the Harris corner detector [3]. Native features square measure then calculated at these points with a range of strategies (several of that square measure compared in [8]), and correspondences between these feature sets are sought-after between all points calculated within the target image and a candidate search image. Finally, ways resembling the generalized Hough Transform or RANSAC are used to calculate the affine transformation between the target and a candidate. These techniques are acceptable and efficient for the corners and blobs detected by Harris and DoG strategies, and are applied to edge features furthermore, for detecting “wiry” objects [7]. We'd like to adopt associate approach that acknowledges that edge features don't possess a

clearly defined “interest point” representation; they are entities that are distributed widely throughout area further as scale. Therefore, we’ll represent edges with 2-dimensional entities throughout this paper. By doing therefore, we have a tendency to put together distinguish our technique from classic form metrics such as the Hausdorff distance that, even when orientation information is enclosed ([9]), describes points which will not be individually strong. Our new methodology is driven by observations within the coefficients of the ILP (Interleave Product), introduced in [1], a measure based upon the Dual-Tree Complex Wavelet [4]. We have a tendency to summarize the properties of the ILP in more detail in section 2, along with the ICP (Inter Coefficient Product), introduced in [2], that is used to verify the particular orientations of these features.

#### B. CLASSIFICATION METHODS

In machine learning and statistics, classification is the downside of distinguishing to that of a group of classes (sub-populations) a new observation belongs, on the idea of a training set of data containing observations (or instances) whose class membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or distribution a designation to a given patient as delineated by discovered characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition. In machine learning terminology,[1] classification is taken as an instance of supervised learning, i.e. learning wherever a training set of properly determined observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into classes based mostly on some measure of inherent similarity or distance. Often, the individual observations are analyzed into a collection of quantitative properties, known variously as explanatory variables or features. These properties could variously be categorical (e.g. "A", "B", "AB" or "O", for blood type), ordinal (e.g. "large", "medium" or "small"), integer-valued (e.g. the number of occurrences of a distinct word in an email) or real-valued (e.g. a measuring of blood pressure). Different classifiers work by examining observations to previous observations by means of a similarity or distance function. An algorithm or formula that implements classification, particularly in a concrete implementation, is known as a classifier. The term "classifier" sometimes additionally refers to the mathematical function, implemented by a classification formula, that maps input data to a class.

#### C. NEURAL NETWORK

The Feed Forward Neural Network with Back Propagation technique has found their approach into several real-time application. Feed Forward Neural Network makes use of activation function. Activation function is employed to proportion the output of various layers in Neural Network. Back Propagation is a common method by which we can train the network. Weight Matrix of Neural Network is adjusted with training method to produce needed results. During this system the value of perceptron is depends upon the inputs and their weight values. In the implementation of perceptron we tend to turn out a threshold value and assume if the result will greater than that value the output will be one otherwise zero. The feed-forward neural network is also a network of perceptron with a differentiable squashing perform, generally the sigmoid perform. The back propagation formula adjusts the weights supported the concept of minimizing error square. The differentiable squashing perform permits the rear propagation formula to control the weights across multiple hidden layers.

$$perceptronoutput = 1 \text{ if } \sum \text{of product of inputs and weights} > theta \quad (1.3)$$

$$\text{otherwise, perceptronoutput} = 0 \quad (1.4)$$

Output is calculated for each input value if the output is correct then no change is required to threshold or weights.

if the output is 1 but it should be 0

then

$$theta = theta + 1 \quad (1.5)$$

and

$$weight_i = weight_i - 1, \text{ if } input_i = 1 \quad (1.6)$$

if the output is zero but it should be one

then

$$\{theta = theta - 1\} \quad (1.7)$$

And

$$\{weight_i = weight_i + 1, \text{ if } input_i = 1\} \quad (1.8)$$

Where, i is a particular input node and weight pair.

#### D. SUPPORT VECTOR MACHINE

In machine learning, support vector machines (SVMs, additionally support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a group of training examples, each marked for belonging to at least one of two classes, an SVM training algorithm builds a model that assigns new examples into one class or the alternative, creating it a non-probabilistic binary linear classifier. An SVM model is a illustration of the examples as points in area, mapped in order that the samples of the separate classes are divided by a clear gap that is as wide as possible. New examples are then mapped into that very same area and foreseen to belong to a class supported based on which side of the gap they fall on. Additionally to playing linear classification, SVMs will with efficiently perform a non-linear classification using what is referred to as the kernel trick, implicitly mapping their inputs into high-dimensional feature areas.

#### E. REGRESSION ANALYSIS

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes several techniques for modeling and analyzing several variables, once the main target is on the link between a dependent variable and one or more further freelance variables (or 'predictors'). More specifically, regression analysis helps one understand however the typical value of the dependent variable (or 'criterion variable') changes once anyone of the freelance variables is varied, whereas the various freelance variables are held fixed.

Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the freelance variables – that's, the average value of the freelance variable when the freelance variables are mounted. Less commonly, the main target is on a quantile, or alternative location parameter of the conditional distribution of the dependent variable given the freelance variables. Altogether cases, the estimation target are a function of the freelance variables referred to as the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be delineating by a probability distribution. Regression analysis is widely used for prediction and forecasting, wherever its use has substantial overlap with the sector of machine learning. Regression analysis is additionally used to understand which among the independent variables are related to the dependent variable, and to explore the kinds of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the freelance and dependent variables. But this may cause illusions or false relationships, therefore caution is advisable; for example, correlation does not imply exploit.

## II. RELATED WORK

Tan[1] define here we need development of effective strategies for the correct iris recognition from far distance face or eye pictures. Here Author using Stabilized Iris Encoding and Zernike Moments Phase Features for the correct Iris Recognition at a Distance. Author defines here the nonlinear approach. A nonlinear approach at the constant time account for each native consistency of iris bit and also the general quality of the weight map. Here we tend to use the Zernike moment based phase encoding of iris features for the classification of native iris features. Here the author uses the algorithm based on three basic databases: 1) UBIRIS.v2; 2) FRGC; and 3) CASIA.v4-distance. Features are extracted using 1D log-Gabor filter and the parameter wavelength of the 3 employed databases.

Doyle [2] has been worked on the Robust Detection of Textured Contact Lenses in Iris Recognition Using BSIF. Once author creating an algorithm for the robust detection of textured contact lenses in iris recognition pictures then 3 problems arises. The primary issue is whether or not the correct segmentation of the iris region is needed as to achieve the correct detection of textured or unsmooth contact lenses. The experimental results counsel that correct iris segmentation isn't needed. The second issue is whether or not an associate degree algorithmic program trained on the pictures get from one device can well generalize to the pictures get from a unique device. The results counsel that due to sensor specific features trained model does not generalize with the same accuracy to another different sensor. 3<sup>rd</sup> issue is how better a detector generalizes to a full of textured contact lenses that aren't seen in training information.

Gale [3] has worked on the review on advance strategies of feature extraction in iris recognition system. Iris recognition is one altogether the foremost correct biometric identification system. The authors have given a outline of the newest analysis of feature extraction of iris recognition. Here the author presents the analysis of various feature extraction methods which are based on CASIA database. Here author uses the Local Binary Pattern and combined LVQ classifier. And uses different iris datasets like CASIA, MMUI, MMU2 and LEI.

Li, Peihua [4] has been present the Iris recognition algorithm in non-ideal imaging conditions. Here author discuss the downsides that arise when pictures are captured in non-ideal conditions. Noisy factors like the off-axis imaging, pose variation, image blurring, illumination modification, occlusion, reflective highlights and noise therefore due to these downsides iris recognition becomes difficult. We tend to introduce a robust algorithm based on the Random Sample Consensus (RANSAC) for localization of ellipsoid iris boundaries. Random Sample Consensus method can detect the iris boundaries a lot of accurately than the strategies based on the Hough transform. Author describes an pictures registration technique which is based on the LucasCKanade algorithm. Author defines this technique to account for iris pattern deformation. This system works on filtered iris images. This technique solves the registration problem for every small sub-image and divides the small pictures into other small sub-pictures. Here author use sequential forward selection method and Gabor filters.

Santos [5] define iris recognition is more preferred because of its uniqueness, period of time stability and its stable shape which helps us in correct segmentation and recognition. Author has presented a fusion approach to unrestricted iris recognition. Main motive of research with the use of new techniques lowering these constraints that aren't impacting performance whereas increasing system usability and new approaches have rapidly emerged. Here author describes the iris boundary, iris normalization, feature extraction. Here author present a completely unique fusion of several recognition techniques to check the downsides of no cooperative iris recognition using nonideal visible-wavelength pictures clicked in unrestricted environment.

### III. FINDINGS OF LITERATURE STUDY

For Iris recognition existing system use the Zernike moments with the stabilized Iris encoding. Completely Different dataset are used to evaluate the results and there average accuracy has been recorded UBIRIS v2 with 54.3%, FRGC with 32.7% and CASIA with 42.6%. The proposed model focuses on the Iris recognition accuracy improvement because existing models offers the accuracy below 54.3%. So, planned model use the robust feature extractor according to the adaptability of feature descriptor methodology to increase the accuracy Iris features. The planned system uses the support vector machine (SVM) for Iris classification. Neural network is additionally use to boost the accuracy of probabilistic classification of SVM.

Due to elevated false rate evaluated by the Zernike moment based feature descriptor, existing system faces the lower accuracy. There is two types of errors are available type 1 and type 2 to describe the false positive. Wrong Iris recognition is taken as the false positive. To avoid the false positive neural network is used because of their robust and accurate nature. To do so neural network uses the probabilistic classification and learning based classification of SVM. SVM removes the non-matching or less-matching samples for the preprocessing classification. After this, most matched template will be short-listed using the fuzzy set which short listed template will be sent to the SVM. Fuzzy sets preprocessing makes the process faster and more accurate as compared to the existing model.

### IV. CONCLUSION

The biometric authentication using the Iris features requires a lot of statistical and mathematical operation to achieve the classification results. The Iris recognition models are considered accurate with the controlled feature selection based databases. The controlled feature selection databases consider the similar feature acquisition in the similar scenario for all persons. For the web based database collection or open database platform for based collection, it is never possible to keep the similar position based features. The more strong and extremely dynamic systems are needed for such databases. The main issue lies with the angular shift of the Iris features in the Iris recognition systems, which is the primary concern of the proposed solution during this paper. The angular shift feature (ASF) along with support vector machine based early feature elimination procedure in multi-step classification is being incorporated during this paper for the upper accuracy and strength.

#### ACKNOWLEDGMENT

I am heartily thankful to my supervisor, Er. Rajan Goyal, Assistant professor of Computer Engineering section whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of my work.

#### REFERENCES

- [1] C. W. Tan, and A. Kumar, "Accurate Iris Recognition at a Distance Using Stabilized Iris Encoding and Zernike Moments Phase Features." *Image Processing, IEEE Transactions on* 23, no. 9 (2014): 3962-3974.
- [2] J. S. Doyle, and K.W. Bowyer, "Robust Detection of Textured Contact Lenses in Iris Recognition Using BSIF." *Access, IEEE* 3 (2015): 1672-1683.
- [3] G. Gale, and S. S. Salankar, "A Review on Advance Methods Of Feature Extraction In Iris Recognition System." *IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE) e-ISSN* (2014): 2278-1676.
- [4] L. Peihua, and M. Hongwei, "Iris recognition in non-ideal imaging conditions." *Pattern Recognition Letters* 33, no. 8 (2012): 1012-1018.
- [5] G. Santos, and E. Hoyle, "A fusion approach to unconstrained iris recognition." *Pattern Recognition Letters* 33, no. 8 (2012): 984-990.
- [6] K.Y. Shin, G. P. Nam, D. S. Jeong, D. H. Cho, B. J. Kang, K. R. Park, and J. Kim, "New iris recognition method for noisy iris images." *Pattern Recognition Letters* 33, no. 8 (2012): 991-999.
- [7] R. Szewczyk, K. Grabowski, M. Napieralska, W. Sankowski, M. Zubert, and A. Napieralski. "A reliable iris recognition algorithm based on reverse biorthogonal wavelet transforms." *Pattern Recognition Letters* 33, no. 8 (2012): 1019-1026.
- [8] W. Y. Han, W. K. Chen, Y. P. Lee, K. S. Wu, and J. C. Lee, "Iris Recognition based on Local Mean Decomposition. " *Appl. Math* 8, no. 1L(2014):217-222.

# Logo Image Based Approach for Phishing Detection

Himani Thakur  
Computer Science Engineering  
PURCITM Mohali  
[thakurhimani3@gmail.com](mailto:thakurhimani3@gmail.com)

Dr.Supreet Kaur  
Computer Science Engineering  
PURCITM Mohali  
[skaur.gujral@gmail.com](mailto:skaur.gujral@gmail.com)

**Abstract**— Phishing is a cyber attack which involves a fake website mimicking the some real legitimate website. The website makes the user believe the website being authentic and thus online user provides their sensitive information like password, PIN, Social Security Number, and Credit Card Information etc. Due to involvement of such high sensitivity information, these websites are a huge threat to online users and detection and blocking of such website become crucial. In this thesis, we propose a new phishing detection method to protect the internet users from such attacks. In particular, given a website, our proposed method will be able to detect between a phishing website and a legitimate website just by the screenshot of the logo image of it. Due to the usage of screenshot for extracting the logo, any hidden logo will not be able to spoof the algorithm into considering the website as phishing as happened in existing methods. In first study focus was on dataset gathering and then the logo image is extracted. This logo image is uploaded to Google image search engine using automated script which returns the URLs associated with that image. Since the relationship between logo and domain name is exclusive it is reasonable to treat the logo image as identity of original URL. Hence the phishing website will not have the same relation to the logo image as such and will not get returned as URL by Google when search for that logo image. Further, Alexa page rank is also used to strengthen the detection accuracy.

**Keywords**— Anti-phishing; Website logo; Google image search

## I. INTRODUCTION

Phishing is the act of mimicking a trusted website to gain sensitive information from online users like detail of credit card, personal identification number etc. Since APWG reports claim that 40-50% of phishing attacks are based on common legal web sites, we decided to check this and so we compiled a list of target words which included many popular phishing targets, such as Ebay and paypal [20]. In most of cases criminals make web pages by copying legitimate or make a little change in page content to gain user's sensitive information. For example, a system can be technically secure enough against password stealing, however uninformed end users if click on the Hypertext Transfer Protocol (HTTP) link may leak their passwords, which ultimately threatens the overall security of the system. There are many solution exist to detect phishing attack but no one bullet proof solution yet to present which detect all type of phishing attack. In order to detect whether the website is phishing website, the first question to ask is: how to discriminate phishing website and the legitimate website as the reason is that the phishing website is look alike to the legitimate website. Where if we have the portrayed identity of the query website then we can find out that if it a legitimate or a phishing website. (if the doubt website is a phishing website, the portrayed identity will be the identity of the besieged legitimate website)[1], we can then differentiate the phishing website from the legitimate website. Knowing that the phishers will use the optical factors ripped off from the legitimate website, especially the logo, in their phishing websites, this inspires to propose an anti-phishing method based on the recognition of website identity through the logo. This is rational, as the logo usually symbolizes the identity of a legitimate website.

**Query website:** The website which is under test to check if it is a phishing website or a legitimate website.  
**Portrayed identity:** The trademark or entity for which a legitimate website is emphasize to. As for example, the domain <http://www.ebay.com> where the portrayed identity of the legitimate website is ebay. Likewise, for e.g., domain <http://www.www1-ebaee.com> is a phishing website which mimic the ebay website, where the portrayed identity is ebay.  
**Real identity:** That is the actual identity of a query website. For example, the domain <http://www.ebay.com> where ebay is the real identity of a legitimate website. Whereas the domain <http://www.www1-ebaee.com> is a phishing website which mimic the ebay website, its real identity is www1-ebaee. There are various type of phishing such as email phishing, malware based phishing, keylogger and screenloggers (particular type of malware that track the keyboard input and send the relevant information to hacker via the internet), man in the middle phishing, etc. There have been several anti phishing technique developed in last few years.

## II. Related work

While there exists numerous different techniques in phishing detection. These are as following:

Kang Leng Chiew et.al [1] used a logo image to find out the identity reliability between the real and the portrayed identity of a website. reliable identity point towards a legitimate website and incompatible identity point towards a phishing website. The proposed technique consists of two procedures, that is logo extraction and identity

verification. The first procedure will identify and take out the logo image from all the downloaded image resources of a webpage. In order to identify the right logo image, the method make use of a machine learning technique. Based on the take out of a logo image, the second procedure will utilize the Google image search to retrieve the portrayed identity. Since the connection between the logo and domain name is special, it is logical to treat the domain name as the identity. Hence, a difference between the domain name returned by Google with the one from the query website will allow us to distinguish a phishing from a legitimate website. The carry out experiments show consistent and promising results. This proves the efficiency and possibility of using a graphical element such as a logo to identify a phishing website.

J. Hong and L. Cranor et al. [7] The majority of the proposed techniques in the literature, the unmediated heuristic approach. One of the popular methods is CANTINA. This method will calculate the TF-IDF from the content of a webpage, and produce a lexical signature. The technique will utilize the generated lexical signature to do a web search through the Google search engine. The returned outcome will be used to conclude the authority of a website. Even though this technique can carry out logically well in the finding of phishing. J. Lee et al. [8] Proposed a more recent research study on the characteristics-based heuristic approach is the one proposed by the main objective of the technique is to detect the identity of the phishing target when a phishing webpage is identified. The technique is based on the design of a self-organised semantic data model, labeled as the Semantic Link Network which is frequently used in organising web resources. While this method is using different detection mechanisms, its basis starts from the textual elements (i.e., the taking out of hyperlinks, keywords and textual contents for the process of link relations, search relations and text relations, respectively). Cao Y et al. [9] Proposed one can gather a list of legitimate URLs. This method is recognized as whitelisting, and it is also a kind of list-based approach. An example of a whitelisting technique is the research proposed by the authors developed an automated technique that maintains and stores a whitelist at the client side. Prakash P et al. [10] proposed a more active and flexible list-based approach is called PhishNet. This technique uses several URL variant heuristics to procedure the existing blacklisted URLs and make multiple variant URLs. The produce URLs will form a analytical blacklist. The results explain that it can successfully detect new and old phishing websites. While a list-based approach provides ease in design and is easy to put into operation, keeping the list complete and up-to-date needs great attempt, and always go through from incompleteness. Tout and Hafner et al. [4] Proposed one of the popular techniques is blacklisting. Many well-liked web browsers are utilize this approach to detect phishing website in this technique, a query website is checked with a list (i.e., a list is recognized as phishing URLs), which is collected and upheld by some association or organisation. If the checking returns a match, then the website will be labeled as phishing. Sadia Afroz et al. [23] proposed one of the popular techniques is PhishZoo. This paper proposes a phishing detection approach—PhishZoo that utilizes profiles of reliable websites outer shells to detect phishing. The advantage is that it can categorize zero-day phishing attacks and embattled hits against minor sites (such as corporate intranets). A key role of this paper is that it comprise a presentation study and a structure for making use of computer vision techniques in a sensible way. Zhuang et.al [12] proposed a technique that is deliberated and applied an intelligent model for detecting phishing websites. In this model, they take out 10 different types of kind such as heading, keyword and connection text information to stand for the website. Various classifiers are then build stand on these dissimilar features. They proposed a ethical ensemble classification algorithm to join the expect results from different phishing detection classifiers. Hierarchical clustering technique has been working for mechanical phishing categorization. Case studies on great and actual daily phishing websites composed from King soft Internet Security Lab demonstrate that their proposed model outperforms other commonly used anti-phishing methods and tools in phishing website finding. Bian et.al [14] proposed a method to assess the effectiveness of three popular online resources in identifying phishing sites-viz, Yahoo! Inlink data and Yahoo! directory service, Google PageRank system. Their results point towards that these online resources can be used to boost the accuracy of phishing site detection when used in combination with existing phishing countermeasures. The proposed loom involves investigate the following three attributes of a goal site (site being check up): (1) the reliability of the target sitepsilas hosting domain, (2) the reliability of in-neighbor sites that link to the hosting domain, and (3) the connection between the aim sitepsilas web category and its hosting domainpsilas web kind. The abovementioned online resources by themselves are insufficient to concentrate on the phishing attack problem. This approach provide convention on how each of those resources may be included with existing phishing detection techniques to offer a more efficient solution. Ali et.al in [15] proposed a approach of confidentiality in Instant Messengers (IM) by means of Association Rule Mining (ARM) method a Data Mining approach included with Speech Recognition system. verbal skills are acknowledged from words with the help of FFT spectrum analysis and LPC coefficients methodologies. Online criminal's at the present time modified voice chatting technique along with text messages collaboratively or either of them in IM's and squashing out personal information direct to intimidation and barrier for privacy. To facilitate centre of attention on privacy preserving this approach residential and try out Anti Phishing Detection system (APD) in IM's to detect

unreliable phishing for text and audio collaboratively. Tan et.al in [16] proposed an anti-phishing method to protect users against phishing attacks in the internet. The scope of this approach study focuses mainly on the detection of phishing websites with English content. In order to encourage users on whom the website claims to be, phishers usually place brand names in different parts of the URL. They oppressed this phishing pattern by conveying weights to words taken out from the HTML content, based on their co-appearance at path, hostname and file names of URLs. These weights are then supplementary to their equivalent TF-IDF weights. The most likely words are particular and submitted to Yahoo Search to recover the highest frequency domain name amongst the top 30 search results. A WHOIS lookup is carried out to disclose the vendor behind the selected domain name. A phishing website can be easily illustrious if the vendor of query domain name be different from the owner of domain name returned by the search engine.

Fang et.al in [17] proposed a approach of an artificial protected system for phishing detection. The system is to sense phishing emails throughout mature detectors and memory detectors. The memory detectors are produced from the training data set, which consecutively contains the phishing emails up to that time seen by the system. The immature detectors are replicate through the system's mutation procedure. To the best of this approach facts that this is the first time such a system is ever projected. They assumed that the system is more adaptive than any other active phishing detection techniques. Nguyen et.al in [18] proposed an efficient approach for identifying phishing websites foundation on the single-layer neural network. Particularly, the proposed technique calculates the value of heuristics impartially. Then, the weights of heuristic are produced by a single-layer neural network. The proposed technique is assessed with a dataset of 11,660 phishing sites and 10,000 legitimate sites. Jo et.al in [19] proposed a approach to consider websites' identity claims. Their phishing detection system copy this human expert behavior. Given a website, their system study the identity that this website assert, and figure the documentary significance between this claimed identity and other description in the website. Their phishing detection system then employ this textual significance as one of the sort for classification. DeBarr et.al [3] proposed a approach as a first step the exercise of Spectral Clustering to analyze messages based on traffic behavior. Specifically, Spectral Clustering analyzes the association between URL substrings for web sites originate in the message contents. Cluster membership is then employ to assemble a Random Forest classifier for phishing. Data from the Phishing Email quantity and the Spam killer Email quantity are used to evaluate this approach. Performance assessment metrics include the region Under the receiver operating characteristic Curve (AUC), as well as accurateness, exactness, evoke, and the (harmonic mean) F measure. Presentation of the incorporated Spectral Clustering and Random Forest loom is found to provide important developments in all the metrics listed, contrasted to a satisfied filtering technique such as LDA joined with text message deletion done arbitrarily or in an adaptive fashion using adversarial learning. The Spectral Clustering approach is strong against the lack of content. Gowtham et.al [2] proposed a study, the features of legitimate and phishing webpages were examined in depth, and support on this analysis, this approach proposed heuristics to take out 15 characters from such webpages. These heuristic results were fed as an contribution to a trained machine learning algorithm to identify phishing sites. To the webpages before alarmed heuristics, this approach worn two initial screening modules in this system. The first component, the preapproved site identifier, verify webpages against a confidential of white-list maintained by the user, and the second part, the Login Form Finder, categorize webpages as legitimate when there are no login appearances present. Deshmukh et al.[11] proposed a approach as cyber crime is technology based fault committed by technocrats. This paper deals with modification of cyber crime like Packet Sniffing, Salami Attack, Bot Networks and Tempest Attacks. It also contains real world cyber crime suitcases their situation and modus operandi. The worldwide malware, rate spam rate and phishing rate is rising speedily. And there is a latent shock of cyber crime on consumer trust, economics and production time. The contradict ways similar to Intrusion Detection, GPRS Security architecture and Agent Based Distributed Intrusion Detection System and prevention System are utilized for safety reason. Verma et.al.[13] propose a approach that merged statistical examination of website URLs with machine learning methods will give a additional precise classification of phishing URLs. Employing a two-sample Kolmogorov-Smirnov examination along with other description. Thus, correctness of phishing URL categorization can be very much improved through the use of these statistical measures.

### III. PROPOSED METHODOLOGY

The proposed problem aims to study the phishing detection by using web logo approach. The methodology comprises of following two processes:

First process will capture the screenshot and perform the approach directly to extract the logo. This approach has a few advantages. As the research work will focus on as a replacement for finding the logo image from a pool of downloaded images (image income of a query webpage). Will be capture the screenshot and directly extract the logo. This approach has a few advantages. As, the captured screenshot is actual offer the web content, which means there is no other secret image. By directly capturing the screenshot will provide the actual web content which is

usually used to optimise website loading speed. Google image search will provide the undesired result by using sprite type of images as a query result even though the logo existed within the sprite image. Another advantage is the logo removal from the poster image of a website will be more precise. In other words, by directly extracting the logo images will provide no other non-logo images. Second process will utilize the Google image search to retrieve the actual identity. As the link between the logo and domain name is special, it is realistic to treat the domain name as the identity. As a contrast between the domain name returned by Google with the one from the query website will allow us to differentiate a phishing website from a legitimate website. Using a graphical element such as a logo to detect a phishing website. Alexa rank of the website is extracted and matched under the range less than 10000 for providing more accuracy in phishing detection.

Consistent identity indicates a legitimate website and inconsistent identity indicates a phishing website.

#### A. Design Consideration

In the design consideration, the proposed approach considered the structure of data flow for the design of the experimental setup. We started by analysing the requirements. The requirements can be listed as follow:

1. Database of Phishing and Non-Phishing websites.
2. Screenshots of the Website under consideration.
3. Processing Tool to Extract Website Logo Image.
4. Automated tool to Detect the Phishing Site by Logo Image Screenshot.

To verify these design considerations, we started by collecting the database of phishing and non-phishing sites. Phishload [5] is an open source database that have been used in the work. Screenshots of the webpages are collected from PhishTank [6]. A url from PhishTank returns the screenshot of the webpage if available. Processing Tool to extract website logo image is developed in Java. It is an assisted cropping tool and the user has to draw a rectangle around the logo image and the image is extracted. The detection of Phishing Site is done by another tool developed in java.



Figure 1: A Google Image Search Result

In figure 1 shows that after searching the logo image of a website in the google image search, it shows the result that whether it is a phishing website or a legitimate website. It shows the best guess of the searched logo image.

#### B. Flowchart Of Concept

The flowchart of the concept is mentioned in the figure 2. As per the flow of code, we will load a phishing url (known to us only). We will load an phishtank id from database and we will perform a search to get its screenshot image. After getting the screenshot image, we will open the Crop Image Tool and Crop the Logo Region. In each step, a logo image is taken from the database which is cropped and stored in a database. The image is uploaded to google image search website. The google returns the results in terms of a best guess value and some number of urls (search results of websites). If the query logo's URL exists in the list of urls returned by google image search, the website is marked as legitimate, instead, if the website is not directly listed in the set of urls in the list, the alexa rank of the website is extracted and matched under the range less than 10000.

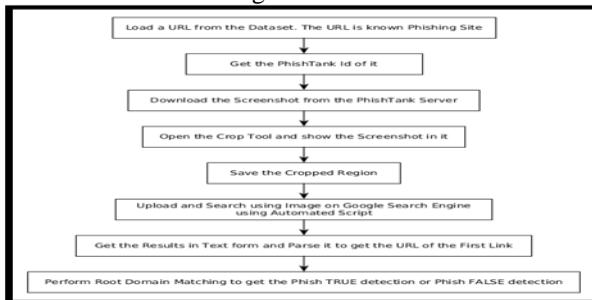


Figure 2: Flowchart of the Proposed Work

### C. Results and Discussions

In each step, a logo image is taken from the database which is cropped and stored in a database. The image is uploaded to google image search website. The google returns the results in terms of a best guess value and some number of urls (search results of websites). If the query logo's URL exists in the list of urls returned by google image search, the website is marked as legitimate, instead, if the website is not directly listed in the set of urls in the list, the alexa rank of the website is extracted and matched under the range less than 10000.

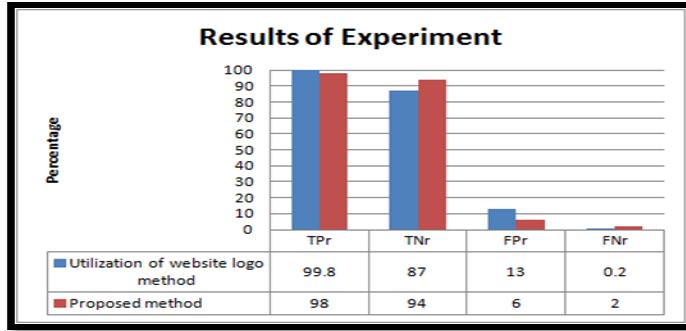


Figure 3: Graph of Performance Analysis of Our System compared to Utilization of website logo based Method

In Figure3 the True Positive rate of our system is 98% while the True Positive rate of Utilization of logo image based method is 99.8 %, further, the True Negative rate of our system is 94% whereas the true negative rate of Utilization of logo image based method is 87 %.

### IV. Conclusion

In this work we have developed a new method to detect phishing websites based on the logo image and the base url of the website. The system has shown a 98% detection rate of phishing website because the logo used by phishing website returns the search of original websites or other websites that have backlinked the original website but the test website's url never appears in the search result. Thus making a 98% accurate detection because some of the alexa ranks were skewed towards base URL. The False Positive Rate is imported from previous work by more than 50 %. But since many websites mention the same logo image to backlink a popular website, the masking effect happens and thus real websites are detected as phishing website too.

### V. Future Scope

In the future work, we can add more parameters like Google PageRank, number of backlinks etc in order to increase the overall confidence towards phishing as well as non-phishing website.

### REFERENCES

- [1] Chiew, K.L., Chang, E.H. and Tiong, W.K., 2015. Utilisation of website logo for phishing detection. *Computers & Security*, 54, pp.16-26.
- [2] Gowtham, R. and Krishnamurthi, I., 2014. A comprehensive and efficacious architecture for detecting phishing webpages. *Computers & Security*, 40, pp.23-37.
- [3] DeBarr, D., Ramanathan, V. and Wechsler, H., 2013, June. Phishing detection using traffic behavior, spectral clustering, and random forests. In*Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on* (pp. 67-72). IEEE.
- [4] Tout, H. and Hafner, W., 2009, August. Phishpin: An identity-based anti-phishing approach. In *Computational Science and Engineering, 2009. CSE'09. International Conference on* (Vol. 3, pp. 347-352). IEEE.
- [5] Phishload. 2016. *Phishload*. [ONLINE] Available at: <http://www.medienifi.lmu.de/team/max.maurer/files/phishload>. [Accessed 01 July 2016].
- [6] PhishTank | Join the fight against phishing. 2016. *PhishTank / Join the fight against phishing*. [ONLINE] Available at: <http://www.phishtank.com>.
- [7] K. L. Chiew, E. H. Chang, S. N. Sze, and W. K. Tiong, "Utilisation of website logo for phishing detection," *Comput. Secur.*, pp. 1–11, 2015.
- [8] J. Hong and L. Cranor, "CANTINA : A Content-Based Approach to Detecting Phishing Web Sites," pp. 639–648, 2007.
- [9] J. Lee, D. Kim, and L. Chang-Hoon, "Heuristic-based Approach for Phishing Site Detection Using URL Features," *Adv. Comput. Electron. Electr. Technol.*, pp. 131–135, 2015.
- [10] Cao Y, Han W, Le Y, "Anti-phishing based on automated individual white-list," Proceedings of the 4th Workshop on Digital Identity Management., pp. 51e60,2008.
- [11] Prakash P, Kumar M, Kompella RR, Gupta M. PhishNet, "predictive blacklisting to detect phishing attacks," INFOCOM 2010 29<sup>th</sup> IEEE International Conference on Computer Communications., pp. 346e50,2010.
- [12] Deshmukh, J.J. and Chaudhari, S.R., 2014. Cyber crime in Indian scenario—a literature snapshot. *International Journal of Conceptions on Computing and Information Technology*, 2(2).
- [13] Zhuang, W., Jiang, Q. and Xiong, T., 2012, June. An intelligent anti-phishing strategy model for phishing website detection. In *Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on* (pp. 51-56). IEEE.

- [13] Verma, R. and Dyer, K., 2015, March. On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers. In *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy* (pp. 111-122). ACM.
- [14] Bian, K., Park, J.M., Hsiao, M.S., Belanger, F. and Hiller, J., 2009, July. Evaluation of online resources in assisting phishing detection. In *Applications and the Internet, 2009. SAINT'09. Ninth Annual International Symposium on* (pp. 30-36). IEEE.
- [15] Ali, M.M. and Rajamani, L., 2012, March. Deceptive phishing detection system: from audio and text messages in instant messengers using data mining approach. In *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on* (pp. 458-465). IEEE.
- [16] Tan, C.L. and Chiew, K.L., 2014, December. Phishing website detection using URL-assisted brand name weighting system. In *Intelligent Signal Processing and Communication Systems (ISPACS), 2014 International Symposium on* (pp. 054-059). IEEE.
- [17] Fang, X., Koceja, N., Zhan, J., Dozier, G. and Dipankar, D., 2012, June. An artificial immune system for phishing detection. In *Evolutionary Computation (CEC), 2012 IEEE Congress on* (pp. 1-7). IEEE.
- [18] Nguyen, L.A.T., To, B.L., Nguyen, H.K. and Nguyen, M.H., 2014, October. An efficient approach for phishing detection using single-layer neural network. In *Advanced Technologies for Communications (ATC), 2014 International Conference on* (pp. 435-440). IEEE.
- [19] Jo, I., Jung, E.E. and Yeom, H.Y., 2010, August. You're Not Who You Claim to Be: Website Identity Check for Phishing Detection. In *Computer Communications and Networks (ICCCN), 2010 Proceedings of 19th International phishtank.com/*. [Accessed 01 July 2016].
- [20] Anti-Phishing Working Group, 2014. Phishing Activities Trends Report. avail-able at [http://docs.apwg.org/reports/apwg\\_trends\\_report\\_q1\\_2014.pdf](http://docs.apwg.org/reports/apwg_trends_report_q1_2014.pdf).
- [21] Satane, V.V. and Dasgupta, A., 2013. Survey Paper on Phishing Detection: Identification of Malicious URL Using Bayesian Classification on Social Network Sites. *International Journal of Science and Research (IJSR)*.
- [22] Khonji, M., Iraqi, Y. and Jones, A., 2013. Phishing detection: a literature survey. *IEEE Communications Surveys & Tutorials*, 15(4), pp.2091-2121.
- [23] Afroz, S. and Greenstadt, R., 2011, September. Phishzoo: Detecting phishing websites by looking at them. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on* (pp. 368-375). IEEE.

# Implementing Modified Cluster Head Selection in Existing TEEN Routing Protocol

Gurjeet Singh<sup>1</sup>, Er. Karandeep Singh<sup>2</sup>

<sup>1</sup>M.Tech Student, <sup>2</sup>Assistant Professor

<sup>1</sup>sohigurjeet@gmail.com, <sup>2</sup>karan\_rob7@yahoo.co.in

Department of Computer Engineering, Punjabi University, Patiala, Punjab (India)

**Abstract-** The study of various protocols in wireless sensor networks shows that energy is the key concern in WSN due to the limited irreplaceable power sources of sensor nodes. The main reason for that is large amount of energy used in data transmission. Various researches take place to improve the energy efficiency and lifetime of WSN by developing routing algorithms. In this paper, Modified cluster head selection implemented in existing TEEN routing protocol to enhance the energy efficiency of the existing protocol in wireless sensor networks. A variable value of threshold energy is used for the nodes to aggregate data and transmit it to further nodes, which leads to increase the residual energy of each node and hence the performance. We evaluated the results after implementing the modified cluster head selection in existing TEEN routing protocol and observed that it performs better than existing TEEN routing protocol.

**Keywords**-Cluster Head Selection, TEEN, Residual Energy, Dead Nodes, ELF, Energy Efficiency, Lifetime.

## I. INTRODUCTION

In now a day, various applications of wireless sensor networks developed to allow the human being to interact with environment [7]. Wireless sensor network assist to gather the sensed information about the environment and deliver it to the BS or upper level node. Sensor nodes can be homogeneous or heterogeneous depending upon the requirement of the application and deployed at random or predefined locations. In recent days, many protocols designed to improve the scalability of sensor nodes and the energy efficiency of the WSN [12]. The enhancement in energy efficiency leads to improve the lifetime of the wireless sensor network. Hence, hierarchical clustering comes into account for better network routing and performance. In hierarchical clustering, all the sensing devices placed randomly in the specified area to sense the required information about it [9-11]. That informational data processed and transmitted to upper level nodes by the cluster representative. Cluster representative known as cluster head. Clusters are the segments of the system formed by some neighborhood nodes [4]. A cluster head is selected to handle the work of transmission and aggregation as a cluster representative. Each node selected as CH to balance the energy consumption among all the nodes of the network. The procedure of cluster head selection is the main concern for most of the protocols.

## II. RELATED WORK

Heinzelman et. al [12] Introduced LEACH (*Low-Energy Adaptive Clustering Hierarchy*) routing protocol, which is a proactive network protocol used to collect the information about the environmental and other physical parameters by sensing the specified area periodically. LEACH was the first protocol to use the hierarchical clustering technique [8]. Thus it is the typical hierarchical clustering protocol that improves the lifetime of the network by using

mechanism of cluster head rotation, data fusion and aggregation technologies [12]. The process of cluster formation and head selection performed according to this formula. Here,  $T(n)$  is , p is the probability and n is number of nodes.

$$T(n) = P \div r \bmod \frac{1}{p} \quad \text{if } n \in G \quad (1)$$

$$T(n) = 0 \text{ if } n \text{ not belongs to } G \quad (2)$$

However, it contains some limitations i.e. not capable to perform well for large networks and it gives information on consistent basis [11]. Any sudden change in the network parameter not reported immediately. To overcome these issues, TEEN (*Threshold Sensitive Energy Efficient Network*) routing protocol is introduced by Manjeswar et. al [11]. Being reactive network protocol is to react immediately, when a sudden change in the parameter achieved. This property of TEEN of being reactive done by introducing hard threshold and soft threshold values of the parameters to be sense in the given field. TEEN uses same clustering technique for sensing data and transmitting information as done in LEACH routing protocol [8][11]. Environment sensed by the cluster members of the cluster does not sent the sensed data to the CH until threshold values not achieved. When the thresholds achieved CH receives data from cluster nodes. CH further performs the task of data aggregation and fusion to avoid the redundancy of the sensed data and then transmit it to the upper level node in the network [10]. The major drawback of the TEEN routing protocol is that there will be no information about the area until the threshold values not achieved. It does not give the sensed information on consistent basis [10-11]. However, it improves the lifetime of the network by reducing the number of transmissions. Time line for TEEN routing protocol (Fig. 01) shows the sequence of activities occurred in TEEN with respect to time [11].

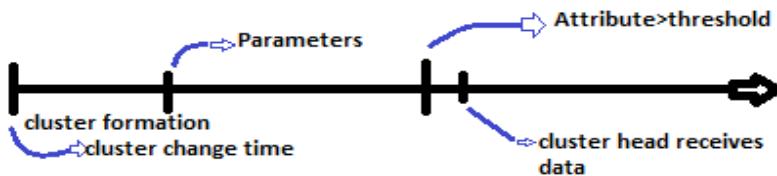


Fig. 01 Time line for TEEN routing protocol

Manjeswar et. al [10] introduced Adaptive Periodic TEEN (APTEEN) to eliminate the limitation of TEEN which does not transmit signal on consistent basis. A hybrid network protocol is better having more qualities i.e. APTEEN can act as both proactive and reactive depending upon the requirements. APTEEN includes two more parameters to handle the problems occurred in TEEN [8-10].

- 1) TDMA Schedule
- 2) Count Time

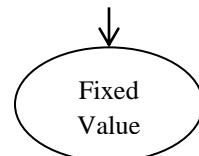
It can also handle all three types of queries i.e. historical, one-time and persistent. TDMA schedule used to avoid the collision of data transmitted by different at same time [10]. These two parameters in APTEEN increase its complexity than other hierarchical wireless sensor network protocols.

### III. PROPOSED METHOD OF IMPLEMENTING MODIFIED CLUSTER HEAD SELECTION IN TEEN:

Existing TEEN uses the same hierarchical clustering technique, which LEACH protocol used for cluster formation and head selection. Cluster head changed if its residual energy is less the required value of energy for data aggregation and transmission, which is necessary for node to be CH [1][3]. In existing TEEN, cluster head selection uses same fixed (static) value of threshold energy for every round as shown below.

*Residual Energy of CH node =*

*Initial Energy of CH node – ELF × Energy required for data aggregation and transmission*



In Modified cluster head selection, dynamic value for threshold energy used instead of fixed value of threshold energy. Because as the nodes starts to be dead. There will be less sensed data in the cluster due to the lesser sensor nodes and hence, lesser energy will be required to transmit it [1][3]. The process of modified cluster head selection introduced in [1] explains the change in value of threshold energy required for CH, obtained by introducing ELF (Energy Loss Factor), which depends on following terms.

$$\text{Energy Loss Factor(ELF)} \propto 1 \div \text{No. of Dead Nodes} \quad (2)$$

$$\text{Rounds} \propto \text{No. of Dead nodes} \quad (3)$$

Method to use the ELF value for change in threshold energy, as in [1]:

*Residual Energy of CH node =*

*Initial Energy of CH node – ELF × Energy required for data aggregation and transmission* (4)

The value of ELF derived from (2) and (3) is as follows.

$$\text{ELF} = 1 - (\text{Dead} \div N) \quad (5)$$

Here, Dead= number of dead nodes, N= total number of nodes

Equation (5) used to find the value of ELF. This shows that the initial value of ELF is one because the value for number of dead nodes is zero. This ELF value goes on to decrease as the number of dead nodes increases. ELF decreases the value of energy required for data aggregation and transmission. Thus, nodes will have better residual energy and they will not die soon.

#### IV. SIMULATION AND RESULTS

During simulation, we evaluated the performances of both the methods. The simulation performed in MATLAB 2012a (version 7.14) and the nodes placed randomly in the network area to sense temperature values. Base Station placed in the center of the network. Network Parameters are as follows:-

Parameter	Value
Network Size	100m*100m
Number of nodes	200
Initial Data Aggregation Energy	8μJ
Number of Rounds	100
Hard Threshold	100°C
Soft Threshold	2°C
Initial Temperature	50°C
Final Temperature	200°C
Initial Energy of nodes	0.5 J
Transmitter Electronics	50nJ
Receiver Electronics	50Nj

Here are some graphs of simulation results.

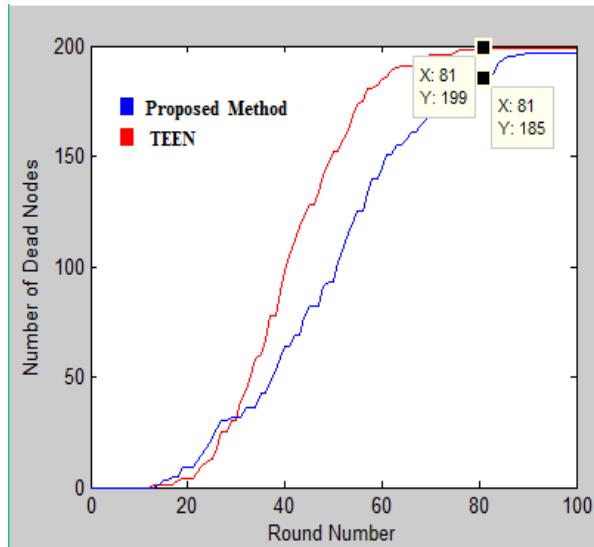


Fig. 02 Number of dead nodes

Fig. 02 shows that after applying modified cluster head selection i.e. Modified TEEN has lesser number of dead nodes as compared to existing TEEN. The data tip is showing the Value for number of dead nodes for proposed method after 81 rounds is 185 and 199 for existing TEEN. Thus, there are lesser dead nodes, which improves the working time of the network system.

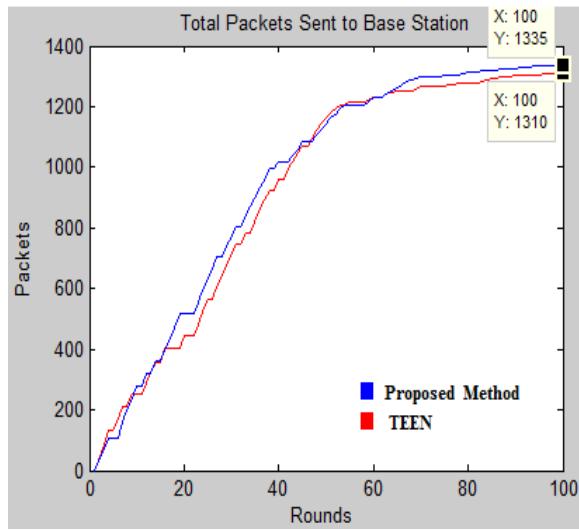


Fig. 03 Total packets sent to BS

Fig. 03 shows that modified cluster head selection increases the total packets sent to BS for TEEN routing protocol. Here 25 more packets being sent in modified TEEN than existing TEEN.

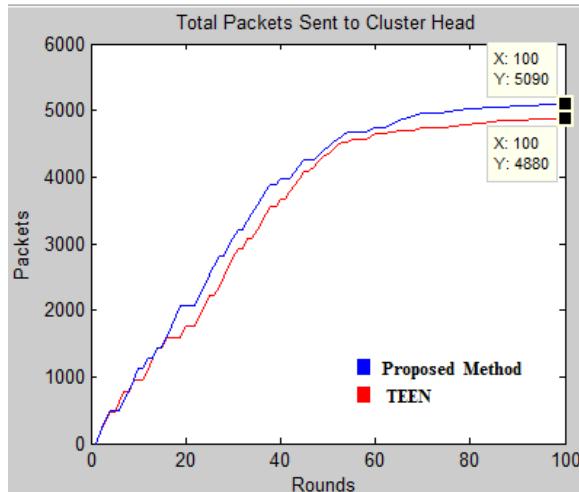


Fig. 04 Total packets sent to CH

Total packets sent after 100 rounds are 5090 for modified TEEN and 4880 for existing TEEN. Hence, 210 more packets sent to CH by modified TEEN. That shows the improvement in the performance of sensor nodes as they are transmitting more information about the area under supervision.

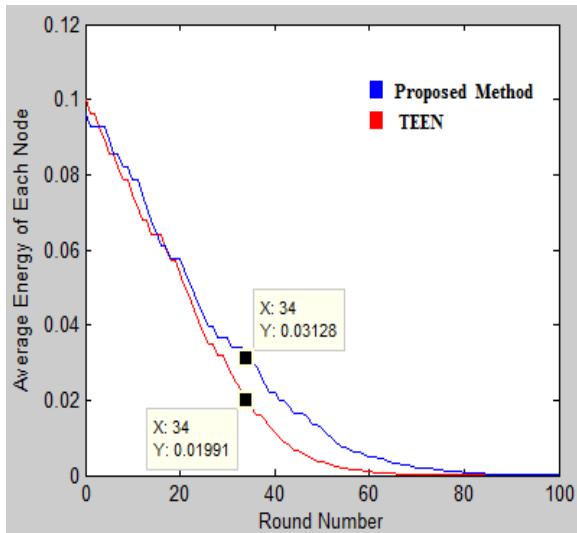


Fig. 05 Average energy of each node

The graph in above Fig. 05 shows the average energy of each node. Modified technique is showing better results than existing TEEN. Hence, it is clear that implementing modified cluster head selection in existing TEEN routing protocol increases the lifetime by reducing the number of dead nodes and improves the energy efficiency by sending more packets to the BS.

## V. CONCLUSION AND FUTURE SCOPE

In this paper, modified clustering technique implemented in existing TEEN routing protocol, which uses dynamic value of threshold energy required to aggregate and transmit data. This shows the improvement in the performance as number of dead nodes reduced by 5% as compared to existing TEEN and average energy of each node improved by 30%. This process increases the energy utilization and lifetime of the network by saving the residual energy of each node after data transmission. However, this technique increases the complexity of the protocol. Thus, there is a future scope of reducing the complexity of the protocol for better results. Here is scope of further research to find out the optimal value of threshold energy for a node to be CH in every round.

## REFERENCES

- [1] Singh, Gurjeet and Er. Karandeep Singh, "Modified Cluster Head Selection to Improve the Energy Efficiency of APTEEN routing protocol", *International Journal of Innovations & Advancement in Computer Science*, IJIACS, 2347-8616, Vol. 5(7), pp. 11-15, 2016.
- [2] Pachori, Nisha and Vivek Suryawanshi, "Cluster Head Selection Prediction in Wireless Sensor Networks", *International Journal of Computer Science and Information Technologies*, IJCSIT, Vol.6 (2), pp. 1033-1035, 2015.
- [3] Singh, Jyoti, Bhanu Pratap Singh and Shubadra Shaw, "A New LEACH-based Routing Protocol for Energy Optimization in Wireless Sensor Network", *5<sup>th</sup> International Conference on computer and Communication Technology IEEE*, pp. 181-186, 2014.
- [4] Sehgal, Nikita and Gurwinder Kaur, "Improved Cluster Head Selection Using Enhanced LEACH Protocol", *International Journal of Engineering and Innovative Technology*, IJEIT, Vol.3 (3), September 2013.

- [5] Zahra Razaei, Shima Mobininejad, "Energy Saving in Wireless Sensor Network", *International Journal of Computer Science and Engineering Survey*, Vol. 3(1), February 2012.
- [6] Parminder Kaur, Mrs. Mamta Katyari, "The Energy Efficient Hierarchical Routing Protocol for WSN", *International Journal of Advanced Research in Computer Science and Software Engineering*, IJARCSSE, Vol. 2(11), 2012.
- [7] Arboleda, Liliana MC, and Nidal Nasser, "Comparison of clustering algorithms and protocols for wireless sensor networks", *2006 IEEE Canadian Conference on Electrical and Computer Engineering*, CCECE, 2006.
- [8] Shashidhar Rao Gandham, Milind Dawande, Ravi Prakash, S. Venkatesan, "Energy Efficient Schemes for Wireless Sensor Network with Multiple Mobile base station", *Association for Computer Machinery*, ACM, 2002.
- [9] Heinzelman, Wendi B., Anantha P. Chandrakasan, and Hari Balakrishnan, "Application-specific protocol architecture for wireless microsensor networks", *Wireless Communications*, IEEE Transactions on 1.4, pp. 660-670, 2002.
- [10] Manjeshwar, Arati, and Dharma P. Agrawal, "APTEEN: A hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks", *International Parallel and Distributed Processing Symposium*, Vol. 2, IEEE Computer Society, 2002.
- [11] Manjeshwar, Arati, and Dharma P. Agrawal, "TEEN: a routing protocol for enhanced efficiency in wireless sensor networks", *International Parallel and Distributed Processing Symposium*, Vol. 3, IEEE Computer Society, 2001.
- [12] W.Heinzelman, A.Candrakasan, and H.Balakrishnan, "Energy-efficient Routing Protocols for Wireless Microsensor Networks", *33rd Hawaii Int. Conf. in Proc. System Sciences* (HICSS), 2000.

# Secured ATM Transaction Using Fingerprint and Voice Recognition Interface

Jaswinder Singh<sup>1</sup>, Jaswinder Kaur<sup>2</sup>  
*Department of Computer Engineering,  
Punjabi University, Patiala,  
Punjab, India,  
jaswindersinghmtech@gmail.com<sup>1</sup>  
jasvindernehal0@gmail.com<sup>2</sup>*

**Abstract-** This paper open a new vista for financial operation. The safe ATM system based on biometric identification and will bring its safe use for the customers. It will also bring the negative use of card based transactions. Now a day's new kind of ATM Machine fraud has come into high in the field of banking system. The life has become so fast that people prefer ATM transactions since it has many benefits. It relieves the people from long queue in the banks.

Moreover it opens our uses. People trust on ATM's. Traditional ATM systems are activated through card and PIN no. But now a day, ATM card and PIN no. does not give exact identity of customers. In order to reduce such kind of fraud and misuse. We can secure transaction using fingerprint and voice recognition. C# programming language is used to make the interface and MS-Access for the database.

**Keywords-** Fingerprint interface, ATM, Voice interface, Authentication, Security.

## I INTRODUCTION

The existing self banking system is very popular. The main aim of ATM machines is to ease banking financial operation at any time. But, in the modern ATM system has some defects. We have to carry ATM cards in our daily life. There are more possibilities to lose them or card could be stolen. ATM card and four digits PIN cannot verify the client's identity exactly. This article by me will expose how fingerprint and voice recognition technology be helpful for secure and easeful ATM operations. Fingerprint recognition technology gives the unique pattern with unique features. Voice recognition technology analyzes the unique information about genuine user. Fingerprint and voice recognition it becomes more secure for users and also bankers from theft and can make easy transactions.

The basic aim of this research paper is secure ATM transactions using fingerprint and voice recognition interface. The bank will collect the customer finger prints and any personal information related to customer while opening the accounts. When a customer goes in ATM center and places his finger on the fingerprint scanner. It will access area automatically and will generate a template and will compare to the stored template in the database. This authenticity of customer will be evaluating.

## II. RELATED WORK

Due to the gently importance and impact of biometric techniques in the security field, In this paper biometric data is used along with PIN number and if biometric data of user is matched with stored biometric data then user will allow to do the transactions. there are lots of work is done by many researchers by using following biometric techniques:

**TABLE I**  
**Types of biometric techniques in the security field**

<b>Title</b>	<b>Author</b>	<b>Year of publication</b>	<b>Method</b>	<b>Significance</b>	<b>Strengths</b>
Formation of Elliptic Curve Using Finger Print and Network Security	B. Thiruvaimalar Nathan	2006	Finger Print technique s	Minutiae co-ordinate points are extracted from Biometric templates and Elliptic curve algorithm is applied	Finger Print is used as biometric template
Efficient Finger Print Image Classification And Recognition using Neural Network Data Mining	K.Uma maheswari	2007	Finger Print technique s	Uses minutiae and combines with data mining techniques	Fast,, Reliable, Accuracy.
A Fast Fingerprint Classification Algorithm by Tracing Ridge flow Patterns	Neeta Nain	2008	Finger Print	Using Tracing Ridge-flow algorithm	Accuracy is perfect (98.75)
A New Approach of Fingerprint Recognition based on Neural Network	Behnaz Saropourian	2009	Finger Print	Based on View and groove pattern	works fine on binary images and gray scanned, Large accuracy
Fingerprint Identification Based on the Model of the Outer Layers of Polygon Subtraction	Nae Myo	2009	Finger Print	Uses model of multi-layers of convex polygon to implement fingerprint verification.	Extraction is based in a specific area in which the dominant brightness value of fingerprint ranges.
Security System Using Biometric Technology:Design and Implementation of Voice Recognition System(VRS)	Rozeha A. Rashid	2008	Voice	Recognize an individual's Unique voice characteristics	Hacking is much Complicated and possible only if u know the word
Microprocessor Based Voice Recognition System Realization	Nihat Ozturk	2010	Voice	Uses PIC18F452 microcontroller	Low cost and easily Applied prototype, Reliable, Accuracy

### **III. MOTIVATION**

The motivation of this work is to increase the security of ATM transactions. ATM Machine frauds have increased tremendously. So, we proposed the model for dual security. This can be done using fingerprint recognition and voice recognition.

### **IV. METHODOLOGY**

The security feature for enhancing the ATM transaction was designed using the client/server approach. There will be a link between the current customer (client) and bank database (server). Microsoft Access 2007 as a database software is used to create a database to store customer's information. The work is implemented using Visual Studio 2010 software tool, C# language is used to design the user interfaces and customer interaction with the ATM Machine.

## V. THE CHARACTERISTICS OF THE SYSTEM DESIGN

The primary functions based on fingerprint and voice recognition are shown as follows:

- a) Fingerprint recognition: in which enroll his/her fingerprint into the fingerprint device/reader adapter into the system. After which the fingerprint database compares the live sample provided by the customer with the template in the database.
- b) Matching and Database Verification:- After the feature extraction the users are authenticated by fingerprint recognition systems. On confirmation that the information provided is true, that customer is granted access to the ATM system.
- c) Client/Server authentication: System puts some question to client and client answer the question. The server check reply is correct or not with data base stored information.

## VI. DESIGN IMPLEMENTATION

By using C# language, the entire design was carried out on VB framework i.e. Visual basic Network. The application was made in six interfaces: login interface, enroll fingerprint interface, Transaction mode interface, Voice interface, PIN no. and transaction type selection interface.

- 1) **User login interface:** when the client enters his ATM card in the machine slot.. A dialogue box on the screen appears to show either the card number is valid or an invalid , If it is valid, next phase appears on the screen.
- 2) **Enroll fingerprint interface:** In this interface ,an image of a person's finger tips is taken and recognize its characteristic. After this, the system authenticates the samples stored in the database. If a match is confirmed. The authenticate customer is allows for further accessing.

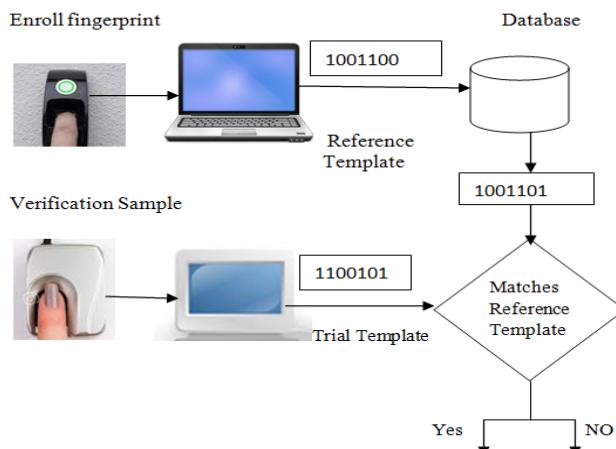


Fig 1 Fingerprint Authentication System

- 3 ) **Transaction mode interface:** In this interface system allows the customer to select the transaction phase means voice or PIN mode of transactions. If the user select PIN number mode. Then user follow step 4 and if the user select voice mode. Then user follow step 5.
- 4 ) **PIN mode interface:** In this phase user enter the PIN. If the customer enters an invalid PIN number, a dialogue box appears says an invalid PIN number . If customer enter valid no by checking the system works and the customer enters the transaction selection process.

5) **Voice mode interface:** In this phase. We use the Conversational voice recognition Technique in which verifies identity of the user by inquiring about the knowledge that is secret. That information is store bank database that time account is open and unlikely to be guessed by anyone.

6) **Transaction type selection:** In which user select the transaction type means balance inquiry, change pin no., withdrawal etc. If withdrawal type is chosen, print slip and balance comes on the screen.

## VII. EXPERIMENTAL RESULTS

Fingerprint recognition and voice recognition has an advantage the stability and reliability of biometric features for safe ATM operations. It is more safe, reliable and easy to apply for better functions. The functionality of the system will explain by the below steps.

Step 1: Insert ATM card in ATM machine slots.

Step 2: The card comes out and the message send to the Banker .

Step 3: Enroll the finger tips on the fingerprint scanner. The user fingerprint features already saved in the database. If authentication failure means try again. User allows trying maximum three times. If success means next step follows.



**Fig 2. Implementation design for the fingerprint verification process.**

Step 4: Choose user type of transaction. For PIN mode user step-5 follows. Voice mode user means Step-6 follows



**Fig 3. Implementation design for mode of transaction process.**

Step 5: PIN mode select user enter the 4 digit PIN. Correct PIN means step-4 follows false means try again. User allows trying maximum three times. After timeout step 4 follows.



**Fig 4. Design for PIN verification process.**

Step 6: If Voice mode type select user means. The server put some question based on the knowledge that is secret . That information is store bank database that time account is open.



**Fig 5. Design for voice verification process.**

Step 7: Then the transaction begins after completion of the authentication process.



**Fig 6. Implementation design for transactions .**

### VIII. CONCLUSIONS AND FUTURE SCOPE

Several existing methods are used for ATM security. In this proposed approach, we have been able to develop a fingerprint mechanism along with voice and PIN no. at a time. This paper has proposed a system to enhance the security features of the ATM for effective banking transaction for banks.

The proposed method overcomes then limitations that exists in other methods and provides a secured and safe environment that saves the hard earned money of the user. The prototype of the developed application has been found promising on the account of its sensitivity to the recognition of the customers finger print & voice recognition as contained in the database. There are three critical issues that need to be investigated further:

- 1) Embedding schemes for transforming one biometric representation into another.
- 2) Evaluation of the proposed system based on I.P address.
- 3) Methods to improve the security analysis by accurately modeling the biometric feature distributions.

### REFERENCES

- [1]. Akshata Patil, "Voice Enabled ATM Machine with IRIS Recognition for Authentication", Proceedings of the 3rd IRF International Conference, Goa, India, 2011.
- [2]. Mr Abhijeet S. Kale, Prof. Sunpreet Kaur Nanda, " Design of Highly Secured Automatic Teller Machine System by Using Aadhaar Card and Fingerprint", International Journal of Engineering Science Invention , ISSN: 2319 – 6726, Volume 3 Issue 5, PP.22-26, May 2014.
- [3]. Akilan, K.Gunasekaran, " Design of Two Tier Security ATM System with Multimodal Biometrics By Means of Fuzzy Logic", International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 1, February 2014
- [4]. Bhawna Negi 1 , Varun Sharma, "Fingerprint Recognition System", International Journal of Electronics and Computer Science Engineering , ISSN- 2277, 2011.
- [5]. Dileep Kumar, Yeonseung Ryu, "A Brief Introduction of Biometrics and Fingerprint Payment Technology", International Journal of Advanced Science and Technology, Vol. 4, March, 2009.
- [6]. Jaswinder singh , Jasvinder kaur, " Proposed Security System to Embed Fingerprinting and Voice Recognition for ATMs" IJARCSSE, Volume 5, Issue 5, May 2015.
- [7]. Khatmode Ranjit , Kulkarni Ramchandra, " ARM7 Based Smart ATM Access & Security System Using Fingerprint Recognition & GSM Technology" , .IJETAE, Volume 4, Issue 2, February 2014.
- [8]. Mahalakshmi , "Enhanced voice recognition to reduce fraudulence in ATM Machine", IJCNS, Volume 4. No 1. 2012.
- [9]. Mr. John Mashurano1, Mr. Wang liqiang , "ATM Systems Authentication Based On Fingerprint Using ARM Cortex-M3", International Journal of Engineering Research & Technology Vol. 2 Issue 3, ISSN: 2278-0181, March - 2013.
- [10]. Murugesh, Rishigesh. "Advanced biometric ATM machine with AES 256 and steganography implementation", Fourth International Conference on Advanced Computing (ICoAC), 2012.
- [11]. Myo, N."Fingerprint Identification Based on the Model of the Outer Layers of Polygon Subtraction", International Conference on Education Technology and Computer, JATIT '09, Page(s): 201 – 204, 2009.
- [12]. Nain, N. Bhadviya, B. Gautam, B. Kumar, Deepak, "A Fast Fingerprint Classification Algorithm by Tracing Ridge- Flow Patterns", IEEE International Conference on Signal Image Technology and Internet Based Systems, JATIT '08. Page(s): 235 – 238,2008.
- [13]. Nathan, B.T., Meenakumari, R., Usha,S., "Formation of Elliptic Curve Using Finger Print for Network Security", International Conference on Process Automation, Control and Computing (PACC) , Page(s): 1 – 5, 2011
- [14]. Ozturk, N., Unozkan,U., "Microprocessor based voice recognition system realization", 4th International Conference on Application of Information and Communication Technologies (AICT), JATIT, Page(s): 1 – 3,2010.
- [15]. Ibidapo, Akinyemi, Zacheous, Omobadegeun, M. Oyelami, "Towards Designing a Biometric Measure for Enhancing ATM Security in Nigeria E Banking System", International Journal of Electrical & Computer Sciences IJECS-IJENS Volume : 10 No: 06.
- [16]. Pennam Krishnamurthy, Mr. M. Maddhusudhan Redddy, "Implementation of ATM Security by Using Fingerprint recognition and GSM", International Journal of Electronics Communication and Computer Engineering ,Volume 3, Issue (1) NCRTCST, ISSN 2249 – 071X,2012.
- [17]. Rashid, Mahalin, N.H. Sarjari, Abdul Aziz, "Security system using biometric technology: Design and implementation of Voice Recognition System (VRS)", International Conference on Computer and Communication Engineering, Page(s): 898 – 902, 2008.

- [18]. S.S. Das, Debbarma , “Designing a Biometric Strategy fingerprint Measure for enhancing ATM Security in Indian e-banking system”, International Journal of Information and Communication Technology Research, Volume.1,no.5,pp.197-203,2011.
- [19]. Salil Prabhakar, Sharath Pankanti, Anil K. Jain, “Biometric Recognition: Security and Privacy Concerns”, IEEE Security & Privacy, Vol. 1, no.2, pp. 33-42, March-April 2003.
- [20]. Shimai Das, JhunuDebbarma, “Designing a Biometric Strategy (Fingerprint) Measure for Enhancing ATM Security in Indian E-Banking System”, International Journal of Information and Communication Technology Research , Volume 1 No. 5, September 2011.
- [21]. S. Pravintha and K. Umamaheswari , "Multimodal Biometrics for Improving Automatic Teller Machine Security", International Journal of Advances in Image Processing ,Volume 1, December, 2011.
- [22]. Santhi and Kumar, “Novel Hybrid Technology in ATM Security Using Biometrics”, Journal of Theoretical and Applied Information Technology(JATIT) , 2012.
- [23]. Saropourian, B., “A new approach of finger-print recognition based on neural network”, 2nd IEEE International Conference on Computer Science and Information Technology, JATIT 2009, Page(s): 158 – 161, 2009.
- [24]. Umamaheswari, K., Sumathi, S., Sivanandam, S.N., Anburajan, K.K.N., "Efficient Finger Print Image Classification and Recognition using Neural Network Data Mining" Signal Processing, International Conference on Communications and Networking, 2007. JATIT '07,Page(s): 426 – 432,2007.
- [25]. V.Padmapriya, S.Prakasam, “ Enhancing ATM Security using Fingerprint and GSM Technology”, International Journal of Computer Applications (0975 – 8887) Volume 80 – No 16, October 2013.
- [26]. Yun Yang., "ATM terminal design is based on fingerprint recognition", 2nd International Conference on Computer Engineering and Technology, 04/2010.

# A REVIEW ON VIDEO FUSION USING DIFFERENT TECHNIQUES

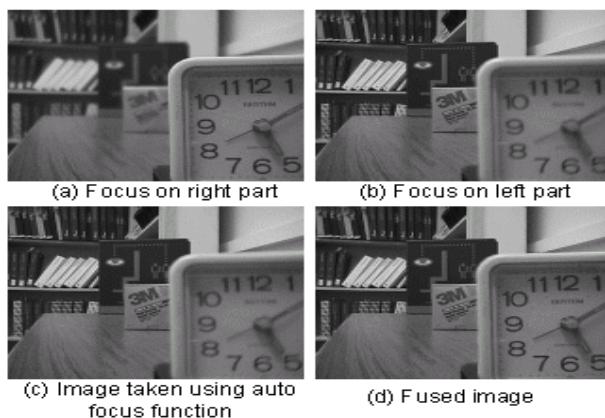
Gurmeet Singh  
Student of M.tech (C.S.E.)  
YCOE ,Talwandi Sabo  
[chauhangurmeet49@gmail.com](mailto:chauhangurmeet49@gmail.com)

Ashok Bathla  
Assistant Professor (C.S.E.)  
YCOE ,Talwandi Sabo  
[ashokashok81@gmail.com](mailto:ashokashok81@gmail.com)

**Abstract:** Video learning combination should in the meantime take under thought every fleeting and spatial measurements, thus a remarkable spatiotemporal video-combination principle bolstered movement remuneration inside the wavelet-change area is anticipated amid this study. There is human visual image and objective analysis criteria connected issues once the fusion of 2 videos occurred that I even have studied within the literature survey. The objective in video fusion is to cut back uncertainty associate degress minimize redundancy within the output whereas maximising relevant info explicit to an application or task. To enhance the videos while not destroying any issue of the videos like Hue , Saturation and Intensity and compare results with some existing techniques.

## I. INTRODUCTION

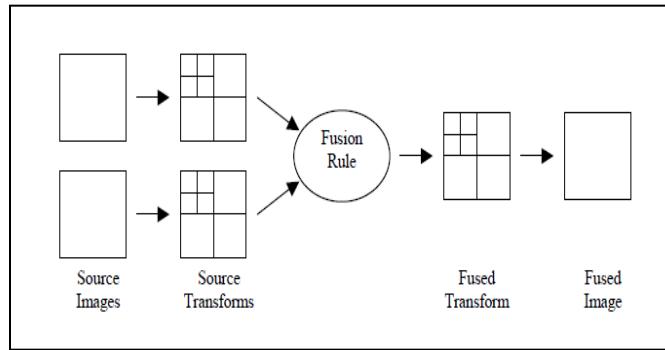
Video successions caught by one gadget regularly can't particular all the information in an exceptionally scene. to get the whole information of a scene, various recordings range unit caught in the meantime for the substance of a comparative scene. To utilize the learning from different recordings adequately and with proficiency, the substance should be joined into a video grouping. Video combination methodologies will adequately consolidate corresponding data from various supply recordings into an amalgamate video that contains extra verdant substance a couple scene than any of the individual video. Up to now, steady with studies inside the writing, a spread of video-combination methodologies are produced and used in a few certifiable fields like protection police examination, medicinal imaging, remote detecting and pc vision. Be that as it may, most existing combination approaches change static recordings. Compelling video-succession combination techniques zone unit still uncommon. In certifiable applications, video process has turned into an empowering space, drawing in wide consideration contrasted and propelled video-preparing fields. Among them, video combination might be an appallingly inventive innovation. For video or element video combination, beyond any doubt bland necessities should be pondered. The combination strategy should protect the greatest sum accommodating data as feasible inside the composite video grouping and may not present any ancient rarities or irregularities. also, an exceptional condition for video combination is that the amalgamate video should be transiently steady and per the info recordings. besides, the video-combination calculations should have high system intensity and low house necessities.



**Figure 1. Example of Fused Video**

## VIDEO FUSION

Figure :2 below demonstrates the general Video fusion process. This figure illustrates the fusion process for two source videos, but the process can be implemented for combining multiple source videos.



**Figure -2 General video fusion processes**

As a rule, the reflection space of video learning must be redesigned to the recurrence area of learning of data as an aftereffect of the recurrence style of information gives extra adaptability and common sense than the deliberation information for video process. The recurrence space of information is gotten utilizing a disintegration subject like separate Fourier redesign, riffle rebuild, and so forth riffle rebuild is comprehended to be higher than Fourier redesign. This video combination framework can utilize a Haar riffle to break down supply recordings and create amalgamate redesign, amid a transient legitimization of video combination technique, supply recordings ar spoiled to supply rebuilds furthermore the amalgamate change is made with supply changes upheld combination principle. the main combination tenet is selecting the one with bigger extent; The amalgamate video comprises through the amalgamate rebuild.

## II. LITERATURE SURVEY

**Liang Xu et al. (2015)** [7] have concentrated on Video learning combination should in the meantime take under thought every fleeting and spatial measurements, thus a remarkable spatiotemporal video-combination principle bolstered movement remuneration inside the wavelet-change area is anticipated amid this study. The combination system joins movement pay furthermore the wave rebuild, along these lines making full utilization of spatial geometric information and between edge transient information of info recordings. The anticipated strategy enhances the worldly solidness and consistency of the blended video contrasted with option existing individual casing based combination ways. The standard first decays the picture outlines that are prepared by Associate in Nursing optic stream movement pay approach then builds up a spatiotemporal vitality based combination guideline to consolidation info recordings.

**S. A. Quadri et al. (2013)** [1] have considered the information combination goes for synergistic use information and learning from very surprising sources to help inside the general comprehension of an improvement. inside the space of remote detecting, wherever recordings square measure noninheritable by various supplies or by a comparative source in numerous procurement connections, the information made possible by entirely unexpected sources square measure corresponding to each option, right combination of the data will bring higher and predictable elucidation of the scene. The paper presents use of Kalman channel at pixel-level combination. The information record gathered from gas watching Instrument (OMI) on NASA's Aura satellite is subjected to the anticipated standard. The execution of the tenet is assessed by few surely understood video quality measurements.

**Rohan Ashok Mandhare et al. (2013)** [2] have anticipated the Remote detecting recordings combination can't exclusively enhance the spatial determination for the principal multispectral video, however should conjointly safeguard the ghostly information to a specific degree. Shading change principally based video combination ways are authorized in various papers and this ways demonstrates shrewd ghastly maintenance. This paper means to execute pixel-level video combination upheld numerical and wave redesign video combination ways and intention their ability to upgrade spatial and ghostly information. For this reason entirely unexpected ways like Averaging procedure, expanding method, Brovey system, and DWT strategy square measure implemented. Execution of this ways is assessed with the help of evaluation parameters such entropy, fluctuation, RMSE and PSNR. Test results demonstrates that recordings converge by expanding and wave essentially based procedure indicated higher spatial determination and higher phantom choices than the main video.

**Mirajkar Pradnya P et al. (2013)** [3] amid this paper, we have a tendency to propose Associate in Nursing video combination approach bolstered Stationary wave rebuild (SWT). Stationary wave redesign (SWT) is initially connected with the main video to impel the sting video information each in level one and level a couple of. Next, every edge recordings square measure blended to affect a whole edge video abuse spatial Frequency movement, that is contrasted and some simple combination ways.

**VPS Naidu et al. (2013)** [5] a special video combination strategy misuse separate trigonometric capacity redesign (DCT) essentially based laplacian pyramid is gave and concentrated on. it's ended that combination with more elevated amount of pyramid gives higher combination quality. The execution time is corresponding to the amount of pyramid levels used in the combination strategy. this framework are frequently utilized for combination of correlative information recordings in like manner as multi model video combination. The anticipated guideline is amazingly simple, easy to actualize and will be utilized for ongoing applications. MATLAB code has been accommodated quick execution and approval of the anticipated principle.

**Changtao He et al. (2013)** [6] Multimodal medicinal video combination, as an intense apparatus for differed clinical analyses, has created with the presence of fluctuated imaging modalities inside the field of medications. This paper proposes a special video combination way to deal with viably understand a concurrent drawback of spatial qualities and otherworldly information inside the mixed video. As we tend to all perceive, the power tint immersion (IHS) rebuild and retina-enlivened model (RIM) combination system will save a great deal of spatial choices and a considerable measure of ghostly information substance, severally. Nonetheless, chief part investigation (PCA) guideline will separate fundamental element to constrict repetition. The anticipated principle incorporates their gifts and effectively enhances mixed video quality to dodge shading twists. The trial exhibits that the anticipated standard beats dynamic combination approaches like PCA, Brovey, RIM, separate wave rebuild (DWT) in light-weight of visual effect and quantitative examination.

### III. PROBLEM FORMULATION

The problems undertaken for the dissertation are given below:

- There is drawback associated with spectral distortion of the videos seems, which implies that the variation on hue before and once the fusion method has appeared.
- The color distortion drawback within the fusion method.
- There is drawback of special characteristics and spectral info within the amalgamate video.
- There may be a color distortion once the fusion is appeared within the color videos.
- In the medical videos ambiguity and redundancy issues.
- There is human visual image and objective analysis criteria connected issues once the fusion of 2 videos occurred that I even have studied within the literature survey.
- The Hue , saturation and also the intensity of the colour videos established as a result of fusion and as a result of noises.

### IV. METHODOLOGY

This treatise is to implement the fusion of various videos .The implementation is performed exploitation Matlab and C artificial language. Matlab provides the convenient facilities to govern videos, and C artificial language allows quick execution. Matlab could be a superior language for technical computing. It integrates computation, visualisation, and programming in an easy-to-use surroundings wherever issues and solutions ar expressed in acquainted notational system. Matlab is that the tool of selection for high-productivity analysis, development, and analysis like science and computation, algorithmic program development, knowledge acquisition, modeling, simulation, knowledge analysis and visualisation, scientific and engineering graphics, etc. Matlab is AN interactive system whose basic knowledge component is AN array that doesn't need orienting . this enables several technical computing issues to be solved , particularly those with matrix and vector formulations. within the video fusion system, the subsequent system is employed to style the video fusion; within the video fusion the subsequent steps ar followed:

#### Steps:

- Take input videos of same size and of same scene or object taken from completely different completely different sensors like visible and below red videos or videos having different focus.
- If the input videos ar color, separate their RGB planes to perform second transforms.
- Apply one in every of the various video fusion techniques.
- Fuse the input video elements by taking any of the pixels merging technique.
- ensuing consolidated remodel elements ar reborn to video exploitation inverse remodel.

## V. CONCLUSION

In this work I have studied different papers and each paper have different problems. There is problems like the color distortion and the contents loss of the videos. In the future work I will use DWT and SWT and PCA to implement the video fusion and get the maximum results.

## REFERENCES

- [1] S. A. Quadri And Othman Sidek " Pixel-Level Video Fusion misuse Kalman equation " International Journal Of Signal procedure, Video process And Pattern Recognition Vol. 6, No. 2, April, 2013.
- [2] Rohan Ashok Mandhare1, Pragati Upadhyay2,Sudha Gupta "Pixel-Level Video Fusion abuse Brovey Transforme And wave Transform" International Journal Of Advanced investigation In Electrical, physical science And Instrumentation Engineering Vol. 2, Issue 6, June 2013.
- [3] Mirajkar Pradnya P., 2sachin D. Ruikar" Video Fusion bolstered Stationary wave rebuild "Mirajkar, Et Al, International Journal Of Advanced Engineering investigation And Studies E-Issn2249-8974.
- [4] Vps Naidu, Bindu Elias" a totally novel Video Fusion Technique abuse Dct based for the most part Laplacian Pyramid" International Journal Of shrewd Engineering And Sciences (Ijes) Issn: 2319-9598, Volume-1, Issue-2
- [5] Changtao He\*1, Guiqun Cao2, Fangnian Langan practical Fusion Approach For Multispectral And Panchromatic Medical Imaging" prescription Engineering investigation March. 2013, Vol. 2 Iss. 1, Pp. 30-36.
- [6] J. H. Jang And J. B. Ra "Pseudo-Color Video Fusion bolstered Intensity-Hue-Saturation Color range "Procedures Of Ieee International Conference On Multisensor Fusion And Integration For Intelligent Systems national capital, Korea, August twenty - twenty two, 2008.
- [7] Wirat Rattanapitak And Somkait Udomhunsakul "Similar strength Of Color Models For Multi-Focus Color Video Fusion "Continuing Of The International Multi Conference In Engineering And researcher 2010 ,imecs-2010.
- [8] H. Li, B.S. Manjunath And S.K. Mitra, "Multisensor Video Fusion misuse wave Transform", Graph. Models Video strategy, 57(3), Pp.235-245, 1995.
- [9] Vps Naidu And J.R. Raol, "Pixel-Level Video Fusion misuse Wavelets And Principal component Analysis – A Comparative Analysis" Defense Science Journal, Vol.58, No.3, Pp.338-352, May 2008
- [10] A. Toet, L.J. Van Ruyven And J.M. Valetton, "Combining Thermal And Visual Videos By A qualification Pyramid", Opt. Eng. 28(7), Pp.789-792, 1989.
- [11] Rick S. Blum, "Hearty Video Fusion utilizing a connected math Signal procedure Approach", Video Fusion, 6, Pp.119-128, 2005.
- [12] Vps Naidu, "Discrete cos Transform-Based Video Fusion", Special Issue On Mobile Intelligent Autonomous System, Defense Science Journal, Vol. 60, No.1, Pp.48-54, Jan. 2010.
- [13] N. Ahmed, T. Natarajan And K.R.Rao, "Discrete cos Transform", Ieee Trans. On Computers, Vol.32, Pp.90-93, 1974.
- [14] Shutao Li, James T. Kwok And Yaonan Wang, "Mix Of Videos With various Focuses misuse The spatial Frequency", information Fusion, 2(3), Pp.167-176, 2001.

# A Mechanism to Enhance Lifetime of NEW LEACH Protocol in WSN

Ramandeep Kaur<sup>1</sup>, Er. Navroz Kaur<sup>2</sup>

<sup>1</sup>Research Scholar, <sup>2</sup>Assistant Professor

<sup>1</sup>raman\_dhillon91@yahoo.com, <sup>2</sup>kahlon. navroz3@gmail.com

Department of Computer Engineering, Punjabi University, Patiala-147002 (India)

**ABSTRACT-** Due to the extensive range of applications the use of wireless sensor networks (WSNs) in past few years have increased a lot and it has become a hot research area now a days. One of the important issues in wireless sensor network is the intrinsic limited battery power in network sensor nodes. Since most of the energy is consumed by the transmission and reception, energy efficient routing protocol is required to enhance the lifetime of WSN. In this paper we presents the review of existing hierarchical routing protocol NEW LEACH and proposed a new approach that is better and more energy efficient than the existed approaches. In the proposed work, two enhancements are made. These enhancements are efficient cluster head replacement technique by using dynamic threshold value and the concept of upper threshold for sending data. Both NEW LEACH and Proposed protocol are simulated in MATLAB. The result shows that our proposed algorithm performs better than the NEW LEACH in terms of network lifetime

**Keywords**— WSN; Hierarchical Routing; NEW LEACH Protocol; Proposed Method; Dynamic Threshold.

## 1. INTRODUCTION

A WSN is a set of sensors, which are deployed in a sensor field to observe specific characterization of the environment [1]. To evaluate that characteristic and collect the data related to the phenomena. Basically Sensors are small devices with restricted resources such as, little computing capability, low memory, limited battery power ,very low data rates, low bandwidth processing and variable link quality [20]. Although these constraints, when sensors are deployed in great numbers, they can give us with a very real image of the area being sensed. WSNs can provide a region coverage that was not probable with other wired and wireless networks. Sensor networks are generally unattended and need to be fault-tolerant so as necessitate for maintenance is minimized. They can be deployed in various environments which can be permanently attended or can be left unattended once deployed. Wireless sensor networks typically have power constraints [1].

### *1. 2 Components of WSN*

There are five components of wireless sensor network which are sensor node, cluster, cluster head, base station and end user which [19]are as follows:-

#### *1) Sensor Node*

A sensor node is the core component of a WSN. Sensor nodes can take on multiple roles in a network, such as simple sensing; data storage; routing; and data processing [20].

2) *Clusters*

Clusters are the organizational unit for WSNs. The dense nature of these networks requires the need for them to be broken down into clusters to simplify tasks such a communication.

3) *Cluster heads*

CHs are the organization leader of a cluster. The function of the CH is to perform common functions for all sensor nodes in the cluster, as aggregation the data before sending it to the BS.

4) *Base Station*

The BS is at the upper level of the hierarchical WSN. It provides the communication link between the sensor network and the end-user [9].

5) *End User*

The data in a sensor network can be used for a wide-range of applications. Therefore, a particular application may make use of the network data over the internet, using a PDA, or even a desktop computer.

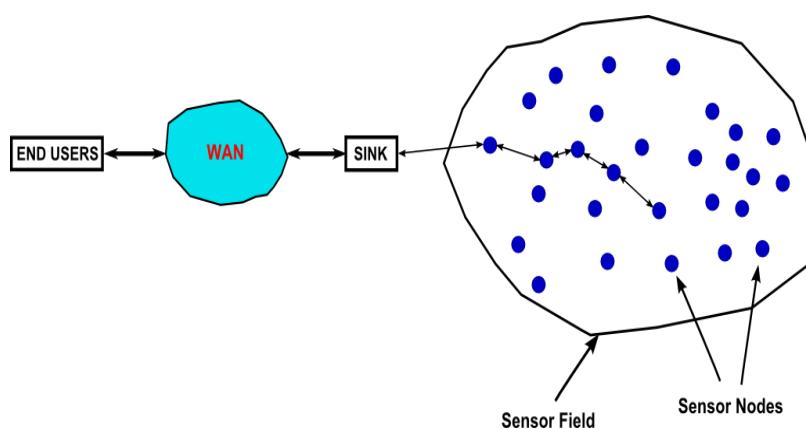


Fig. 1 Architecture of WSN

## 2. RELATED WORK

Our proposed algorithm is based on an energy-efficient hierarchical routing protocol LEACH [23]. The main aim of hierarchical routing is to efficiently maintain the energy consumption of sensor nodes by involving them in multihop communication within a particular cluster and by performing data aggregation and fusion in order to decrease the number of transmitted messages to the sink. LEACH is one of the first hierarchical routing approaches for sensors networks. The idea proposed in LEACH has been an inspiration for many hierarchical routing protocols. The major problem in wireless sensor network is the battery consumption. To increase lifetime of the sensor networks, technique of clustering is to be applied. Traditionally in LEACH protocol the cluster head selection was done on the basis of the probability. So to change cluster head after each round was a problem . After this new approach was proposed in which cluster head selection is done on the basis of static threshold value. If the residual energy of the CH is greater than the threshold then the CH will not be changed. If cluster head has less energy than the static threshold then it will be replaced according to the LEACH. But in case all nodes having energy less than the static threshold value, it creates the problem [2]. Because at time random election of CH will be done, sometimes that CH will be selected which has not enough energy to perform communication. So

there is loss of data when CH dies in the middle of the any round. Hence it will reduce efficiency of the network. So in this way the static threshold was a problem. In such condition there is a need of the dynamic threshold which should be based on the residual energy of the whole network. So considering all these problems we propose a new approach of clustering that is better and more efficient than traditional approach.

### 3. PROPOSED WORK

We have proposed an updated version of NEW LEACH. To make the network more energy efficient we have considered heterogeneous network with nodes having two different amounts of energy - normal nodes having a lesser amount of energy will be used as the cluster members and advanced node possessing double energy than that of normal nodes will be elected as CH in initial rounds. The CHs will aggregate the received data and transmit it to the BS. Inside this protocol in a very first round cluster head selection was done on the basis of the probability from a given set of advanced nodes. We have estimated to be 5% of the total number of nodes as cluster heads. During reclustering residual energy of a CH will be checked. If the residual energy of the CH is greater than or equal the predefined threshold then it will be continue as cluster head in next round also. The predefined threshold defines the minimum energy level which must be possessed by a CH to communicate with the BS directly. Here for simplicity we have considered initially threshold which is equal to be half of the initial energy of the normal nodes. Every time when the residual energy of the CH becomes less than predefined threshold than a new dynamic threshold value will be calculated that is equal to the half of the mean residual energy of the whole network. So in this way every time that CH will be chosen this has maximum energy in that particular cluster. This method is not only use the remaining energy of cluster head but also minimize the overhead associated with the CH formation in every round. Basically, in proposed work, two modifications/ enhancements are made. These enhancements are

- 1) Efficient cluster head replacement technique.
- 2) Introducing the concept of upper threshold.

In our proposed protocol, the CH broadcasts the following threshold value to its cluster members at every cluster setup phase in the protocol. By applying upper threshold concept in proposed protocol make the protocol reactive in nature.

*Upper threshold (UT):* This is a threshold value for the sensed attribute. It is the absolute value of the attribute through which the nodes that sensing this value must transmit to its transmitter and report to its cluster head [26].

The nodes sense their surroundings continuously. When the sensed data reaches its upper threshold value the node transmits the sensed value. The sensed data value is stored in a variable known as *sensed value* (SV). The node will transmit the data in current round only when, the sensed data is greater than the upper threshold, thus, the upper threshold tries to reduce the data communication by allowing the nodes to send only when the sensed attribute is in the range of interest [1]. Reactive nature of this routing algorithm is not only minimizing the routing overhead, it also gives better network life time [7]. Different steps involved in the proposed method are followings:

#### 3. 1 Set-Up Phase

1. In the first round all the advanced nodes generate a random number between 0 and 1 and if it is less than the threshold T (n) then the node is selected as a CH. Go to step 3.

$$T(n) = \frac{p}{1 - p \left( r \bmod \frac{1}{p} \right)} \text{ if } n \in G \\ (1)$$

$$T(n) = 0 ; \quad \text{otherwise (1)}$$

Here

n= number of nodes r =the current round

p= probability of a node to become CH;

G = refers to the set of nodes that have not been elected as CH in the last rounds.

2. In every other round the residual energy of CH checked. If it is greater than a predefined threshold then it will continue as the CH in the next round also. So go to steady-state otherwise a new dynamic threshold value will be calculated. Thus in this way every time that CH will be chosen which has maximum energy in that particular cluster.
3. CHs broadcast its advertisement to remaining nodes.
4. Clusters are formed depending on the signal strength a normal node receives from different CHs.
5. The normal nodes send a join message to the corresponding cluster heads that in turn create TDMA schedule for data transmission and broadcast it to the members.

### 3.2 Steady-State Phase

1. All the cluster members will send data to their related cluster-heads in their allotted time slot.
2. CHs aggregate the received data and transmit it to the BS.
3. Once all the CHs finish the control returns to steady phase again.

## 4. PERFORMANCE EVALUATION

We have simulated our proposed protocol in MATLAB 2012. To evaluate the performance of our proposed technique, we compared it with the existing protocol NEW LEACH. Simulation parameters are shown in table1.

TABLE I. NETWORK PARAMETERS FOR SIMULATION

Network Parameter	Value
Network Size	100m*100m
Number of Nodes (N)	100
Energy of normal nodes ( $E_0$ )	0. 5J
Energy of advanced nodes ( $E_1$ )	1. 0J
Amplification Energy when $d \geq d_o$ ( $E_{fs}$ )	10pJ/bit <sup>2</sup>
Amplification energy when $d \leq d_o$ ( $E_{mp}$ )	0. 0013pJ/bit
Packet Size	4000bits
Upper Threshold ( $U_T$ )	100
Number of rounds( $r_{amx}$ )	3000
Transmitter Electronics (ETX)	50nJ/bit
Receiver Electronics (ERX)	50nJ/bit
Data Aggregation Energy(EDA)	5nJ/bit

The network consists of hundred nodes that are randomly deployed in the given network of  $100 \times 100 \text{ m}^2$ . We have assumed five advanced nodes and the remaining nodes are normal nodes. If a node's energy is less than zero, we describe that as a dead node. As soon as all nodes in the network are dead, we define it as network failure. Base station is assumed to be fixed at the centre location in our WSNs. All the sensor nodes are permanent once they are deployed on their locations. The Fig. 2 below shows how the nodes are distributed randomly.

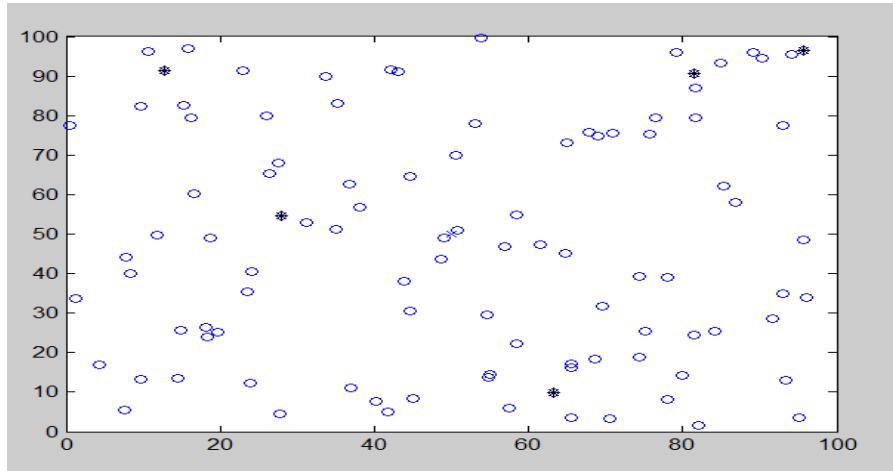


Fig. 2 Random Distributions of Nodes

#### 4. 1 Result

The simulated results depicted in Fig. 3 and Fig. 4 compare network life time of NEW LEACH, and proposed protocol by showing number of dead and alive nodes respectively. In the proposed method the first node dies later than NEW LEACH because it uses the concept of upper threshold to send data to CH and efficient cluster head replacement technique is followed by calculating dynamic threshold every time when CH energy goes below the predefined threshold.

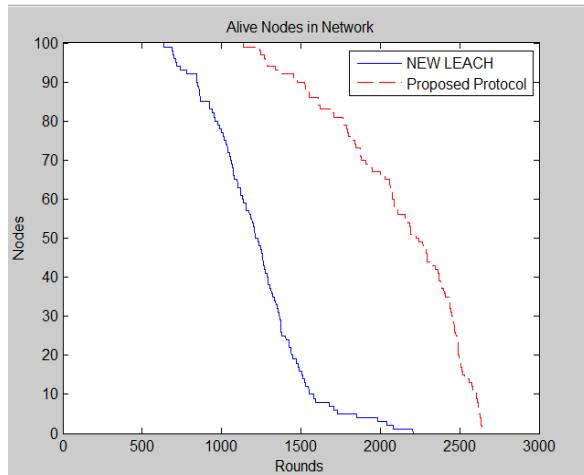


Fig. 3 Alive nodes in NEW LEACH and proposed protocol

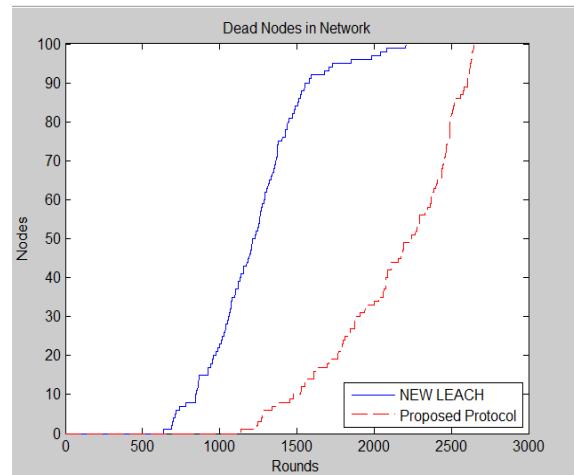


Fig. 4 Dead nodes in NEW LEACH and proposed protocol

The node will transmit the data in current round only when, the sensed data is greater than the upper threshold, thus, the upper threshold tries to reduce the number of transmission by sending only when the sensed data is in the range of interest. The proposed method increases the lifetime of WSN by 20.03% as compared to NEW LEACH shown in Table II. From the simulation results shown above, an improvement table is drawn below which shows lifetime improvement on the basis of first node dead and all nodes dead. Table II shows the rounds when nodes start dying and all the nodes are dead in NEW LEACH, and proposed protocol.

TABLE II. COMPARISON OF THE PROTOCOLS

Protocol	Rounds when nodes start dying	Rounds when all nodes are dead
New Leach	635	2206
Proposed Protocol	1138	2648

As we have already mentioned that in our Proposed Protocol the nodes started to die later due to proper use of upper threshold and CH energy. Hence the overall lifetime of network is increased by 20.03% than NEW LEACH this can be explained mathematically as below:

$$((\text{Last Round of Proposed Protocol} - \text{Last Round of NEW LEACH}) / \text{Last Round of NEW LEACH}) * 100 = ((2648 - 2206) / 2206) * 100 \\ = 20.03\%$$

So as 20.03% improvement over NEW LEACH in overall lifetime of WSN is obtained by our proposed Protocol. Besides network life time the other parameter which we have considered is the throughput of the routing protocols. From the simulated results in Fig. 5 and Fig. 6 shows throughput in NEW LEACH and Proposed Method.

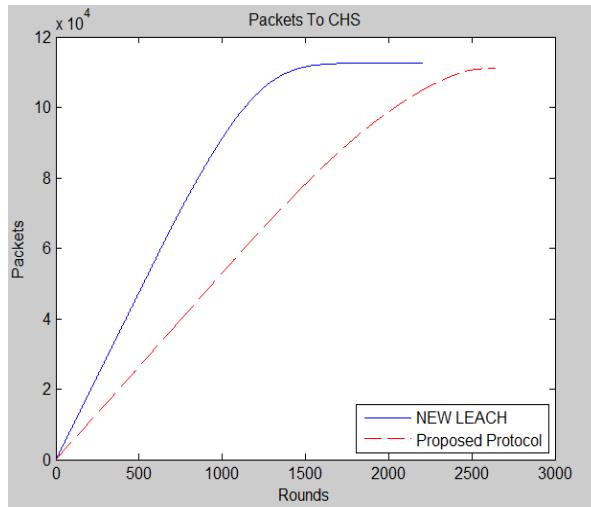


Fig. 5 Packets to CH in NEW LEACH and Proposed Protocol

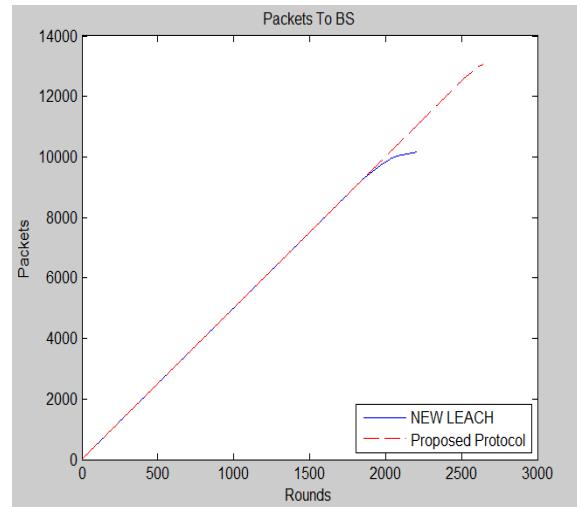


Fig. 6 Packets to CH in NEW LEACH and Proposed Protocol

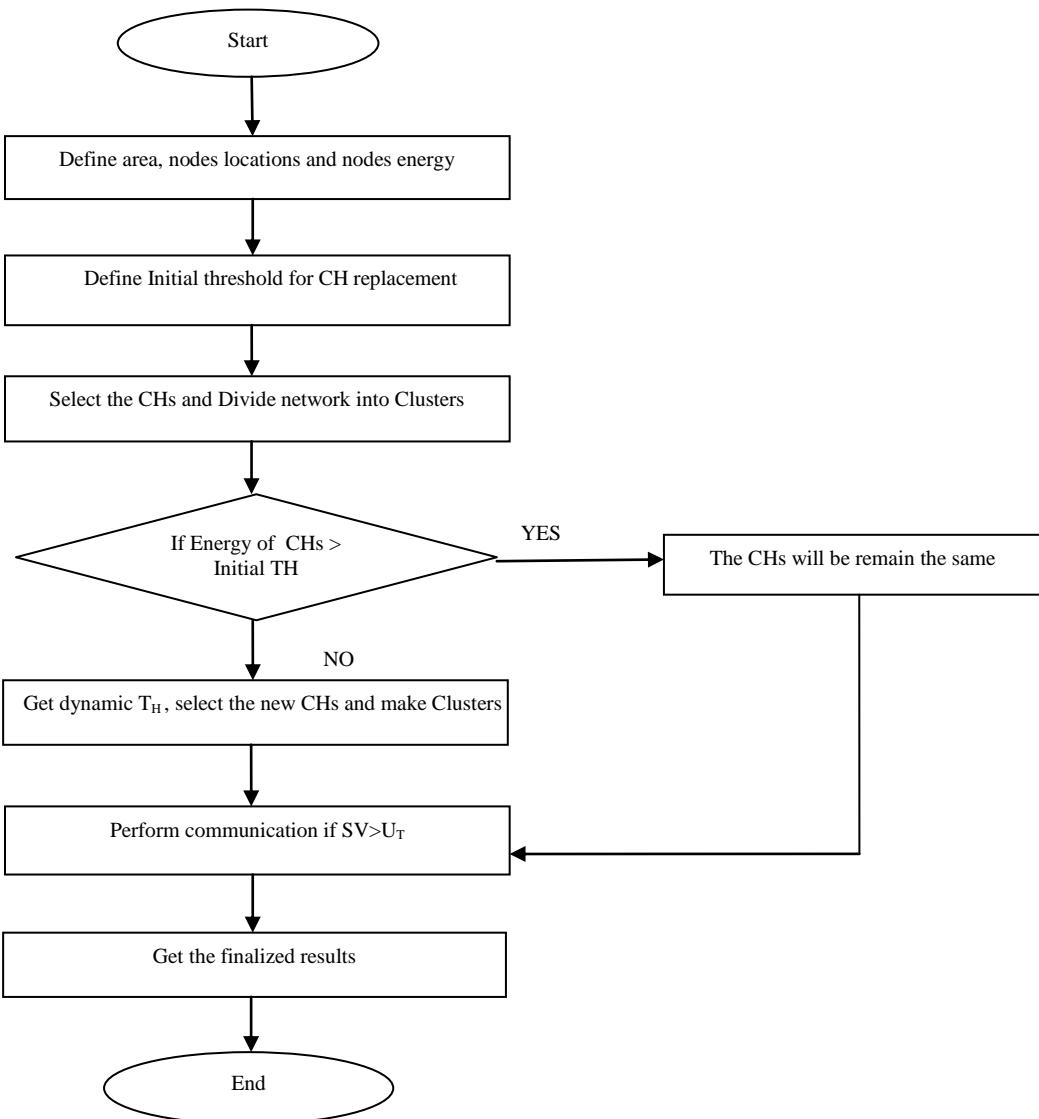


Fig. 7 Flowchart for the working of proposed Protocol

## 5. CONCLUSION AND FUTURE SCOPE

In this research work we discussed the working of NEW LEACH and Proposed Protocol. We have compared these two routing protocols and concluded that the performance of Proposed Protocol is better than NEW LEACH. We compared both the algorithms with the help of different parameters like lifetime and throughput by using these parameters Proposed Protocol better than NEW LEACH. This improvement is achieved by using the concept of upper threshold to send data to CH and efficient cluster head replacement technique is followed by calculating dynamic threshold every time when CH energy goes below the predefined threshold. The node will transmit the data in current round only when, the sensed data is greater than the upper threshold, thus, the upper threshold tries to reduce the number of transmission by sending only when the sensed data value is in the range of interest.

The future work includes the concept of mobility can be incorporated to our Protocol and further deployment of nodes based on the locations can also be introduced.

REFERENCES

- [1] Sandeep Wraich, Jasdeep Kauri, "Comparative Analysis of Various Energy Efficient Protocols for Wireless Sensor Networks". International Journal of Computer Applications(IJCA) Volume 115, April 2015, pp. 25-31
- [2] Jyoti Singh, Bhanu Pratap Singh, Subhadra Shaw, "A New LEACH-based Routing Protocol for Energy Optimization in Wireless Sensor Network", Proceeding of IEEE International Conference on Computer and Communication Technology, 2014, pp. 181-186.
- [3] Mohammad Shurman, Noor Awad , Mamoun F. AI-Mistarihi, and Khalid A. Darabkh, "LEACH Enhancements for Wireless Sensor Networks Based on Energy Model ", IEEE, 2014, pp. 1-4.
- [4] Sapna Gambhir, Nida Fatima. "Op-LEACH: An Optimized LEACH Method for busty traffic in WSNs", Proceeding of fourth International IEEE Conference on Advanced Computing & Communication Technologies, 2014, pp. 222-229.
- [5] Salim EL KHEDIRia, Nejah NASRI, Anne WEI, Abdennaceur KACHOURi, "A New Approach for Clustering in Wireless Sensors Networks Based on LEACH ", Science Direct, 2014, pp. 1180-1185.
- [6] Padmalaya Nayak , Pallavi Shree, "Comparison of Routing Protocols in WSN using NetSim Simulator: LEACH Vs LEACH-C ", International Journal of Computer Applications (IJCA) Volume 106, 2014, pp. 1-6.
- [7] D. Mahmood, N. Javaid, S. Mahmood, S. Qureshi, A. M. Memon, T. Zaman, " MODLEACH: A Variant of LEACH for WSNs", Eighth IEEE International Conference on Broadband, Wireless Computing, Communication and Applications, 2013, pp. 159-163.
- [8] Anjali Bharti, Kanika Sharma, " Comparative Study of Clustering based Routing Protocols for Wireless Sensor Network", International Journal of Computer Applications(IJCA) Volume 66, 2013, pp. 9-15.
- [9] Pooja A. Vaishnav and Naren V. Tada," A New Approach to Routing Mechanism in Wireless Sensor Network Environment ", Nirma University IEEE International Conference on Engineering , 2013, pp. 1-5.
- [10] Saewoom Lee, Youngtae Noh, and Kiseon Kim, "Key Schemes for Security Enhanced TEEN Routing Protocol in Wireless Sensor Networks", International Journal of Distributed Sensor Networks, 2013, pp. 1-8.
- [11] Jia Xu,Ning Jin, Xizhong Lou,Ting Peng,Qian Zhou and Yanmin Chen, "Improvement of LEACH protocol for WSN ", International IEEE Conference on Fuzzy Systems and Knowledge Discovery, 2012, pp. 2174-2177.
- [12] J. Gnanambiga,Dr. N. Rengarajan,K. Anbukkarasi, "Leach and Its Descendant Protocols: A Survey", IJCSI Volume 01, 2012, pp. 15-21.
- [13] Sandeep Sharma, Sapna choudhary, "Heterogeneous Multi-hop LEACH routing protocol", IEEE, 2012, pp. 1181-1187.
- [14] Vipin Pal , Girdhari Singh, R P Yadav, " Energy Efficient Cluster Head Selection Scheme: AChange in Node Death Scenario of LEACH for Surveillance Wireless Sensor Networks", 2nd IEEE International Conference on Parallel, Distributed and Grid Computing, 2012, pp. 865-869.
- [15] Kiran Maraiya, Kamal Kant, Nitin Gupta, "Application based Study on Wireless Sensor Network", International Journal of Computer Applications(IJCA) Volume 15, May 2011, pp. 9-15.
- [16] Rajashree. V. Biradar, Dr. S. R. Sawant, Dr. R. R. Mudholkar , Dr. V. C. Patil, "Multihop Routing In Self-Organizing Wireless Sensor Networks", IJCSI Volume 8, 2011.
- [17] Vinay Kumar, Sanjeev Jain2 and Sudarshan Tiwari, "Energy Efficient Clustering Algorithms in Wireless Sensor Networks: A Survey", IJCSI Volume 8, 2011, pp. 259-268.
- [18] M. Bani Yassein, A. Al-zou'bi, Y. Khamayseh, W. Mardini, "Improvement on LEACH Protocol of Wireless Sensor Network (VLEACH)", International Journal of Digital Content Technology and its Applications , 2009.
- [19] Fan Xiangning, Song Yulin, "Improvement on LEACH Protocol of Wireless Sensor Network", International IEEE Conference on Sensor Technologies and Applications, 2007, pp. 260-264.
- [20] Liliana M. Arboleda C. and Nidal Nasser, "Comparison of clustering algorithms and protocols for wireless sensor networks", IEEE, 2006, pp. 1787-1792.
- [21] Mhatre, Vivek, and Catherine Rosenberg, "Homogeneous vs heterogeneous clustered sensor networks: a comparative study ", IEEE, 2004, pp. 3646-3651.
- [22] Seema Bandyopadhyay and Edward J. Coyle, " An Energy Efficient Hierarchical Clustering Algorithm for Wireless Sensor Networks ", IEEE, 2003.

- [23] Wendi B. Heinzelman, Anantha P. Chandrakasan, and Hari Balakrishnan, "An Application-Specific Protocol Architecture for Wireless Microsensor Networks", IEEE, 2002, pp. 660-670.
- [24] Arati Manjeshwar and Dharma P. Agrawaly, " APTEEN: A Hybrid Protocol for Efficient Routing and Comprehensive Information Retrieval in Wireless Sensor Networks ", IEEE International Parallel and Distributed Processing Symposium (IPDPS. 02), 2002.
- [25] Kemal Akkaya and Mohamed Younis, "A Survey on Routing Protocols for Wireless Sensor Networks ", IJCSI Volume 8, 2002, pp. 259-268.
- [26] Arati Manjeshwar and Dharma P. Agrawal, "TEEN: A Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks", IEEE, 2001.
- [27] Wendi Rabiner Heinzelman, Anantha Chandrakasan, and Hari Balakrishnan, "Energy-Efficient Communication Protocol forWireless Microsensor Networks", 33<sup>rd</sup> IEEE International Conference on System Sciences , 2000, pp. 1
- [28] <https://scholar.google.co.in/>
- [29] <https://www.mathworks.com/>



# Noise Reduction in Compressed Images Using Improved Fuzzy Transform Technique

Gaganpreet Kaur<sup>#1</sup>, Priyanka Jarial<sup>\*2</sup>  
<sup>#</sup>Department of Computer Engineering,  
University College of Engineering,  
Punjabi University, Patiala, India  
<sup>1</sup>gaganpreetkaur75@gmail.com  
<sup>2</sup>jarial.priyanka@gmail.com

## ABSTRACT

In the area of digital image compression, computer algorithms are used to perform processing of images and compression. It deals with developing a digital system that perform operations on digital image. It has many advantages using in digital camera, film, satellite, X-ray and many more applications. Image compression is a technique used to save the storage space normally used to compress images and videos. Number of compression algorithms are used like run length encoding, huffman coding, discrete cosine transform, vector quantization, fuzzy transform. This gives a brief idea on improved fuzzy technique to reduce noise in compressing image. There are so many techniques for compression but in this only present the techniques of improved fuzzy method to reduce noise and compressed the image by using edge detection. The main idea behind applying this is to preserve the well significant edges as Jpeg is the popular standard but at low bit rate Jpeg exhibits blocking artifacts means noisy effects that affect the visual image quality so as to produce high visual quality image at low bit rate, the algorithm is efficient and simple. The proposed algorithm consists of three steps. First, image is preprocessed using competitive fuzzy edge detection. Second, based on edge information image is compressed and decompressed using improved fuzzy transform. Third, reconstructed image is postprocessed using hybrid median filter for artifact reduction. Analysis proves the superiority of proposed algorithm. The results of different number of coefficients are compared with the value of PSNR, MSE of algorithm. After comparison of techniques it is found to be efficient for visualisation.

**KEY WORDS -** Edge detection, Improved fuzzy, Artifact reduction, LV, MV, HV, CFED.

## I. INTRODUCTION

### OVERVIEW

Compression is very important for data storage and transmission. Incase of general data compression, it has to be a lossless one. It means, we have ability to recover the original data 1:1 from the compressed file. In this area, use something which is called a lossy compression. Our main aim is not to recover data 1:1, but keeping them visually similar.

Fuzzy is the method of real value functions. It is human ability based.

Using fuzzy transform in which improved fuzzy is used consists of three steps-

- 1 Preprocessing using competitive fuzzy edge detection.
- 2 Compression and decompression using improved fuzzy transform.

### 3 Postprocessing( noise/ artifact reduction using hybrid median filter)

Its contribution is well known that for image blocks with many edge pixels have more information and they should be less compressed and the blocks with smooth regions should be more compressed.[2]

Fuzzy transform compress each image block into the same level without taking the edge information.

The algorithm further used which compresses block by taking into account the edge information contained in the block.The algorithm gives better quality of compressed images with well preserved edges and reduced artifacts.The algorithm provides improvement in visual quality and quantitative results should be calculated.In fuzzy transform,first an image is preprocessed using fuzzy edge detection which detect the edge pixels in the image.Second on the basis of edge information,image is compressed and decompressed using improved fuzzy transform.Third,reconstruction is postprocessed using hybrid median filter for artifact reduction.The analysis proves the superiority of proposed algorithm. Transformations use fuzzy logic functions of local areas by their membership values which use basic functions for generalized functions and this technique can be successfully applied for comparing the approximate derivatives of the initial function as well as comparing the definite integrals.

#### *1.1 FUZZY TRANSFORM(F-TRANSFORM)*

Fuzzy transform is soft computing method with many applications.Showing this technique with applications to data analysis. The F-transform (Fuzzy) establishes a link between a set of continuous functions of real numbers and the set of n-dimensional (real) vectors. The inverse F-transform(inversion formula) converts real vectors into some continuous function which approximates the original function. The advantage of the inversion formula is that the F-transform is simply the approximate value representation of the original function. So in complex computations can use the inversion formula instead of representation of the original function.Fuzzy is easy to understand based on natural language. Flexible and tolerance of imprecise data.It is human logic ability based and its efficiency is very high and is very easy to use .Fuzzy logic is used in fuzzy set theory their success is for closeness to human ability as well as simplicity. It provide high knowledge base that is easy to understand and maintain .

Image compression is used in applications like televideo-conferencing, remote sensing and documents.Its main aim is to reduce redundancy of images for storing and transmitting data in an efficient manner.Uncompressed image/data require more space and time.Two types-Lossy and lossless. In lossless techniques, original image does not loss any data..In lossy technique some unnecessary information can be lost and original image cannot be recovered.There are many techniques in lossless compression like run length encoding,huffman,arithmetic and LZW coding .In lossy ,transformations used are DCT,DWT,FFT etc in transformation domain..The increasing demand for multimedia content like digital images and video has led to take interest in research on compression techniques like medical imaging, fax transmission .Fuzzy is also efficient and reliable technique used for compression.Motivation Jpeg image based on Dct is popularly used standard.A low bit rate Jpeg based compression exhibit blocking artifacts that affect the visual image quality.Fuzzy is simple and more efficient for this. Its main objective is to reduce redundancy in images for storing and transmitting the data . Uncompressed data like uncompressed video and audio, the data takes space and time . Digital image requires large space for storage and greater bandwidth for transmission. The main aim is to reduce the memory space of data so that transmission times are reduced.

Fuzzy transformation is our proposed work to compress the images because

- 1 Its efficiency is very high as compared to efficient result in Dct,Dwt etc.
- 2 This transformation is very easy not too much calculations so give accurate result.
- 3 Its main advantage is that it is human ability based or human logical reasoning based.
- 4 Its understanding is very easy.
- 5 There is less work done on fuzzy transform because it is a new logical based transform used for the accurate results.

Its drawbacks are -

- 1 It is not robust .
- 2 It considers only min max rule that is 0 and 1 values.
- 3 If we want to imitate human reason ,the minmax rule is definitely not the way.

The focus of study is to reduce the noise effect in compressed images using this transform.During compression of images, edges should be well preserved for human perception.So there are many standards such as Jpeg is most popular standard to compress images.Jpeg exhibits blocking artifacts that affect the visual quality of images. Improved fuzzy is simple and more efficient for reducing artifacts .It consists of three steps-

- 1 Image is preprocessed based on edge detecting method called competitive fuzzy edge detection.
- 2 Second,based on the edge detection image should be compressed and decompressed.
- 3 Last,used hybrid median filter for artifact reduction .This gives improvement in quality of image used for compression at a low bit rate.Fuzzy has an advantage of producing a unique representation of an original function that makes complex equations easier.It is to compress image at a very low bit rate.

## II. PROPOSED WORK

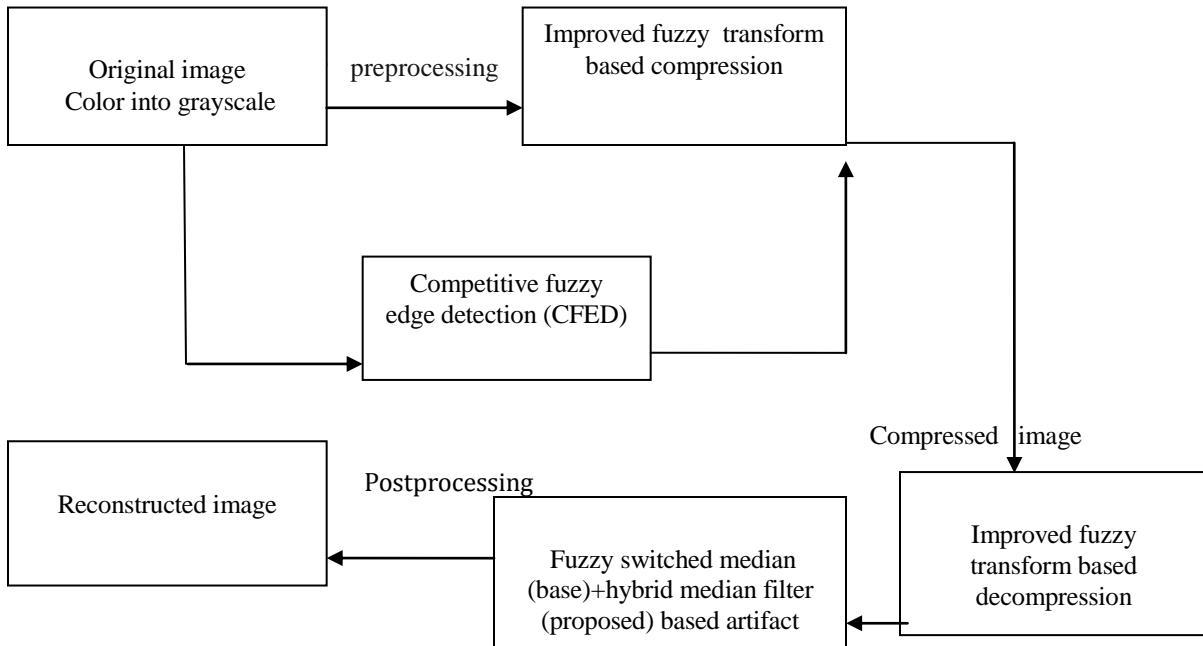


Figure1

## 2.1 Convert Color Image Into grayscale image



Figure2

## 2.2 Preprocessing Based On Edge Detecting Method Called Competitive Fuzzy Edge Detection(CFED)

In this, CFED assigns each pixel to one of the six classes depending on its neighbourhood. Classes are then classified into Low (LV), Median(MV) and High(HV) variation blocks depending on the number of edge pixels. CFED detect edge pixels by using different fuzzy membership functions based on neighbourhood situation in different directions and used for segmenting the blocks. CFED accept input in form of vectors. Each input is assigned a class depending on max/min fuzzy membership functions.

**2.2.1 4d Feature Vector-** Computing for each pixel in input image by adding the gray level magnitude difference of pixels in four directions.

$$v1 = |x(i,j) - x(i+1,j-1)| + |x(i,j) - x(i-1,j+1)|$$

$$v2 = |x(i,j) - x(i-1,j-1)| + |x(i,j) - x(i+1,j+1)|$$

$$v3 = |x(i,j) - x(i,j-1)| + |x(i,j) - x(i,j+1)|$$

$$v4 = |x(i,j) - x(i-1,j)| + |x(i,j) - x(i+1,j)|$$

### 2.2.2 Edge Classification-

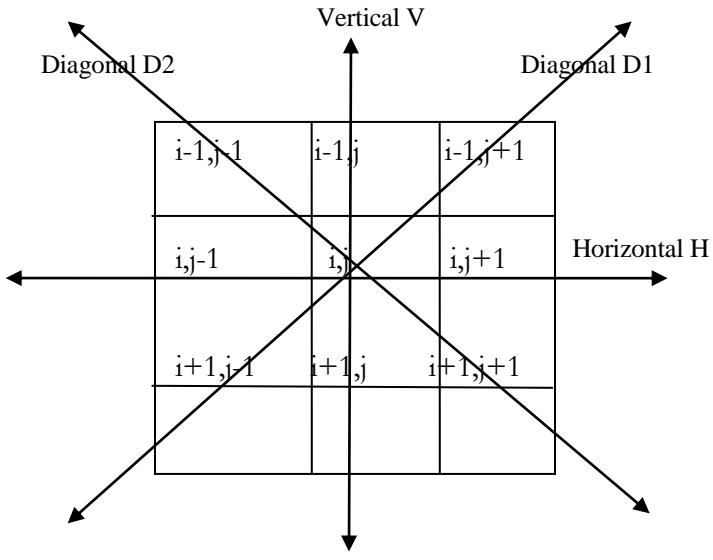


Figure 3

Class 1 is assigned to pixel which has low magnitude intensity difference along diagonal direction D2 and high magnitude intensity difference along rest of directions. Second ,Class2 is assigned that has low magnitude difference along vertical direction V and high on rest of directions. Class3 is assigned that has low along diagonal D1 and high along remaining directions. Class4 is assigned that has low along horizontal H and high along remaining directions. Then classes assigned to it and competitive rules are applied according to class assigned. Only the pixels

classified as edge pixels and sure edge pixels are assigned the value high(white) in output and all other are assigned low(black) in output image.This results in white edges on black background.[2]

### III. FUZZY RULE FIRING

Before classifying edge pixel to either black or white, it is compared with neighbourhood edge pixel .

Rule 1- If x belongs to smoothly class(class 0 ) then change color of pixel to black.

Rule 2- If x belongs to edge class 1,compare v1 with its neighbouring pixels along diagonal D2.If v1 is large then change pixel value to white else black.

Rule 3- If x belongs to edge class 2,compare v4 with neighboring on vertical V.If a large then change pixel to white else black.

Rule 4-If x belongs to edge class 3, compare v2 with neighboring pixels along diagonal D1. If it is large then change to white else black.

Rule 5-If x belongs to edge class 4,compare v3 with neighboring pixel along horizontal H.If it is large then change pixel to white else black.

Rule 6-If x belongs to sure edge class (class 5), then change pixel to white.

In this way , finally achieved the edge image.CFED can detect edges in image block. These blocks then classified into LV,MV and HV blocks depending on no of edge pixels.

Experimenting CFED on different images,it is concluded that treating blocks with less than 20 % of edge pixels as LV block, between 20 and 70 % as MV block and rest as HV block.It yields an optimum quality of compressed image at low rate.[2]

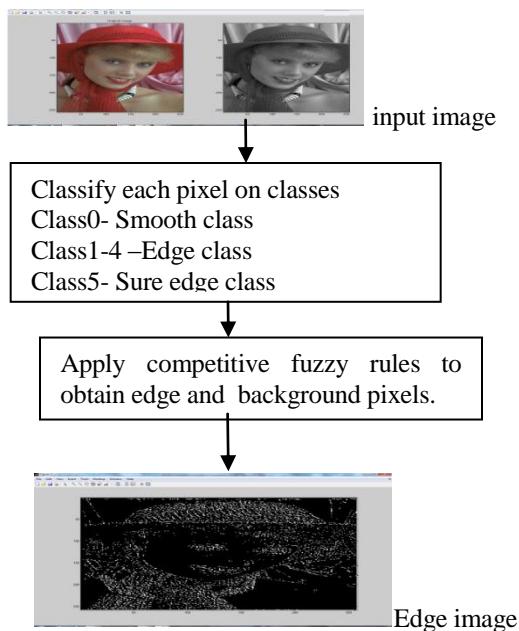


Figure4

#### IV. COMPRESSION AND DECOMPRESSION USING IMPROVED FUZZY TRANSFORM

Fuzzy transform gives better results than FEQ(fuzzy relation equations) .The value obtained using fuzzy transform are similar or slightly less as compared to Jpeg based compression result.So to provide an algorithm that perform better than JPEG and fuzzy transform.The improved fuzzy is the technique which takes each of the LV,MV and HV blocks differntly to use for compression.These blocks are compressed to different size blocks so that maintaining an average value of compression rate and performing better than JPEG standard used for compressing images at similar compression rates.

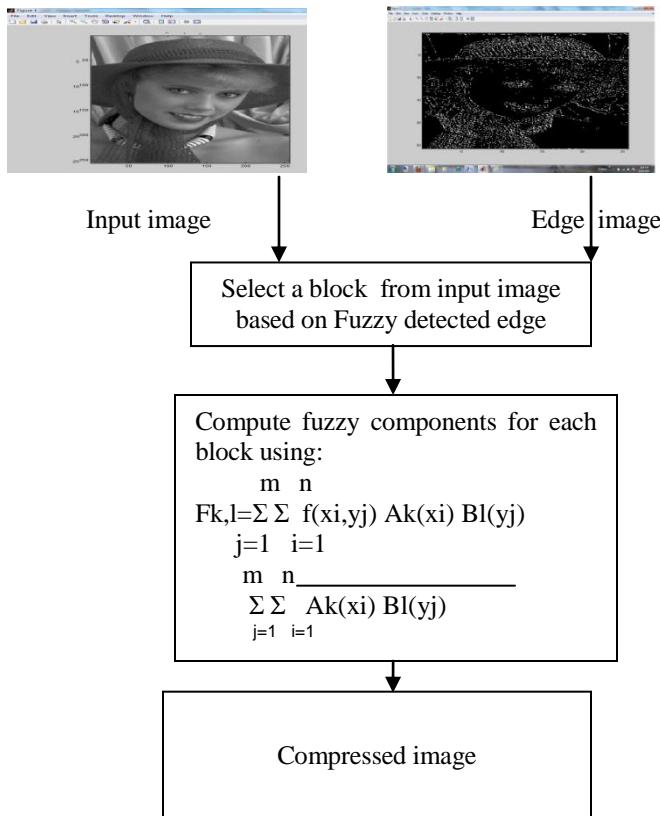


Figure 5 Improved fuzzy transform compression

Then after compressing apply inverse transform and obtain the original image.Decompressed image is achieved after that use the fuzzy switched median filter to reduce artifacts in decompressed image . After that use another filter is hybrid median filter to obtain much more better result and reducing artifacts more. PSNR and MSE calculated after proposed work gives better results.

#### V. NOISE REDUCTION

To reduce volume of a sound, soundproofing is used. To reduce the noise of machinery and products, have the example of noise control.Noise reduction is the method of removing noise from a signal.All recording devices have ability to make susceptible to noise. Noise has types it can be of random noise which are randomly occur or white noise.In electronic recording devices,the important type of noise is which is caused by random electrons that are affected by heat, stray from their designated path. These stray electrons thus create detectable noise due to voltage influence.Noise

is introduced due to the grain structure of the medium in films. The size of the grains in the film defines the film's sensitivity ,more sensitive film having larger sized grains. Filters are used to reduce noise so we are using filter after decompressed image.

*Fuzzy Switched Median Filter:-* Artifacts reduced by using a square filtering window  $W_{(2M+1) \times (2M+1)}$  where M is even integer and  $(2M+1)$  rows ,  $(2M+1)$  columns centered at pixel  $x(i,j)$  positioned at  $i,j$ .When window is placed on the right (left) boundary of image,then neighbouring pixels on left (right) considered to be free from artifact.The value of pixel calculated based on median value.

$$\text{Med}(i,j)=\text{Median}\{x(i+k,j+l) \quad \text{For } k,l= -M \text{ to } +M\}$$

After this calculate  $D(i,j)$  that provides local information of window.  $D(i,j)=\text{Max}\{|x(i+k,j+l)-x(i,j)|\}$

1 Pixels with value of  $D(i,j)$  between 0 and  $TH_1$  are non-edge pixels and assign zero membership value.

2 Pixels with value of  $D(i,j)$  between  $TH_1$  and  $TH_2$  are edge pixels with membership value between 0 and 1 and calculated by  $mD(i,j)+c$

$$\text{where } m= \frac{1}{\overline{TH_2-TH_1}} \quad c=\frac{(-TH_1)}{\overline{(TH_2-TH_1)}}$$

3 Pixels with value greater than  $TH_2$  are sure edge pixels and assign value 1.

These values are denoted by  $u(i,j)$ . $TH_1$  is value between 5 to 15, and  $TH_2$  is value between 25 to 35.

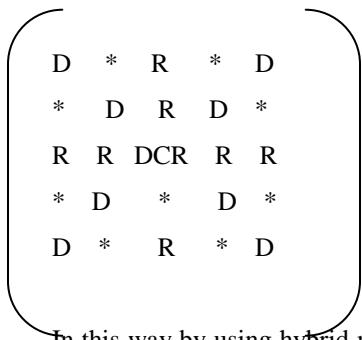
Finally artifact calculated by

$$Y(i,j)=[(1-u(i,j))]x(i,j)+u(i,j)\text{Med}(i,j)]$$

This is base work filter.

## VI. PROPOSED WORK FILTER

*Hybrid Median Filter :-*In median filter, the neighbourhood pixels should give rank according to brightness or intensity and median value becomes new value for central pixel.Median filters are best to reduce "shot" noise or impulsive noise in which some individual pixels has extreme values.It can erase lines narrower than half width of neighbourhood.They can be of roundoff corners.The hybrid median filter is a three step ranking process and preserve edges better than square kernel median filter.Three median values are calculated so called hybrid median. MR is median of horizontal and vertical R pixels,MD is the median of diagonal pixels D and centered pixel C.So filter is the median of two median values and center pixel C:  $\text{Median}([\text{MR},\text{MD},\text{C}])$ . For example n=5



In this way by using hybrid median filter calculate the median of got median values i.e median of median values so this gives better result by preserving edges in image and result calculated for different images whereas, CFED in which detector takes low=0, high=20 and width (w)=256 and results show high PSNR and low MSE value.

## VII. RESULTS AND DISCUSSIONS

To evaluate the performance taken the images of size 256x256 of Jpeg and Bitmap type of images and experiment is performed. For compression, images are divided into blocks of LV, MV and HV (low, medium and high). The total number of 8x8 size blocks for 256x256 size image is  $(256 \times 256) / (8 \times 8) = 1024$ . Experiment has been found that block containing less than 20% of edge pixels treated as LV block, between 20 and 70 % of edge pixels as MV block and rest are HV block. Hybrid median filter is used to reduce noise in compressed images. The proposed algorithm is analyzed and results obtained are compared with previous used filter. The figures first show the original color image we convert it into grayscale then achieved edge image by using CFED technique then compressed and decompressed using fuzzy method achieves the decompressed image. Base method which had used fuzzy switched median filter image and proposed method image that used hybrid median achieved better result (free from artifacts image) than base method. Artifacts are more reduced by using this method. Jpeg is standard to compress images but suffer from artifacts which affect the quality. Fuzzy transform also exhibits artifacts but smaller than from Jpeg compressed images. Hybrid median filter has more less artifacts which is our proposed work. It gives better visual quality. Hybrid median filter, is non-linear filter that easily removes impulse noise while preserving edges. Hybrid one has better corner preserving characteristics. The basic idea behind filter is for any element of the signal (image) apply median technique several times and then take the median of the got median values. Used the parameters for objective analysis PSNR and MSE. PSNR is peak signal to noise ratio. Having 50 pictures. Showing only four. The table describes results for 50 images.

*JPEG IMAGES*



1

2

3

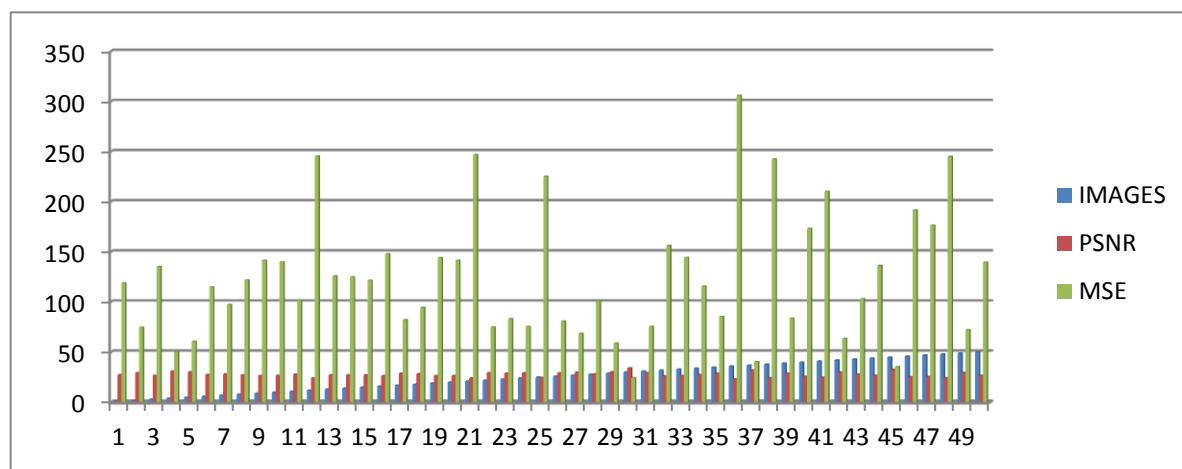
4

TABLE1

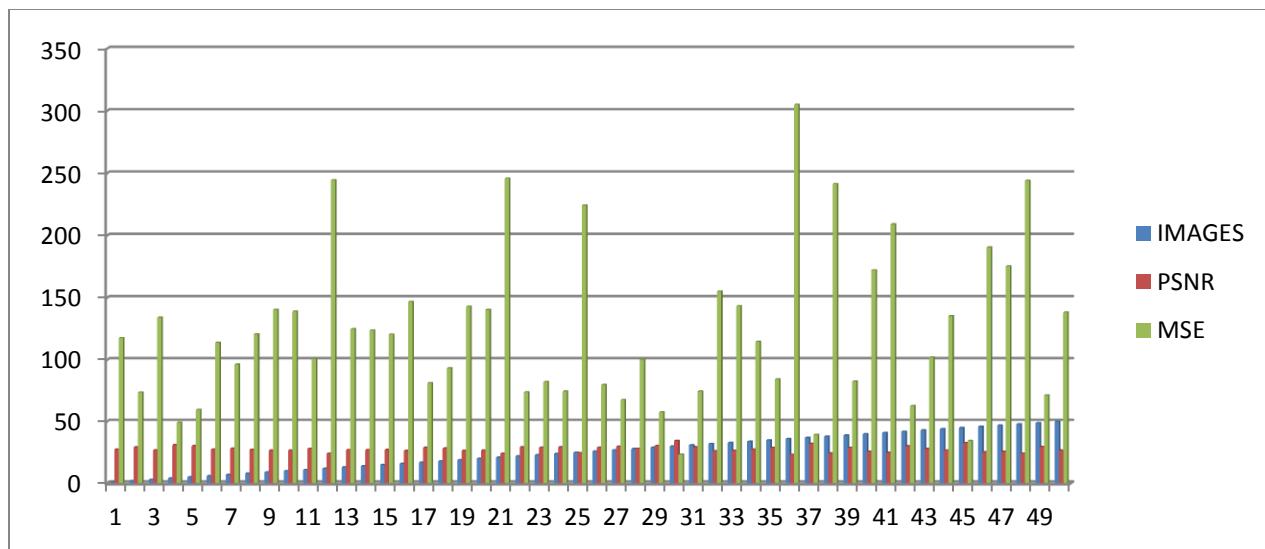
<b>DIFFERENT images for experiment (jpg)</b>	<b>Base method values (Fuzzy switched median filter)</b>		<b>Proposed method values (Hybrid median filter)</b>	
	<b>PSNR</b>	<b>MSE</b>	<b>PSNR</b>	<b>MSE</b>
1.	27.38	118.81	27.41	117.81
2.	29.39	74.76	29.45	73.76

3.	26.81	135.34	26.84	134.34
4.	31.11	50.35	31.19	49.35
5	30.29	60.75	30.36	59.75
6.	27.52	115.00	27.56	114.00
7.	28.23	97.52	28.28	96.52
8.	27.27	121.84	27.30	120.84
9.	26.62	141.48	26.65	140.48
10.	26.67	139.76	26.69	139.22
11.	28.04	102.03	28.08	101.03
12.	24.22	245.53	24.24	244.54
13.	27.12	125.99	27.16	124.99
14	27.16	124.98	27.19	123.98
15.	27.27	121.69	27.31	120.69
16.	26.43	147.93	26.45	146.93
17.	28.98	82.22	29.03	81.23
18.	28.38	94.41	28.42	93.41
19.	26.54	144.11	26.57	143.11
20.	26.62	141.37	26.65	140.54
21.	24.20	246.92	24.22	245.92
22.	29.38	74.94	29.44	73.94
23.	28.92	83.26	28.97	82.26
24.	29.34	75.59	29.40	74.59
25.	24.60	225.27	24.62	224.28
26.	29.05	80.91	29.10	79.91
27.	29.76	68.69	29.82	67.69
28.	28.07	101.36	28.11	100.36
29.	30.43	58.76	30.51	57.76
30.	34.25	24.38	34.44	23.38
31.	29.34	75.61	29.40	74.61
32.	26.19	156.25	26.22	155.25
33.	26.53	144.32	26.56	143.32

34.	27.49	115.87	27.52	114.87
35.	28.81	85.48	28.86	84.48
36.	23.27	306.14	23.28	305.14
37.	32.06	40.38	32.17	39.38
38.	24.28	242.48	24.30	241.48
39.	28.90	83.64	28.95	82.64
40.	25.74	173.22	25.76	172.22
41.	24.90	210.19	24.92	209.19
42.	30.08	63.75	30.15	62.75
43.	27.99	103.09	28.04	102.09
44.	26.77	136.51	26.81	135.51
45.	32.63	35.46	32.75	34.47
46.	25.30	191.59	25.32	190.59
47.	25.66	176.49	25.68	175.49
48.	24.23	245.09	24.25	244.09
49.	29.54	72.29	29.60	71.29
50.	26.68	139.45	26.71	138.45



Fig(A) Showing PSNR And MSE Values Of Base Method For Jpeg (50 Values)

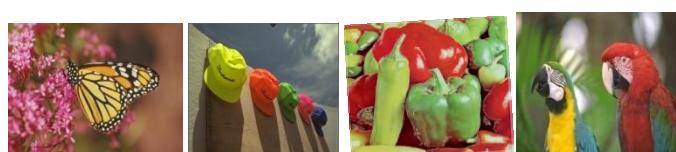


Fig(B) Showing PSNR And MSE Values Of Proposed Method For Jpeg(50 Values)

The comparative analysis of table1 shows that to reduce the noise using proposed filter achieve better results as compared to base method .The value of peak signal to noise ratio achieved using proposed is high value as compared to base method which means image quality achieved is optimum and error achieved is less which means distortion occurred in image is less .Jpeg is lossy type of compression so it achieves better results to reduce the noise with edges well preserved.The bargraph shows the result of better noise reduction at low rate.

Again for bitmap,having 50 pictures.Showing only four.The table describes results for 50 images.

#### BITMAP IMAGES



1

2

3

4

TABLE2

<b>DIFFERENT images for experiment (bmp)</b>	<b>Base method values (Fuzzy switched median filter)</b>		<b>Proposed method values (Hybrid median filter)</b>	
	<b>PSNR</b>	<b>MSE</b>	<b>PSNR</b>	<b>MSE</b>
1.	24.22	245.76	24.24	244.76
2.	29.55	72.06	29.61	71.05
3.	25.89	167.52	25.91	166.52
4.	29.63	70.66	29.70	69.66
5	29.39	74.76	29.45	73.76
6.	26.57	143.02	26.60	142.02
7.	26.06	160.83	26.09	159.84
8.	20.89	529.06	20.90	528.06
9.	30.07	63.87	30.14	62.87
10.	23.03	323.13	23.05	322.13
11.	21.52	458.00	21.53	457.00
12.	26.75	137.17	26.78	136.18
13.	25.93	165.61	25.96	164.61
14	23.02	323.79	23.04	322.79
15.	26.34	150.78	26.37	149.78
16.	21.36	475.14	21.37	474.14
17.	28.43	93.33	28.47	92.33
18.	24.69	220.34	24.71	219.35
19.	22.05	405.17	22.06	404.18
20.	24.76	217.10	24.78	216.10
21.	26.44	147.55	26.47	146.56
22.	27.53	114.73	27.57	113.73
23.	22.87	335.06	22.89	334.06

24.	23.51	289.65	23.52	288.65
25.	27.11	126.28	27.15	125.28
26.	25.15	198.25	25.18	197.25
27.	26.66	140.23	26.69	139.23
28.	22.57	359.53	22.58	358.53
29.	27.96	103.88	28.00	102.88
30.	22.38	375.33	22.39	374.33
31.	28.90	83.74	28.95	82.74
32.	30.75	54.69	30.83	53.69
33.	33.20	31.09	33.34	30.09
34.	27.10	126.60	27.14	125.60
35.	28.21	98.16	28.25	97.16
36.	25.69	175.31	25.70	174.68
37.	26.29	152.75	26.31	151.75
38.	23.91	264.28	23.92	263.28
39.	25.51	182.67	25.53	181.68
40.	27.49	115.68	27.53	114.68
41.	25.97	164.44	25.99	163.44
42.	27.76	108.86	27.80	107.87
43.	30.77	54.36	30.85	53.37
44.	26.02	162.29	26.05	161.32
45.	23.83	269.07	23.84	268.07
46.	22.27	385.45	22.28	384.46
47.	25.96	164.60	25.99	163.60
48.	35.27	19.30	35.50	18.31
49.	26.82	135.01	26.85	134.02
50.	25.59	179.40	25.61	178.40

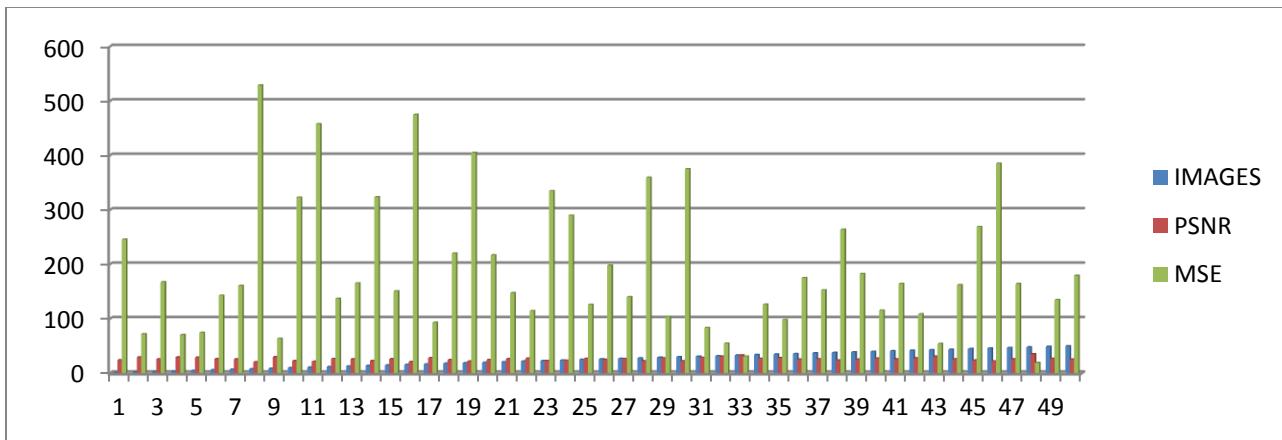
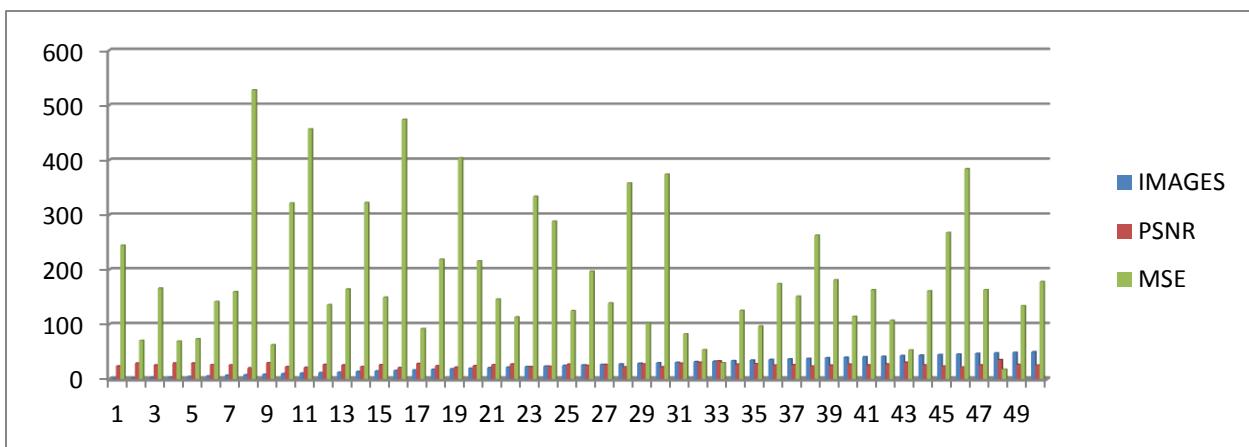


Fig (A) Showing PSNR And MSE Values Of Base Method For Bitmap(50 Values)



Fig(B) Showing PSNR And MSE Values Of Proposed Method For Bitmap(50 Values)

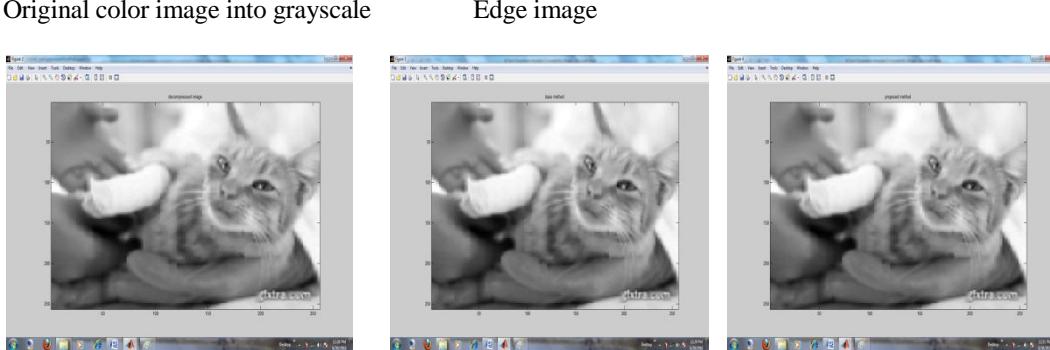
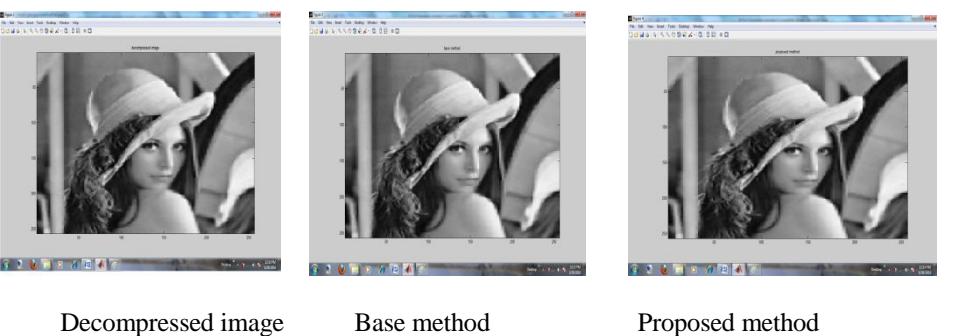
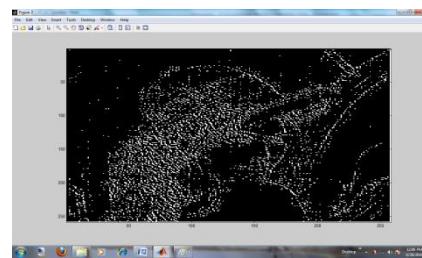
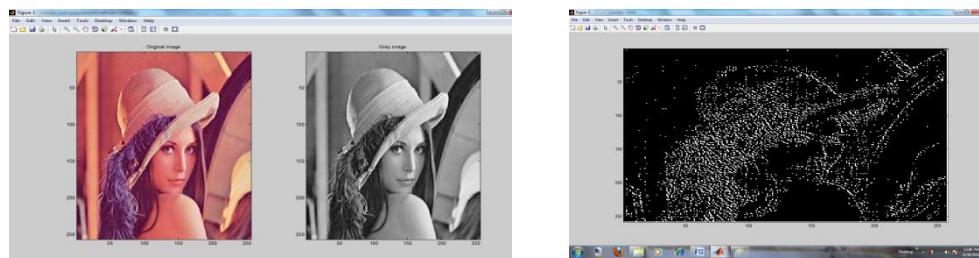
**PSNR(X,Y)=** $10 \log_{10} [L^2 / MSE]$  where L=max possible value of intensity,L=255 and MSE is defined as

$$\text{MSE}(X,Y)=\frac{\sum_{i=1}^n \sum_{j=1}^m [X(i,j)-Y(i,j)]^2}{M \times N}$$

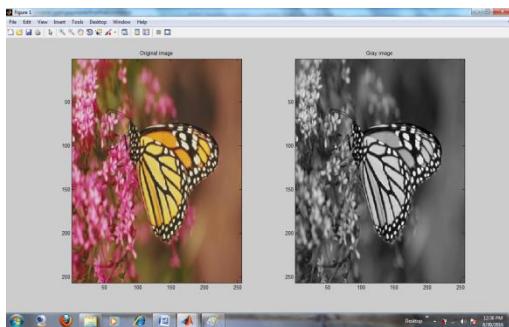
X(i,j) and Y(i,j) are gray level pixel values of original and reconstructed image.

These two are commonly used parameters and easy to calculate and clear physical meaning.These measures are based on pixel level calculation. The comparative analysis of table2 shows that to reduce the noise using proposed filter achieve better results for bitmap type images also .The bitmap images achieves lossless type of compression so it not compresses more as compared to jpeg. Noise occur is not much more but still it achieves better results using proposed method by comparing with the base . The bar graph shows the result of noise reduction at low rate.

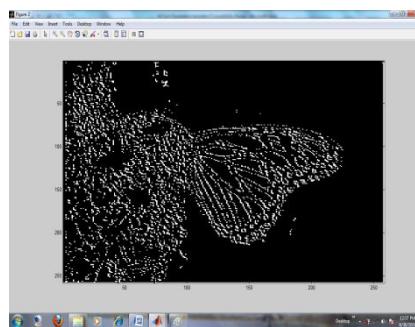
*Results Jpeg*



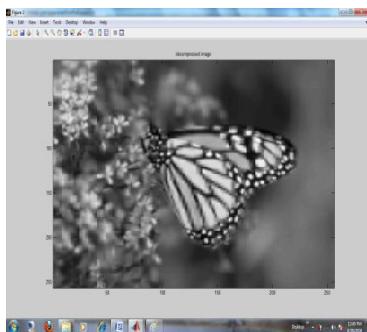
*Results BMP*



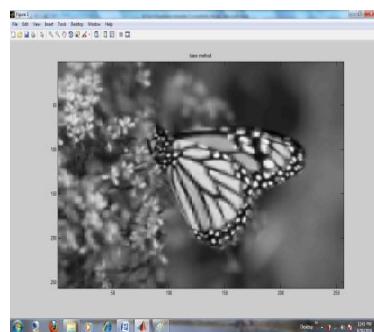
Original color image to grayscale



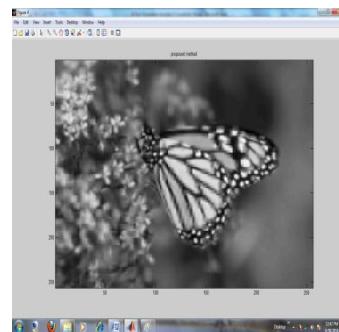
Edge image



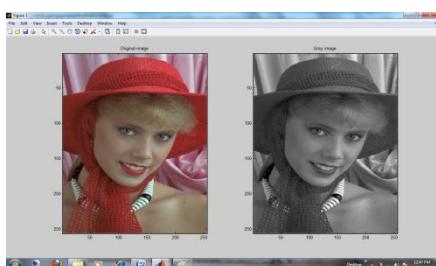
Decompressed image



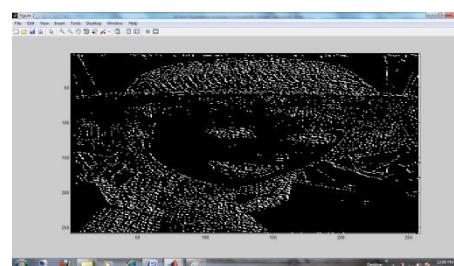
Base method



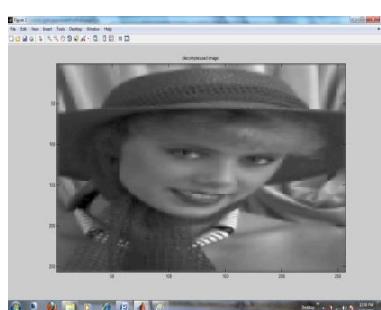
Proposed method



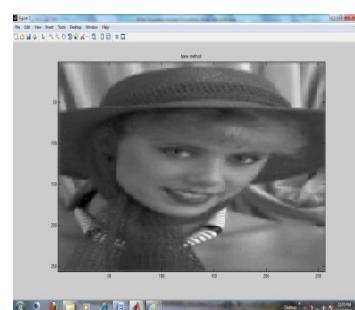
Original color to grayscale



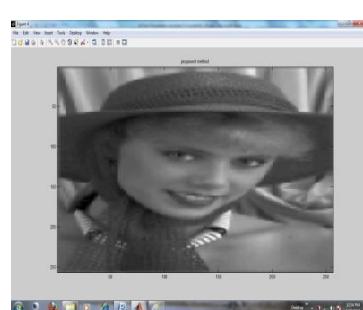
Edge image



Decompressed image



Base method



Proposed method

The image finally achieved is free from artifacts image with high PSNR and low MSE.

### VIII. CONCLUSION

This paper proposes a method to reduce noise in compressed images named Noise reduction in compressed images using improved fuzzy transform technique. The proposed method preprocesses an image using CFED(competitive fuzzy edge detector) that results that each image block classifies as either low variation(LV), medium variation(MV) and high variation(HV). These blocks are then compressed and decompressed using improved fuzzy. The reconstructed image contain blocking artifacts or noise that are reduced using hybrid median filter. The proposed method is subjectively and objectively analysis and compared with previous method and found better in terms of quality and artifact reduction as compared to fuzzy and jpeg. Done work for jpeg and bitmap for 50 values of both. This method achieves better improvement in terms of quality but still it requires more improvement. In future, it has the advantage that the neural network with fuzzy transform technique can be used to develop image compression algorithm at low bit rate. Also by comparing proposed filter with other filter prove best result.

### REFERENCES

- [1] K.Meenakshi;G.N Beena(2014) " Design and simulation of constant bit rate compressor using fuzzy logic," ,IEEE 978-1-4799-3486-7/14.
- [2] Deepak gambir;Navin rajpal (2015) "Improved fuzzy transform based image compression using pair fuzzy," Springer Int. J. Mach. Learn. & Cyber. 6:935–952 DOI 10.1007/s13042-015-0374-1.
- [3] Maneesha gupta;Dr. Amit Kumar garg (2015) "Analysis of Image Compression algorithm Using Dct," by gupta.
- [4] " Efficient image compression using all the coefficients of 16x16 Dct image subblock,"(2015).
- [5] "An Image compression technique based on fuzzy transform ,"(2014).
- [6] " Design and simulation of constant bit rate compressor using fuzzy logic," (IEEE 2014).
- [7] Arun Kumar PS ,Dept. of ECE, NIT Rourkela(2007-2009) "Implementation of Image Compression Algorithm using Verilog with Area, Power and Timing Constraints,"by kumar .[8] Ning Xu, Embedded Networks Laboratory, Computer Science Dept. USC. Los Angeles . " Implementation of Data Compression and FFT".
- [9] M. Klimesh,1 V. Stanton,1 and D. Watola1 (2001) "Hardware Implementation of a Lossless Image Compression Algorithm Using a Field Programmable Gate Array ."
- [10] Sadashivappan Mahesh Jayakar;K.V.S AnandBabu;Dr.Srinivas K (2011) "Color Image Compression Using SPIHTAlgorithm International Journal of Computer Applications,"Volume 16– No.7, pp 34-42.
- [11] David F. Walnut(2003) "An Introduction To Wavelet Analysis,American Mathematical Society" Volume 40, Number 3,Birkhauser, Isbn-0-8176-3962-4.Pp. 421-427.
- [12] M. I. Khali (2010) "Image Compression Using New Entropy Coder, International Journal of Computer Theory and Engineering," Vol. 2, No. 1 1793-8201.
- [13] Andrew B. Watson(1994)" Image Compression Using the Discrete Cosine Transform ,," NASA Ames Research Center .
- [14] Rupinder Kaur, Nisha Kaushal (NITTTR, Chd.) (2007) "Comparative Analysis of Various Compression Methods for Medical Images ."
- [15] Desai U, Masaki I, Chandrakasan A, Horn BKP(1996) " Edge and mean based image compression" IEEE ICASP, p 49.
- [16] De A, Guo C (2014) "An image segmentation method based on the fusion of vector quantization and edge detection with applications to medical image processing. Int J Mach learn 5(4):543-551
- [17] Gambhir D,Rajpal N(in Press) Image Coding using fuzzy edge classifier and fuzzy f- transform: dualfuzzy,Int J fuzzy Comput model ISSN online:2052-3548.
- [18] Rahul Shukla and Narendra Kumar Gupta "Image compression through Dct and Huffman Coding Technique" IJCET impressco(2015).
- [19] Er abhishek Kaushik,Deepti nain "Image compression Algorithms using dct" IJERA vol4 issue 4 (April 2014).
- [20] Priyanka Dixit,Mayank Dixit " Study of jpeg Image compression technique using Dct" IJIRI vol1 issue (1 oct-dec 2013).
- [21] Abhishek sahu,Praveen yadav " Development of constant bit rate jpeg image compression using fuzzy logic" IJSR (2015).

- [22] A.M. Raid,W.M Khedr M.A. El-dosuky and Wesam Ahmed “Jpeg Image compression using discrete cosine transform In IJCSES vol-5 no2,(April 2014).
- [23] Gaganpreet Kaur, Priyanka jarial” A Survey on Dct and Fuzzy image compressionIJIR
- [24] Garima goyal,Ajay Kumar,Manish “Impact and analysis of hybrid median filter” .

# A Review on the movement of object detection and video stabilization for aerial surveillance system

**Jagdeep Kaur (M.Tech Student)**

Department of Computer Engineering,  
Ycoe , Talwandi Sabo, Punjab, India

Email-id: [deep.sandhu256@gmail.com](mailto:deep.sandhu256@gmail.com)<sup>1</sup>

**Er.Ashok Kumar Bathla(Assistant Professor)**

Department of Computer Engineering,  
Ycoe ,Talwandi Sabo, Punjab, India

Email-id: [ashokashok81@gmail.com](mailto:ashokashok81@gmail.com)<sup>2</sup>

**Abstract** — Aerial videos are mostly useful videos for different area work. There is main problem in aerial videos is that the video is not in stable position there are motions are detected in video. To get that video stable we use SIFT and SURF algorithms. Scale Invariant Feature Transform(SIFT) with its high accuracy and relatively low computation time, become the de-facto standard. Speeded Up Robust Feature(SURF) which has been shown to yield comparable or better results to SIFT while having a fraction of the computational cost .Both algorithms give results that are invariant to scale, rotation, occlusion, change in illumination and noise.

**Keywords**— Moving object detection, Aerial Surveillance, Scale invariant feature Transform (SIFT), Speeded up Robust Features (SURF), Digital video stabilization.

## I.INTRODUCTION

Observation could also be the gathering about police work, sagacity what is a lot of police work info (usually visual symbolism or video) from a mobile vehicle—such as an unmanned Aerial vehicle (UAV), helicopter, alternately spy plane and observation aircrafts utilize. Associate in Nursing extend for sensors (e.g. Radar) on screen their planned space through live streaming video, sound conjointly recorded secure info. Advanced imaging technology, miniaturized computers what is a lot of varied completely different innovative developments? In days lapsed decade bring helped quick developments for flying intelligence activity instrumentation maybe micro-aerial vehicles, advanced infrared and high-positioning symbolism practiced of distinctive queries in greatly long distances.

## IMAGE REGISTRATION

Image registration is a method of reworking totally different sets of knowledge into one organization. Information from multiple images, totally different sensors, times, depths or viewpoints. Image registration is a picture process technique accustomed aligns multiple scenes into one integrated image. It helps overcome problems like image rotation, scale and skewness that area unit common once overlaying pictures. Picture enlistment can be the procedure about overlaying two or greater amount pictures of the same scene made from different views, toward different times or with distinctive sensors.



Fig.1 Image Registration

#### *IMAGE REGISTRATION PROCESS*

The image registration process is shown below in fig.2.

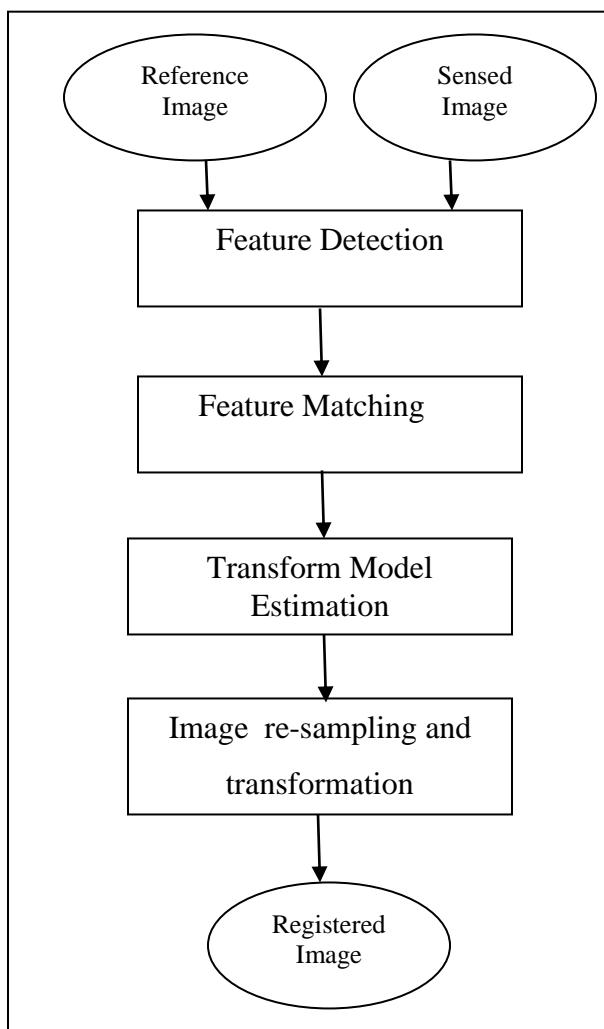


Fig.2 Image Registration Process

*Characteristic identification*-Closed-boundary regions, edges, contours, transport intersections, corners and so forth throughout and observing.

*Characteristic matching*-In this step the correspondence between those offers distinguished in the sensed picture and the individuals distinguished in the reference picture is created.

*Picture re-sampling*-What's more change, those sensed picture will be converted by method for the mapping capacities. Picture values clinched alongside non-integer coordinates are registered by that fitting insertion strategy.

#### **VIDEO STABILIZATION**

Observation are the screening of the behavior, activities or different evolving information typically from claiming people for the explanation for influencing, managing, directing and alternately securing them. Flying observation are presently days actual outstanding to broadcasting, news, shooting or gathering info beginning with air what is additional giving in-depth measure from claiming feature info for a great deal of individuals functions together with look and rescue, military operations, business applications, counter terrorist act and additionally fringe watch. Advanced imaging technology miniaturized computers and varied totally different innovative developments once more days glided by decade bring helped quick developments for flying intelligence operation instrumentality maybe micro-aerial vehicles, advanced infrared and additionally high-positioning symbolism suitable likeness queries toward greatly long distances.

#### *Algorithms used*

##### *SIFT (Scale Invariant Feature Transform)*

The first algorithm used for feature extraction is SIFT. Scale Invariant Feature Transform (SIFT) is used as feature extraction method from images and it is freely available for researchers from University of British Columbia but for commercial purpose needs license to purchase. SIFT extract a large number of distinctive potential key points from an image that is invariant to different viewpoints, rotation, scaling.

##### *SURF (Speeded Up Robust Features)*

Second algorithmic program used for feature extraction is SURF. SURF (Speeded up Robust Features) is additionally a feature extraction algorithmic program works nearly the same as SIFT algorithmic program, however quicker than SIFT [8]. SURF algorithmic program strong to detector and descriptor of potential interest key points.

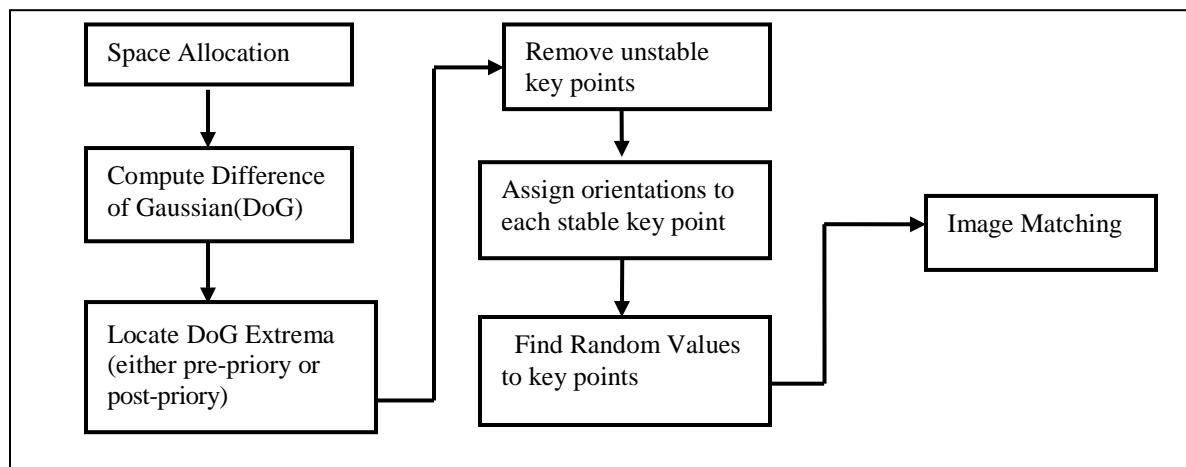


Fig.3 Flow chart of SIFT algorithm

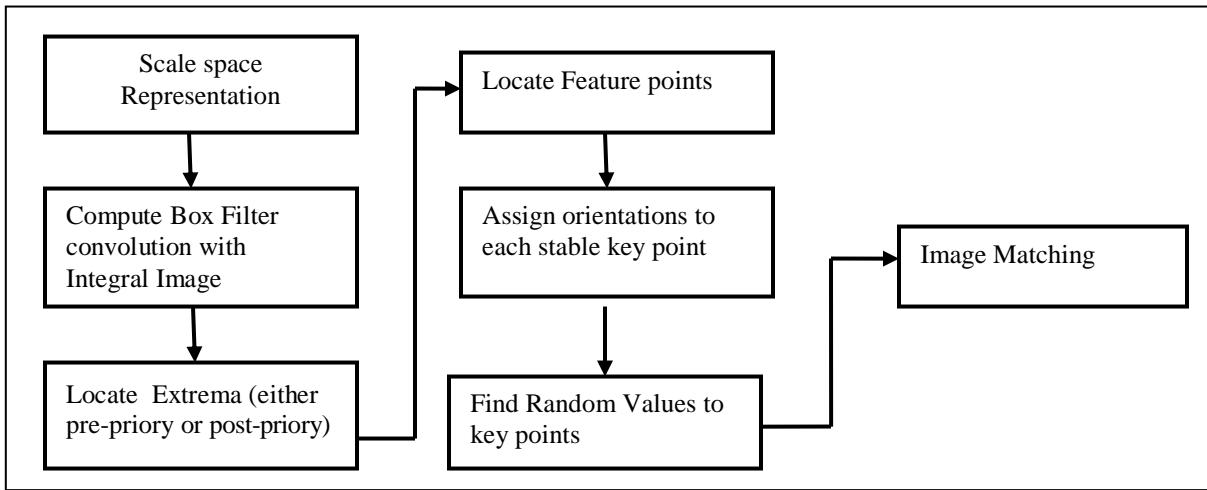


Fig.4 Flow chart of SURF algorithm

#### *Kalman Filter*

Kalman filter is employed to estimate vary to obstacles. The computation of a variety map for a typical scene could involve many hundred Kalman filters. Optical flow computations offer the measurements for every Kalman filter. Kalman filter activity update takes place at fastened intervals.

## II. RELATED WORK

Sridhar et al.(1993) [1] broadened Kalman channel is employed to estimate reach with obstacles conjointly done associate normal Kalman filter, those estimation upgrade takes place in settled intervals. Those overhead and correspondence load connected with some hundred Kalman filters running asynchronously makes it to a fault convoluted what is more may exceed the preferences of the off probability driven Kalman channel. Santhaseelan et al.(2010) [3] Associate in Nursing versatile parameterization system could also be urged can characterize those qualities of the channel wont to eliminate high back elements within the movement manner. Movement of the Polaroid is evaluated utilizing filter characteristic following. Ought to move forward those algorithmic programs on these lines that those characteristic following methodology isn't influenced. Eventually Tom's reading absence of insufflates lighting future partake) energizes this territory additionally incorporates modification of the calculation with attain current execution. Zhang et al. (2010) [5] those objective of feature adjustment principally are ought to fathom those smirched feature initiated Eventually Tom's reading the unwanted Polaroid developments. This paper proposes an algorithmic program, which African sandalwood treat those completely different feature queries singly in sight of their price of majority the info what is more reduce the period of the time lost on the muse locus with success. Compared for existing algorithms, those executions of the calculation urged. Eventually Tom's reading the paper want been considerably improved additionally progressed clenched aboard a number of zones. Wong et al.(2014) [6] author urged Associate in Nursing novel strategy that employments the size regarding matched SURF image offers additionally dynamic run through warp with perform stable restriction. By different SURF characteristic scales between info photos additionally pre-constructed information stable restriction could also be earned while not those necessity on ascertain those key grid. The check outcomes indicate that associate in nursing information image matching preciseness within 2 sequential photos African sandalwood an opportunity to be earned

at associate in nursing rate from claiming half of 1 mile alternately higher. Path distinction rates fluctuate between eighty four can one hundred. Future fill in incorporates moved forward information development and execution dissection utilizing image successions for com wood truth GPS additionally odometer majority of the info. Suaib et al. (2014) [7] creator displays the assessment for Scale Invariant Feature Transform (SIFT) and Speeded up Robust Features (SURF) exhibitions. Those outcomes indicate that surf are beat over filter clenched aboard expression from claiming rate from claiming matched focuses what is more additionally over process probability. Correlation regarding execution the center of filter and SURF calculation within the haul from claiming matching rate that is the rate of range regarding matched focuses contrasted with those range of distinguished focuses bring been news person. Moreover, the time period for matching rework for each calculation { may be additionally} inspected and also blacks.

### III. METHODOLOGY

Feature adjustment is those pre-processing venture that is generally connected on dissect flying feature observation. The plan behind that the video stabilization for a given collection of images or dataset is to recognize similar objects detection. There are Kalman Filter is used to estimate the variation of objects in different places. There are two algorithms are used to do so, SIFT (Scale Invariant Feature Transform) and SURF (Speeded up Robust Features).In the performance of SIFT and SURF, SURF algorithm is faster than SIFT. This area reviews the four essential sorts for feature reconnaissance gear – cameras, camcorders (camera-recorders), recorders/players. Also feature shows (monitors/televisions) – that cam wood a chance to be utilized (in A percentage combination) to structure.

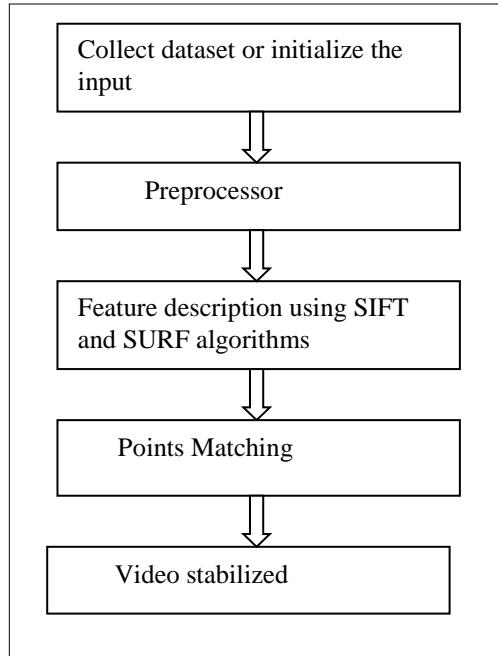


Fig.5 Flow diagram of methodology

#### IV. CONCLUSION

Different algorithms for video stabilization are described in this paper. We tend to attempt re-arrangement of moving object detection into the stabilization rule and to work out that detection once stabilization doesn't work well. We tend to jointly confirm that SIFT as choices unit of measurement durable for video stabilization and moving object detection functions and SURF is to boot a feature extraction algorithmic program works nearly like SIFT algorithmic program, but faster than SIFT [8]. SURF algorithmic program robust to detector and descriptor of potential interest key points.

#### ACKNOWLEDGEMENT

I am Thankful to my respected guide Er.Ashok Kumar Bathla, Assistant Professor (Computer Engineering), Yadavindra College of Engineering, Talwandi Sabo for his invaluable and enthusiastic guidance and useful suggestions.

Lastly and most importantly, I remain indebted to my parents, well-wishers and almighty for always having faith in me and for their endless blessings.

#### REFERENCES

- [1] Sridhar B., Smith P., Suorsa R., and Hussien R. (1993)," Multirate and Event-Driven Kalman Filters for Helicopter Flight" IEEE Control Systems pp. 26-33.
- [2] Pluim J. P.W., Maintz J. B. A., Viergever M.A. (2003), "Mutual-Information-Based Registration of Medical Images: A Survey" IEEE Transactions on Medical Imaging, Vol. 22, No. 8, pp. 986-1004.
- [3] Santhaseelan V. and Asari V.K.(2010)," An Adaptive Parameterization Method for SIFT based Video Stabilization"IEEE 39<sup>th</sup> Applied Imagery Pattern Recognition Workshop (AIPR) .
- [4] Chen Y.H. and Lin H.Y.S., Su C.W.(2014)," Full-frame Video Stabilization via *SIFT* Feature Matching" IEEE Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing pp. 361-364.
- [5] Zhang G., Yu L., Wang W.(2010)," Video Stabilization Algorithm Based on Video Object Segmentation "IEEE 2nd International Conference on Future Computer and Communication Vol. 2 pp. 509-512.
- [6] Wong D., Deguchi D., Ide I., Murase H.(2014)," Single Camera Vehicle Localization using SURF Scale and Dynamic Time Warping " IEEE Intelligent Vehicles Symposium (IV) June 8-11, pp. 681-686.
- [7] Suaib N.M., Marhaban M.H., Saripan M.I., and Ahmad S.A.(2014)," Performance Evaluation of Feature Detection and Feature Matching for Stereo Visual Odometry Using SIFT and SURF" IEEE Region 10 Symposium pp. 200-203.
- [8] P M Panchal, S R Panchal, S K Shah(2013), "A Comparison of SIFT and SURF", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 2, pp 323-327.

# RANDOM PATTERN BASED STEGANOGRAPHY USING SEQUENTIAL BIT ENCODING FOR HIDING DATA IN 3D ULTRASOUND VIDEO

<sup>1</sup>Raghvir Singh Grewal, <sup>2</sup>Gaurav Deep

<sup>1</sup>Research Scholar

(Department of Computer Engineering)

UCoE, Punjabi University, Patiala

Email: raghvir.grewal13@gmail.com

<sup>2</sup>Assistant Professor

(Department of Computer Engineering)

UCoE, Punjabi University, Patiala

Email: deepgaurav48@pbi.ac.in

*Abstract— The hackers always remain the point to grab the chance to steal the important information to make money by selling it or when working for the spy or detective organizations. Hence the data embedding techniques becomes very important for the purpose of information security. The data is being embedded in the several types of data such as image into images, signal into images, text into image, signal or video, image to image, etc. The proposed model focuses upon the embedding of the image data into the video frames. The specialized part of the proposed model lies with the embedding of the secret data in the various segments before embedding and then embeds each segment of data into the different frames of the video to minimize the probability of detection. The proposed model is robust model for the purpose of embedding the secret medical imagery data into the 3-D medical video data. The data security lies within the process of embedding where the data is encrypted before being embedded into the video frames. The proposed model is based upon the deterministic but random pixel list generation, where the heuristic embedding takes place for the embedding of each pixel in the given data. The proposed model has been analyzed for its performance under the various domains such as mean squared error, peak signal to noise ratio, histogram difference etc. The proposed model has performed better on the basis of all of the performance evaluation parameters. The histogram model has shown the least difference between the two histograms obtained from the embedding process before and after the embedding. In the clear words, the proposed model can be declared winner over the previous models of steganography.*

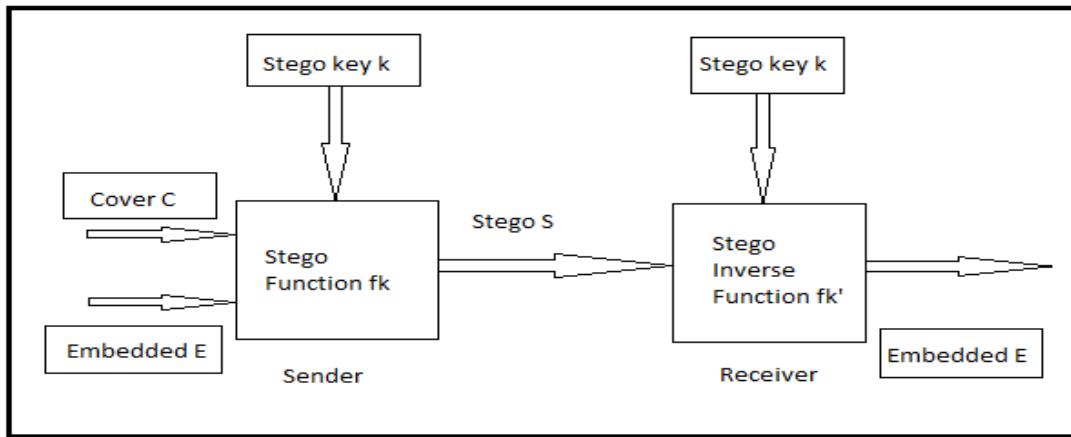
*Keywords— Video steganography; Transform domain embedding; Frequency embedding; heuristic steganography.*

## I. INTRODUCTION

Due to the detonation of the web programs, it leads mankind into the digital world and the transmission which is done through digital information gets repeated. Although some difficulties emerge and have been traversed such as security of the information in digital transmission, hidden transmission by means of digital channel. Now-a-days there are several methods which has been already proposed for the security of the image data based upon the steganography and cryptography models. The standard algorithm for the steganography is based upon the following design in Fig. 1.

It is the foremost wide used technique for secret communication. This method exploits the constraints of Human Visual System [2] [7]. Human eye cannot find the variation in luminousness of color vectors at high frequency facet of the visual

spectrum. An image is diagrammatical by a set of color pixels. A personal picture element are often diagrammatical by its optical characteristics like brightness, chrominance etc.



**Fig. 1: Basic methodology used for Steganography**

A video could be a combination of audio and image. A nonstop flow of image constitutes a video. Therefore, the techniques which will be applied on audio and image on an individual basis, they'll be applied on video conjointly. The most advantage of video is that it's comprised of enormous quantity of knowledge, so the tiny low distortion in information doesn't place any adverse result on video quality and it will go unobserved through human eyes [7].

## II. LITERATURE REVIEW

S. Thilagamani et al. (2011), during this paper there's a survey on completely different agglomeration techniques to succeed image segmentation. Completely different agglomeration techniques are mentioned which might be applied pictures and databases. Agglomeration are often termed here as a grouping of comparable pictures within the information. The method of agglomeration is completed supported completely different attributes of a picture size, color, texture etc.

Pooja Yadav, Nishchol Mishra, and Sanjeev Sharma, In this paper authors have proposed a method of embedding secret video in cover video using LSB technique based upon sequential encoding method. The secret video is encrypted using XOR encryption key before embedding for increasing the security. In this author has used a very common LSB technique which is vulnerable to steganalysis attack. The XOR Encryption key is also very weak and cannot withstand cryptanalysis attack.

Harpreet Kaur, Gaurav Deep, this paper gives introduction to segmentation, clustering and their respective algorithms which are mostly used by researchers. As we know, image is a visual representation of person, place or something. The purpose of image segmentation is to partition an image into meaningful regions with respect to a particular application. Therefore, several image segmentation algorithms such as fast scanning, region growing, region splitting, and merging were proposed to segment an image before recognition or compression. Image segmentation is to classify or cluster an image into several parts according to the feature of image.

### III. EXPERIMENTAL DESIGN

The Random pattern based LSB method has implemented using Pseudo Random Sequential Embedding algorithm is implemented. In proposed model, the ultrasound video frames are evaluated using the similar regions against the text and images using the pixel-based analysis after the frame extraction using histogram based color matching and analysis after the acquisition of the cover and hidden object. An exhaustive search is conducted to pair up the same color pixels within a cluster. Each pixel which is similar in color within threshold value is included in a cluster. After the similar region selection, the region properties are obtained, and their color table is created and maintained based on region texture and color mapping properties. As it is very rare that two or more regions can be of same size, therefore, the regions which contains the largest number of pixels is chosen so that embedding space should be as large as possible. It is useful when embedding a larger message. After the message is embedded into the cluster, the stego-image generated can be sent over the internet in a secure way. Also the XOR based symmetric key cryptography would be implemented over the secret image or text data to create the more secure hidden object by encrypting the secret message (or secret image). The above mentioned method would be designed in the three basic models comprised of encryption & decryption module, embedding & extraction module, data acquisition & data validation module. The proposed model has been designed to realize the robust embedding with maximized non-detect ability to enhance the level of security during the embedding. The proposed model is based upon the multi-layered XOR based symmetric key encryption to add the higher level of the security to the existing model. The proposed model design has been explain the following figure 2.

---

**Algorithm 1: Main data embedding algorithm**

---

- 1) Select the cover video data
- 2) Extract the frames from the video data
- 3) Select and load the secret data
- 4) Count the number of the frames extracted from the video data
- 5) Run the iteration to compute the similarity of each frame with the cover data
- 6) Prepare the comparative similarity matrix
- 7) Find the most similar frame
- 8) Check the size based compatibility of the secret data and the cover data
- 9) If the cover data is found larger than the given condition
  - a) Segment the cover data into the multiple segments
  - b) Embed the  $i^{\text{th}}$  segment into the selected frame
    - i) Input the secret key
    - ii) Encrypt the secret segment data using the secret key
    - iii) Input the random seed pixel
    - iv) Create the random pattern using the random permutation
    - v) Embed the data in the image over the pattern sequence
  - c) If it's the last segment
    - i) Return the stego data
  - d) Otherwise
    - i) Goto 9(a)
- 10) Otherwise
  - i) Input the secret key
  - ii) Encrypt the secret segment data using the secret key

- iii) Input the random seed pixel
  - iv) Create the random pattern using the random permutation
  - v) Embed the data in the image over the pattern sequence
- 11) Return the embedded data

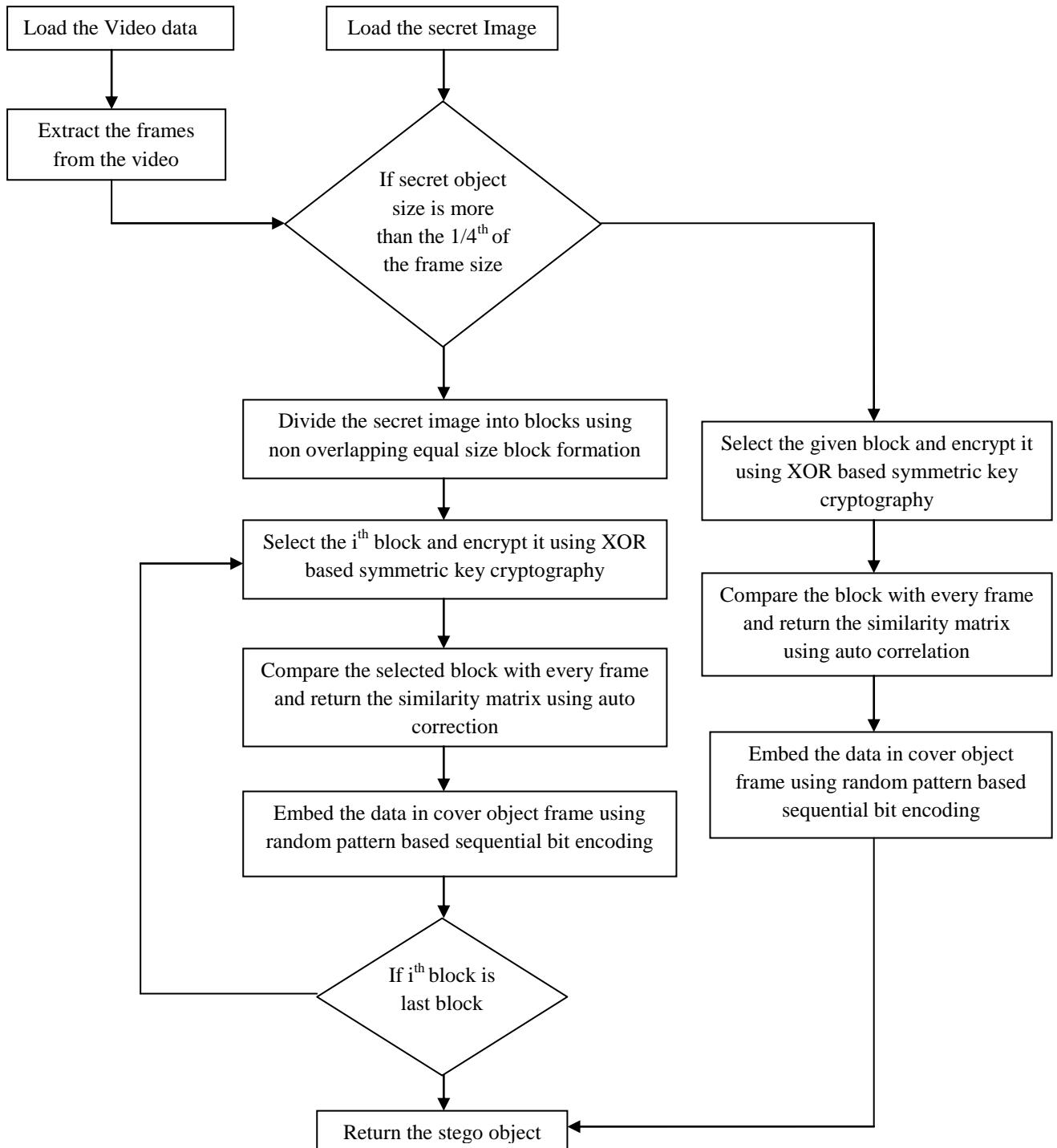


Fig. 2: Working of an Algorithm

**Sample Data:** The sample data from five different 3d ultrasound videos is selected for testing of the system. I have selected five frames randomly as cover image from each video for result analysis by implementing the proposed technique. The secret image of MRI is used for embedding in the selected frames. Sample data images are shown below in fig 3.

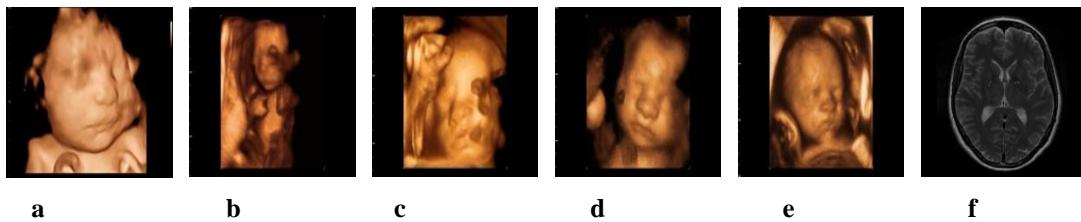


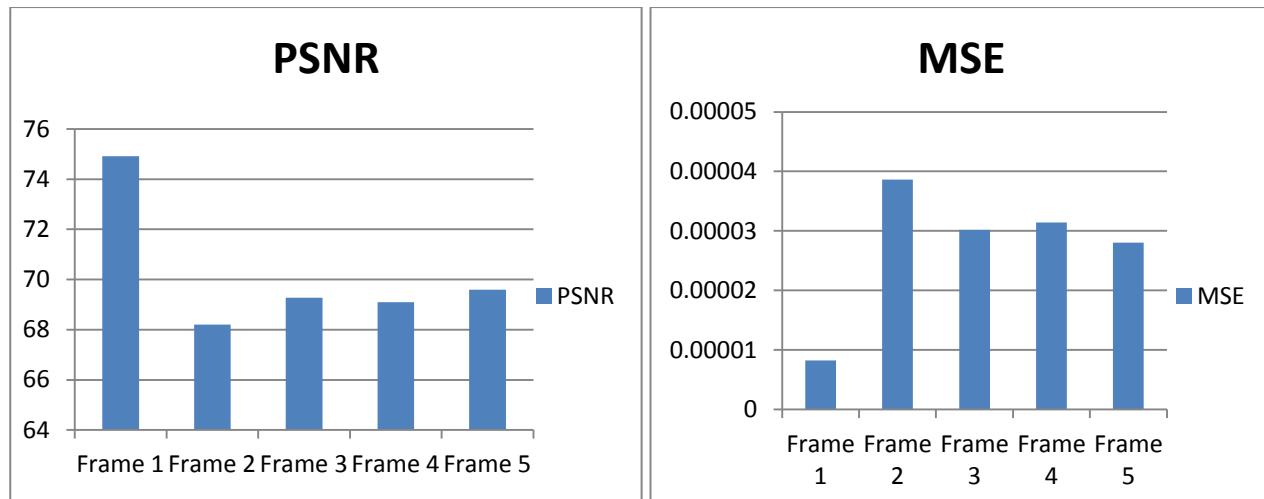
Fig. 3: Cover images (a) frame 1, (b) frame 2, (c) frame 3, (d) frame 4, (e) frame 5, and secret image (f)

**Peak Signal to Noise Ratio:** Peak signal-to-noise ratio, often abbreviated PSNR, is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. PSNR value is drawn by calculating the formula on the extracted image and original secret image. The PSNR values testify the quality of the images produced before and after the algorithm processing. Higher PSNR shows the better quality of the results as shown in table 1 and fig 4.

**Mean squared error (MSE):** The total squared error between the two data matrices is known as the mean squared error, which is capable of showing the error with the positive value and can be easily compared with the other results obtained from the other image matrices. The mean squared error has been recorded between the 0.000008 and 0.00003900 which is considerably very low and lowers the chances of detection of the embedding in the cover data as per shown in the table 1 and fig 5.

	PSNR	MSE
Frame 1	74.90798	0.00000824
Frame 2	68.19689	0.00003862
Frame 3	69.26749	0.00003019
Frame 4	69.09568	0.00003140
Frame 5	69.59123	0.00002802

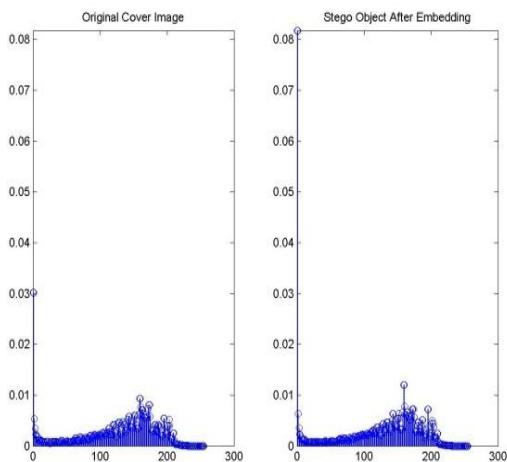
Table 1: Result analysis showing PSNR and MSE Values



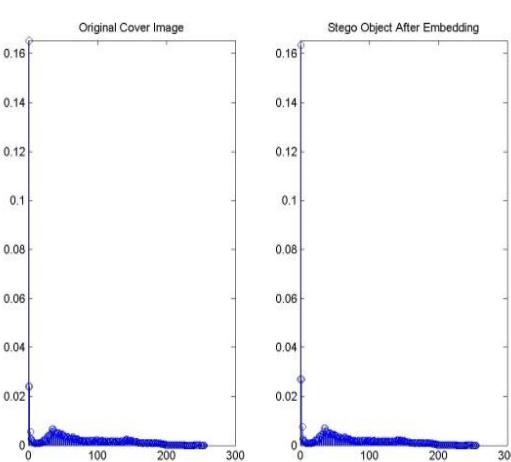
**Fig. 4: Graphs showing PSNR and MSE Values**

**Steganalysis Attack based Analysis:** The proposed model results have been obtained in the form of the histogram method for the embedding detection. The proposed model has been tested over the 3d ultrasound images obtained from the 3d ultrasound video. The following image has undergone the steganalysis attack based histogram method evaluation for the detection of the embedding in the given cover data. Once the embedding is detected, there are several kinds of the attacks to extract the data from the stego object in order to reveal the data unethically after the completion of the steganography based embedding. The proposed model has been made efficient, which is clearly visible from the results.

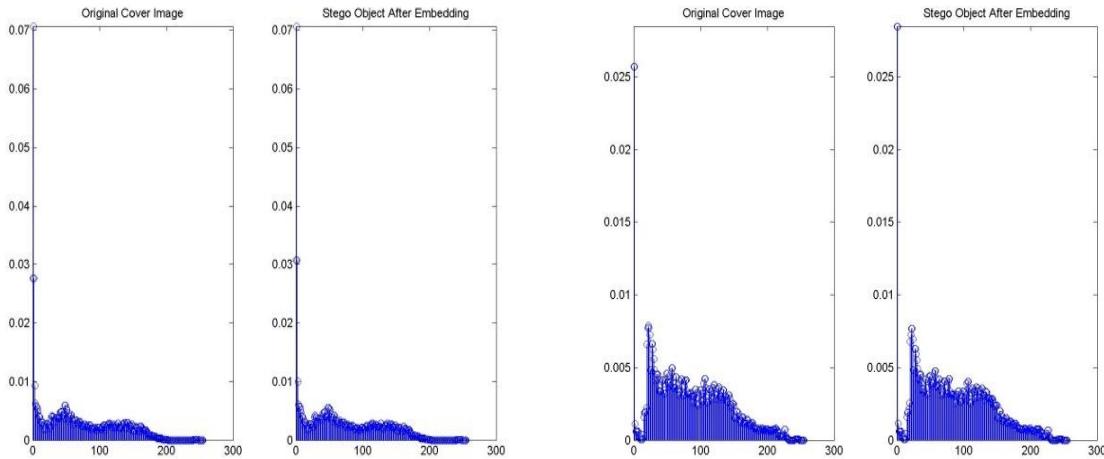
The chi-square method is used for the analysis of the histogram based method to reveal the chances of the embedding in the image data. The figures below show almost no difference between the two images, cover image and the stego image, which are analyzed under the chi-square method in the figures below. The least difference or almost no dissimilarity between the two histograms shows the extremely lower probability of the steganographic method being detected during the analysis.



**Fig. 5: Histogram comparison of frame 1**

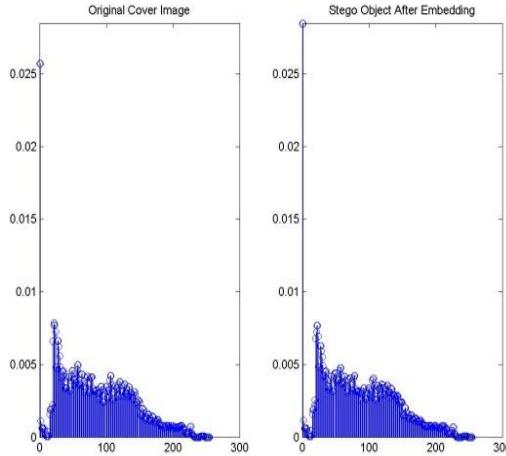


**Fig. 6: Histogram comparison of frame 2**



**Fig. 7: Histogram comparison of frame 3**

**Fig. 8: Histogram comparison of frame 4**



**Fig. 9: Histogram comparison of frame 5**

## V. CONCLUSION

The planned model primarily based upon the multiple rule based embedding within the cowl knowledge for the upper embedding security against the steganalysis attacks. This model focuses upon the compatibility verification between the quilt image and therefore the secret image so as to search out the labile match for embedding. The planned model has been guided to defer the embedding method if the compatibility isn't actually matched and doesn't satisfy the given set of thresholds. This procedure is accountable for the strong analysis of the security embedding by adaptive and flexible cover image based embedding method for the robustness of the embedding security. This model will increase the extent of security that is clearly indicated from the mean squared error and bits per pixel based mostly parameters underneath the result analysis sections. Many experimental results are conducted so as to judge the system performance, which may be clearly foresaid because the strong embedding possibility because it clearly decreases the chance the info disclosure attacks over the stego objects.

## VI. SCOPE FOR FUTURE WORK

In the future, the proposed model can be enhanced by using the high order randomized pattern in order to realize the robust pattern based embedding model. The future improvements can utilize the travelling sales man algorithm to sort the pixels in the perfect order in order to embed the data in the cover image. In future modern cryptography techniques like AES, DES, etc. can be implemented to provide more security. The secret image and secret text can be embedded simultaneously in future.

## REFERENCES

- [1] Mritha Ramalingam, and Nor Ashidi Mat Isa. "A steganography approach for sequential data encoding and decoding in video images." In Computer, Control, Informatics and Its Applications (IC3INA), 2014 International Conference on, pp. 120-125. IEEE, 2014.
- [2] Pooja Yadav, Nishchol Mishra, and Sanjeev Sharma." A Secure Video Steganography with Encryption Based on LSB Technique."International Conference on Computational Intelligence and Computing Research. IEEE, 2013.
- [3] A.K. Jain, M.N. Murty And P.J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, September 1999.
- [4] Prof. Samir Kumar, Bandyo padhyay Barnali, Gupta Banik, "LSB Modification and Phase Encoding Technique of Audio Steganography Revisited", International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 4, June 2012,ISSN : 2278 – 1021.
- [5] Shawn D. Dickman," An Overview of Steganography", JMU-INFOSEC-TR-2007-002.
- [6] Soumyendu Das, Subhendu Das, Bijoy Bandyopadhyay and Sugata Sanyal, "Steganography and Steganalysis: Different Approaches".
- [7] Arvind Kumar and Km. Pooja, "Steganography- A Data Hiding Technique", International Journal of Computer Applications (0975 – 8887) Volume 9– No.7, November 2010.
- [8] Deborah A. Whitiak, "The Art of Steganography".
- [9] Jagvinder Kaur and Sanjeev Kumar," Study and Analysis of Various Image
- [10] Steganography Techniques", IJCST Vol. 2, Issue 3, September 2011, I S S N : 2 2 2 9 - 4 3 3 3 .
- [11] L. Y. POR, B. Delina, "Information Hiding: A New Approach in Text Steganography", 7th WSEAS Int. Conf. on APPLIED COMPUTER & APPLIED COMPUTATIONAL SCIENCE (ACACOS '08), Hangzhou, China, April 6-8, 2008, ISSN: 1790-5117.
- [12] Jonathan Cummins, Patrick Diskin, Samuel Lau and Robert Parlett, " Steganography And
- [13] Digital Watermarking", School of Computer Science, The University of Birmingham.
- [14] Mrs. Kavitha, Kavita Kadam, Ashwini Koshti, Priya Dunghav,"
- [15] Steganography Using Least Significant Bit Algorithm", International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, May-Jun 2012, pp. 338-341, ISSN: 2248-9622.
- [16] Jayaram P, Ranganatha H R, Anupama H S, "INFORMATION HIDING USING AUDIO
- [17] STEGANOGRAPHY – A SURVEY", The International Journal of Multimedia & Its Applications (IJMA) Vol.3, No.3, August 2011, DOI : 10.5121/ijma.2011.3308.
- [18] Hsien-Chu Wu, Hui-Chuan Lin and Chin-Chen Chang, "Reversible Palette Image Steganography Based on De-clustered and Predictive Coding".
- [19] David Houque, "INTRODUCTION TO MATLAB FOR ENGINEERING STUDENTS", version 1.2, August 2005.S. Thilagamani1 and N. Shanthi, "A Survey on Image Segmentation Through Clustering",
- [20] International Journal of Research and Reviews in Information Sciences Vol. 1, No. 1, March 2011.
- [21] Nagham Hamid, Abid Yahya, R. Badlishah Ahmad and Osamah M. Al-Qershi, "Image Steganography Techniques: An Overview".

# A REVIEW ON IMAGE ENHANCEMENT USING INTELLIGENT TRANSPORTATION SYSTEM

Jagdev Singh

Student of M.tech(C.S.E.)

YCOE, Talwandi sabo

[gogachauhan047@gmail.com](mailto:gogachauhan047@gmail.com)

Mr Ashok bathla

Assistance Professor(C.S.E)

YCOE, Talwandi sabo

[ashokashok81@gmail.com](mailto:ashokashok81@gmail.com)

**Abstract:** Enhancement plays an important role in digital image processing. The visibility of images of open air road scenes becomes degraded when captured in severe climate conditions. Drivers frequently turn on the headlights of their vehicles and streetlights are regularly actuated, bringing about confined light sources in images of road scenes in these conditions. Furthermore, dust storms are additionally climate occasions that are regularly experienced when driving in a few districts. A novel and effective haze removal approach to cure these issues created by confined light sources and color shifts, which subsequently accomplishes better restoration results for single hazy images. Poor visibility degrades quality and performance of computer vision algorithms for smart transportation frameworks, for example, traveling vehicle data recorders and traffic surveillance systems, activity observation frameworks which must operate under a wide range of weather conditions. Another issue is that the captured foggy road scene images contains confined light sources or shading movement issues because of dust storm conditions. Movement detection and Darkness are also problems in the captured images. The goal of this work is to enhance the road scene images using different filters and enhancement techniques. The distinctive sorts of parameters are figured that are PSNR, MD, MSE and Processing Speeds.

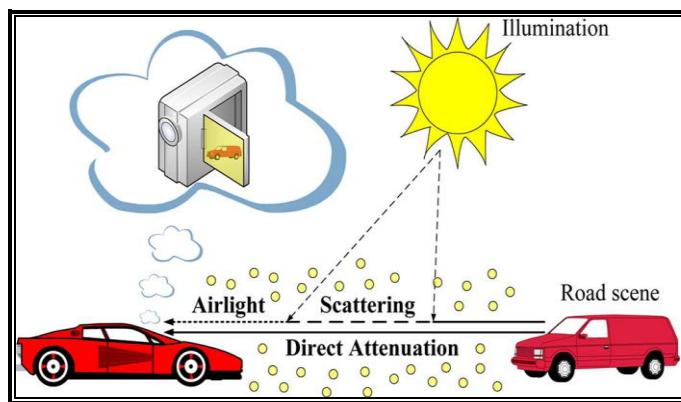
**Keywords:** PSNR, MD, images, HE, BBHE, DPC, RSWHE.

## I. INTRODUCTION

Visibility in road images can be debased because of normal air conditions, for example, dimness, mist, and dust storms. This visibility corruption is because of the ingestion and dissipating of light by barometrical particles. Road image debasement can bring about issues for wise transportation frameworks, for example, voyaging vehicle information recorders and movement observation frameworks, which must work under an extensive variety of climate conditions [1]. The amount of absorption and scattering depends on the depth of the scene between a traffic camera and a scene point; therefore, scene depth information is important for recovering scene radiance in images of hazy environments. Movement scene characterization is a developing subject with extensive significance in the field of wise transportation frameworks [6]. With the expanded accessibility of cameras in vehicles (either on cell phones or as inserted equipment in sumptuous auto models) there are more potential outcomes for improving basic wise transportation undertakings. Fleet management systems are used to track the status of fleets of vehicles belonging to various kinds of companies e.g. taxi, delivery, cargo transport [8]. They utilize GPS sensors to track the vehicle's area, yet have little data about the vehicle's surroundings [5]. Some useful information about the vehicle's surroundings can be inferred by using a camera to record images from the driver's perspective, and then solving a classification problem to detect interesting types of traffic scenes and scenarios [7]. For example, this approach can be used to identify traffic jams, or to differentiate open road environments from urban/rural roads or tunnels [2].

Image classification in general is a common topic in computer vision, extensively researched in great number of papers. Active research focuses mainly on recognizing images in a large number of diverse classes [1]. This enables a simple and meaningful comparison of state-of-the-art methods applied on various domains. However, the scene radiance recovered via the dark-channel-prior-based techniques is usually accompanied by the generation of serious artifacts when the captured hazy road image contains localized light sources or color-shift problems due to sandstorm conditions [10]. This can be problematic for many common road scenarios. For example, in inclement weather conditions, the drivers generally turn on headlights when they are driving in order to improve visual perception, and streetlamps are lit for similar reasons [9]. The techniques based on the dark channel prior cannot produce satisfactory restoration results when presented with these situations.

A novel haze removal approach by which to avoid the generation of serious artifacts by the conjunctive utilization of the proposed hybrid dark channel prior (HDCP) module, color analysis (CA) module, visibility recovery (VR) module, Histogram Equalization(HE), Recursive sub weighing Histogram Equalization(RSWHE) . The proposed technique can effectively conceal localized light sources and restrain the formation of color shifts when the captured road image contains localized light sources or color-shift problems.



**Fig. 1 Pictorial description of hazy image acquisition via the optical model [1]**

## II. LITERATURE SURVEY

Previous Research work on “**Road Scenes Images Enhancement**” in the literature survey was studied which helped to complete work and enhance knowledge. On the basis of studying various papers some of them are given below:

**Shih-Chia Huang et.al [2014]** have presented the visibility of images of outdoor road scenes will generally become degraded when captured during inclement weather conditions. Drivers often turn on the headlights of their vehicles and streetlights are often activated, resulting in localized light sources in images capturing road scenes in these conditions. Additionally, sandstorms are also weather events that are commonly encountered when driving in some regions. In sandstorms, atmospheric sand has a propensity to irregularly absorb specific portions of a spectrum, thereby causing color-shift problems in the captured image. Traditional state-of-the-art restoration techniques are unable to effectively cope with these hazy road images that feature localized light sources or color-shift problems. In response, a novel and effective haze removal approach to remedy problems caused by localized light sources and color shifts, which thereby achieves superior restoration results for single hazy images. The performance of the proposed method has been proven through quantitative and qualitative evaluations. [1].

**Giri Nandan et.al [2014]** In this paper various enhancement techniques are discussed and compared. Various methods for image resolution enhancement had been discussed which shows we can enhance the images on color scale by using different techniques nowadays. Different areas in which image enhancement can be used are compared in this paper. We will discuss the methods which can enhance the resolution of MR images, images taken by regular cameras, Built-in camera image of a Mobile phone, vehicle camera images and an aerial image. [6]

**Ashamdeep Singh et.al [2013]** Digital images are the most common application of now day’s world. In almost every era of life and technology, the digital images are playing their roles. The problem with images is that, their quality depends on a number of other factors like lighting at the image capturing location, proficiency of the operator, and noise. A lot of techniques have been suggested earlier for the enhancement of the color images which works on histogram of the image or on some particular region. Region based techniques using texture analysis are simple and more effective as they work according to the specified regions of the image. Seed selection is an optimal method for initiate any spatial enhancement. This paper suggests a new hybrid approach for enhancement of the digital images. The suggested technique is based on region growing segmentation and works adaptively for enhancement of the image. Further, the technique is seed dependent so selection of seed is very important in this algorithm. A seed chosen in darker regions will give better results than the seed chosen in brighter region, because it is assumed that user will require enhancing the darker portions of the image. In this paper the process of color image enhancement uses three modules. Initial seed selection is our first module. Our second module is region growing it is used to segment the image based on seed regions. The third and last module is region merging and used morphological operations as texture analysis.

**Ajay Raghavan et.al [2012]** have presented Unattended camera devices are increasingly being used in various intelligent transportation systems (ITS) for applications such as surveillance, toll collection, and photo enforcement. In these fielded systems, a variety of factors can cause camera obstructions and persistent view changes that may adversely affect their performance. Examples include camera misalignment, intentional blockage resulting from vandalism, and natural elements causing obstruction, such as foliage growing into the scene and ice forming on the porthole. In addition, other persistent view changes resulting from new scene elements of interest being captured, such as stalled cars, suspicious packages, etc. might warrant alarms. Since these systems are often unattended, it is often important to automatically detect such incidents early [3].

**Fan-Chieh Cheng et.al [2011]** have proposed a novel background subtraction approach in order to accurately detect moving objects. The method involves three important proposed modules: a block alarm module, a background modeling module, and an object extraction module. The block alarm module efficiently checks each block for the presence of either a moving object or background information. This is accomplished by using temporal differencing pixels of the Laplacian distribution model and allows the subsequent background modeling module to process only those blocks that were found to contain background pixels. Next, the background modeling module is employed in order to generate a high-quality adaptive background model using a unique two-stage training procedure and a novel mechanism for recognizing changes in illumination. The overall results show that our proposed Method attains a substantially higher degree of efficacy, outperforming other state-of-the-art methods by Similarity and *F1* accuracy rates of up to 35.50% and 26.09%, respectively. [4].

**Ivan et.al [2010]** has studied work deals with multi-label classification of traffic scene images. We introduce a novel labeling scheme for the traffic scene dataset FM2. Each image in the dataset is assigned up to five labels: settlement, road, tunnel, traffic and overpass. They propose representing the images with (i) bag-of-words and (ii) GIST descriptors[2]. The bag-of-words model detects SIFT features in training images, clusters them to form visual words, and then represents each image as a histogram of visual words. On the other hand, the GIST descriptor represents an image by capturing perceptual features meaningful to a human observer, such as naturalness, openness, roughness, etc. compare the two representations by measuring classification They report good classification results for easier class labels (road,  $F1 = 98\%$  and tunnel,  $F1 = 94\%$ ), and discuss weaker results (overpass,  $F1 < 50\%$ ) that call for use of more advanced methods. [2].

### III. IMAGE ENHANCEMENT TECHNIQUES

#### i) Optical Model:

In computer vision and pattern analysis, the optical model is widely used to describe the digital camera information of a hazy image under realistic atmospheric conditions in the RGB color space as [1]:

$$I^c(x, y) = J^c(x, y)t(x, y) + A^c(1 - t(x, y)) \quad (1) [1]$$

where  $c \in \{r, g, b\}$ ,  $I^c(x, y)$  represents the captured image,  $J^c(x, y)$  represents the scene radiance that is the ideal haze free image,  $A^c$  represents the atmospheric light, and  $t(x, y)$  represents the transmission map describing the portion of the light that arrives at a digital camera without scattering [1]. The first term of (1), i.e.,  $J^c(x, y)t(x, y)$ , represents the direct attenuation describing the decayed scene radiance in the medium. The second term of (1), i.e.,  $A^c(1 - t(x, y))$ , represents the airlight that resulted from the scattered light and leading to the color shifting in the scene.

#### ii) Dark Prior Channel Technique:

The dark prior channel technique [3] can work well for haze removal in single images that lack localized light sources. However, haze removal by the dark channel prior technique [2] usually results in a seriously underexposed image when the captured scene features localized light source. Correct atmospheric light, the use of a large local patch will result in invariable transmission and thereby leads to the generation of halo effects in the recovered image [11]. In contrast, when the dark channel prior technique [1] uses a small patch size, the recovered image will not exhibit halo effects. In other words, the minimum intensity in such a patch should has a very low

$$J^{dark}(x) = \min_{c \in \{r, g, b\}} (\min_{y \in \Omega(x)} (J^c(y)))$$

value. Formally, for an image  $J$ , we define [1]:

where  $J_c$  is a color channel of  $J$  and  $\Omega(x)$  is a local patch centered at  $x$ . Our observation says that except for the sky region, the intensity of  $J^{dark}$  is low and tends to be zero, if  $J$  is a haze-free outdoor image.

#### iii) Hybrid Dark Channel Prior Technique:

HDCP module can produce a restored image that is not underexposed by using a procedure based on the dark channel prior technique [3]. HDCP technique for haze removal in single images efficiently conceals localized light sources and, consequently, accurately estimates the position of the atmospheric light.

However, localized light will be misjudged as atmospheric light. Hence, we present the HDCP module that ensures correct atmospheric light estimation and the subsequent avoidance of halo effects during the haze removal of single images based on the hybrid dark channel prior technique [13]. This technique will be introduced in the following. To effectively estimate the density of the haze featured by an image, we combine the advantages of small and large patch sizes via different weights [10]. In addition, we use the large patch size to acquire the correct atmospheric light during the implementation of the hybrid dark channel prior technique. HDCP module can provide effective transmission map estimation and thereby avoids the production of artifact effects in the restored image [9].

**iv) Color Analysis Model:**

The particles of sand in the atmosphere caused by sandstorms absorb specific portions of the color spectrum [8]. This phenomenon leads to color shifts in images captured during such conditions, resulting in different color channel distributions [15]. To recover from scene radiance problem, we propose the CA module that is based on the gray world assumption. The gray world assumption relies on the notion that average intensities should be equal in each RGB color channel for a typical image [14].

**v) Visibility Recovery Technique:**

The VR module combines the information obtained by the HDCP and CA modules to avoid the generation of serious artifact effects and thus obtain a high-quality haze-free image regardless of weather conditions [3]. It is used to remove haze, fog, mist [12]. It uses visibility restorer to improve the visibility of the input image.

**vi) Histogram Equalization Technique:**

HE is most widely used method for image contrast enhancement.

It uses images cumulative distributive function to improve image contrast. But it has a problem of mean shift in which the mean brightness of input image is different from output image. According to [6], HE introduces two types of artifacts in which over-enhancement is done for the image with more frequent gray levels; and loss of contrast for the image regions with less frequent gray-levels. Consider the input image  $\mathbf{X}$ . Based on the histogram  $H(\mathbf{X})$ , the probability density function (PDF) of the image is calculated and it finds cumulative distributive function (CDF)[14].

**vi) Brightness-Preserving Bi Histogram Equalization:**

BBHE copes up with the mean shift problem encountered in histogram equalization works by segmenting the input histogram into two sub-histograms. BBHE first segments the input histogram and then executes each segmented sub-histogram independently. During segmentation it determines threshold for histogram segmentation in order to minimize the brightness difference between the input image and output image[14].

BBHE first decomposes the input histogram  $H(\mathbf{X})$  into two sub-histograms  $HL(\mathbf{X})$  and  $HU(\mathbf{X})$  by using the input mean  $XM$ , where  $HL(\mathbf{X})$  is associated with the gray levels  $\{X_0, X_1, \dots, XM\}$  and  $HU(\mathbf{X})$  is associated with the gray levels  $\{XM+1, XM+2, \dots, XL-1\}$ [14]. Then it performs conventional histogram equalization on  $HL(\mathbf{X})$  and  $HU(\mathbf{X})$  independently.

**vi) Recursively Separated and Weighted Histogram Equalization:**

It enhances image contrast and also preserves image brightness. This technique covers up the mean shift problem of histogram equalization and is an extension of BBHE. Although BBHE carries out the mean based histogram segmentation only once but RSWHE segments both mean and median based histogram equalization more than once recursively. RSWHE changes the input histogram before running the equalization procedure. This is

the difference between the previous methods and RSWHE. It divides input histogram into two or more than two sub-histograms recursively upto a specified recursion value  $r$  and creates upto  $2r$  sub-histograms. The resultant sub-histograms are then equalized individually[14]. The histogram segmentation module takes the input image  $\mathbf{X}$ , computes the input histogram  $H(\mathbf{X})$  and the histogram weighting module modifies the sub-histograms by using a normalized power law function. Lastly, the histogram equalization module runs histogram equalization individually over each of the modified sub-histograms[14].

#### IV. CONCLUSION

Different enhancement techniques for enhancing road scene images is described in the paper. Image enhancement of road images improves the interoperability or perception in images of road. Diverse issues emerge when the captured foggy road scene images contains restricted light sources or shading movement issues because of dust storm conditions, sunlight, fog or extensive variety of climate conditions. Motion Detection is also among one of the issue areas. Future extension can be to enhance the Road scene images using various methods. The distinctive sorts of parameters are computed that is PSNR, MSE and AMBE and to analyze the results being obtained.

#### REFERENCES

- [1] Shih-Chia Huang, Bo-Hao Chen, Yi-Jui Cheng, "An Efficient Visibility Enhancement Algorithm for Road Scenes Captured by Intelligent Transportation Systems" IEEE Transactions On Intelligent Transportation Systems, Vol. 15, No. 5, pp.2321-2332, October 2014.
- [2] Ivan et.al, "Multi-Label Classification of Traffic Scenes" Proceedings of the Croatian Computer Vision Workshop, pp. 9-14, September 16, 2014.
- [3] Ajay Raghavan, Robert Price, Juan Liu, "Detection of Scene Obstructions and Persistent View Changes in Transportation Camera Systems" 15th International IEEE Conference on Intelligent Transportation Systems Anchorage, Alaska, USA, pp. 957-962, September 2012.
- [4] Fan-Chieh Cheng, Shanq-Jang Ruan, "Accurate Motion Detection Using a Self-Adaptive Background Matching Framework" IEEE Transactions on Intelligent Transportation Systems, Vol. 13, No. 2, pp. 671-679, June 2012.

- [5] J.-E. Ha, W.-H. Lee, "Foreground objects detection using multiple difference images" Opt. Eng., Vol. 49, no. 4, pp. 047-201, April 2010.
- [6] Giri Nandan, "Image Resolution Enhancement Methods for Different Applications" International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, no.17, pp. 1733-1738, 2014.
- [7] C. Stauffer, W.E.L Grimson, "Adaptive background mixture models for real-time tracking" in Proc. IEEE Computing Vision And Pattern Recognition, 1999, Vol. 2, pp. 246–252, June 1999.
- [8] A. Doshi et.al, "Smoothing of optical flow using robustified diffusion kernels" Image Vis. Computing, Vol. 28, no. 12, pp. 1575–1589, Dec. 2010.
- [9] Gokilavani Chinnasamy, Gowtham M., "A Novel Approach for enhancing foggy images" International Journal of Emerging Technology and Advanced Engineering, Vol.4, no.10, pp.88-91, October 2014.
- [10] Bo Li, Shuhang Wang, Jin Zheng, Liping Zheng, "Single image haze removal using content adaptive dark channel and post enhancement" IET Comput. Vis., Vol. 8, no.2, pp. 131-140, July 2013.
- [11] Tae Ho Kil, Sang Hwa Lee,Nam Ik Cho, "A Dehazing algorithm using dark channel prior and contrast enhancement" IEEE Transactions On Intelligent Transportation Systems, pp. 2484-2487,May 2013.
- [12] Harpoonamdeep Kaur, Dr. Rajiv Mahajan, "A Review on various visibility restoration techniques" International Journal of Advanced Research in Computer and Comm. Engg.,Vol. 3,no.5, pp. 6622-6625, May 2014.
- [13] Yadwinder Singh, Er.Rajan Goyal, "Haze removal in color images using Hybrid Dark Channel Prior and Bilateral Filter" International Journal on Recent and Innovation trends in Comp. and Comm., Vol. 2, no. 12, pp. 4165-4171, December 2014.
- [14] Mary Kim, Min Gyo Chung, "Recursively separated Weighing Histogram Equalizaton for brightness preservation and contrast enhancement", Vol. 54, no.3, August 2008.
- [15] Nancy, Er.Sumandeep Kaur, "Image Enhancement Techniques: A Selected Review" IOSR Journal of Computer Engg., Vol. 9, no. 6, pp. 84-88, April 2013.
- [16] Ashamdeep Singh, "Spatial Image Enhancement of Color Images Using Texture Analysis" International Journal of Application or Innovation in Engineering & Management Volume 2, Issue 6, June 2013.

# Comparative Analysis of OLSR and GRP Routing Protocols in Mobile Ad-Hoc Networks

Ravneet Singh Sahota<sup>1</sup>, Madan Lal<sup>2</sup>

<sup>1</sup> M.Tech Student , Department of Computer engineering, Punjabi University Patiala ,India

<sup>2</sup> Assistant Professor, Department of Computer engineering, Punjabi University ,India

## *Abstract:*

Wireless Technology is at its pinnacle. This field of MANET'S has become a hub of invention of latest theories and structures. Mobile Ad-hoc Network may be a special purpose of focus for the researchers. Mobile Ad-Hoc Networks (MANETs) are autonomous and decentralized wireless systems. Mobile Ad hoc Network may be a agglomeration of mobile nodes within which the wireless links are often broken because of mobility and dynamic infrastructure. Routing may be a broader issue and challenge in Mobile ad hoc networks. The main categorization of MANET routing protocols are Proactive, Reactive and Hybrid. Formulating this paper our objective is to compare performance of various routing protocol in particular Geographical-based Routing Protocol (GRP) which is a hybrid protocol and Optimized Link State Protocol (OLSR) which is a proactive routing protocol, whereby routes mostly discovered and updated frequently and accessible on demand .The simulation results showed by various researchers has been reviewed and comparison of these routing protocols has been presented on the basis of performance parameters such as throughput , delay , load , etc .

**Keywords**— MANET, GRP, OLSR, Routing Protocols

## 1. INTRODUCTION

Network structure is changing quickly in recent years. The only network accessible four decades past was wired network. The emergence of wireless networks has gone an extended manner in resolving the growing service demands. The main target of analysis and development has nearly shifted from wired networks to wireless networks. The limitations of wireless network techniques like high error rate, power restrictions, constraints of bandwidth ,etc. has not deterred the expansion of wireless networks[1] MANET may be a wireless network that transmits from pc to pc rather than employing a central base station (Access Point) to which all computers should communicate, this peer-to-peer mode of operation will greatly extend the gap of the wireless network and to achieve access to the web, one amongst the computers will be connected via wire or wireless. MANETs basically classified into 3 classes on the premise of route discovery i.e. Reactive additionally known as on-demand routing protocol . Another technique is proactive additionally called table-driven protocol and another technique is Hybrid protocol.[2] These classification of routing protocols is completed on the premise of network organization as flat primarily based, stratified based and site based techniques. In flat primarily based protocol all the nodes being equal i.e. they play constant role within the network. In stratified protocol totally different nodes play different roles i.e. during this totally different cluster heads being chosen among cluster members. In location primarily based protocol nodes rely on the placement fact and use this fact for communication [3]. This study intends to enhance the performance of OLSR, GRP and ABR routing protocols by performance standardization of those protocols [4].We have reviewed many key literatures within the field of MANET'S, the routing protocols that highlight existing protocols. Manet routing protocol should be equipped to handle the dynamic and unpredictable topology changes related to mobile nodes [5].While conjointly being responsive to the restricted wireless bandwidth and device power concerns which can lead to reductions in transmission vary or output.[6]. Manet routing protocols ought to even be decentralized and self-organizing and ready to exploit multi-hops and load balancing, these needs ensure Manet routing protocols has ability to control and operate autonomously[7].

## 2. ROUTING PROTOCOL's

There are numerous routing protocols for MANETs. Therefore we intends to review the performance of built-in routing protocols and also tuning or optimizing of these protocols is necessary. The main technique for evaluating the performance of MANETs is by employing a simulator. In this paper, we've reviewed the performance of AODV, DSR, OLSR and GRP routing protocols. The performance is supported by varying range of nodes and analysis is performed by means that of Route latency[8], Organisation of network, Route availability, Control traffic, Routing information, Topology propagation, Overhead in communication

### 2.1 Optimized Link State Routing

Optimized Link State Protocol (OLSR) could be a recognized as proactive routing protocol. OLSR is enhancement and extension of a pure link state protocol. The topological be different by causes the flooding of the nodes data to each accessible nodes within

the set-up[9]. To decrease the promising overhead within the network protocol makes use of Multipoint Relays[10]. The theme of MPR is to decrease flooding of broadcasts by reducing the similar broadcast during a few regions within the network. A node senses and selects its MPR's with control messages called HELLO messages. Hello messages are used to ensure a bidirectional link with the neighbor. HELLO messages are sent at a certain interval. Nodes broadcast "TC" or Topology control messages to determine its MPR's.[3]

## **2.2 Geographic Routing Protocol**

GRP may be a position based protocol classified as hybrid routing protocol. In GRP the geographical Positioning System (GPS) is employed to mark the placement of node and also the quadrants optimize flooding. Once a node moves and crosses neighborhood then the flooding position is updated [7]. The neighbor's and their positions are known by the exchange of "Hello" protocol. The idea of route protection ensures that a node will come its packet to consequent node. Once packets make the nodes destination, then End host broadcast a network and data packet to its hosts. The Transmitting node computes the foremost glorious route in accordance with composed Information also as at that time instantly starts to send Data packets [11]. GRP provides a high-quality structure which might work on constant time with the strength of reactive routing protocol and proactive routing protocol that gathers network data at a transmitting node employing a very little organize overheads[12]. The promising routes is equipped by initial nodes on the basis of the gathered data and it continue transmits knowledge packets although the present route could also be disconnected, it ends up in achieving quick packet transfer delay Without unduly compromising on Control overhead performance

## **2.3 AODV: Ad hoc on Demand Distance Vector**

AODV routing protocol be a reactive routing protocol that establish a route when a node need to transmit data packets. AODV is capable of each unicast and multicast routing. The operation of the protocol is split in 2 functions: route discovery and route maintenance[13]. Once a route is required to some destination, the protocol starts route discovery. Then the initial node sends route request message to its neighbor's. And if those nodes don't have any info concerning the destination node, they're going to send the message to any or all its neighbor's and then on. And if any neighbor node has the data concerning the destination node, the node sends route reply message to the route request message initiating node[14]. On the idea of this method a path is recorded within the intermediate nodes. This path identifies the route and is named the reverse path. Since every node forwards route request message to any or all of its neighbor's, multiple one copy of the initial route request message will reach a node. a unique id is chosen, once a route request message is formed. when a node received, it'll check this id and also the address of the instigator and discarded the message if it had already processed that request. Node that has info concerning path to the destination sends route reply message to the neighbor from that it's received route request message. This neighbor will possible. Owing the reverse path it may be possible. Then the route reply message travels back by using reverse path. once a route reply message reaches the initiator the route is prepared and also the initiator node will begin sending information packets[10]

## **2.4 Dynamic Source Routing (DSR)**

DSR is also a reactive routing protocol which based on concept of source routing. In this type of routing the sender knows information about hop-to-hop route to the destination. All the routes are maintained in the route cache. When a node attempts to transmits a data packet to a destination for which it is not aware of that route. In DSR every node maintain a route cache along-with route entries which being updated continuously and route learns about new routes[15]. The advantage of DSR is that no periodic routing packets is required.. Unlike other protocols The sender of the packets chooses and controls the route used for own packets. All routes used are guaranteed to be free of loops as the sender avoids duplicate hops in the chosen routes[16].

## **3. COMPARISON OF PROTOCOLS-**

The comparison of various categories of the protocols has been reviewed [17]and on basis of the reviewing we have mentioned the advantages and disadvantages of various protocols in Table 1.

### **3.1 Comparison on the basis of its categorical type-**

Table 1 depicts the advantages and disadvantages of Proactive, Reactive, Hybrid routing protocols.

Protocol	Advantages	Disadvantages
<b>Proactive</b>	Information is always available when needed. Latency is retrench in the network.	Overhead is high, Routing information is usually flooded in the network
<b>Reactive</b>	Path also being available when needed but overhead is low and it is free from loops.	Latency is enhanced in the Network
<b>Hybrid</b>	Suitable for large and complex networks and information is available up to date	Complexity increases

**TABLE- 1- Categorical comparison**

### **3.2 Comparison on the basis of parameters-**

Table 2 depicts the comparison on the basis of parameters i.e Proactive , Reactive , Hybrid routing protocols. The parameters have been reviewed and compared on the basis of its accessibility[18]. The parameters such as Route latency, Organisation of network, Route availability, Control traffic, Routing information, Topology propagation , Overhead in communication are represented in the table.

PARAMETERS	PRO-ACTIVE	RE-ACTIVE	HYBRID
Route latency	Available	Available (when needed)	Both
Organization of network	Flat/ hierachal	Flat	Flat hierachal
Route availability	Available	Computed(when needed)	Both
Control traffic	High	Low	Lower from both
Routing information	Always stored in routing table	It does not store	Depends on requirements as needed
Topology propagation	Periodic	On-demand	Both
Overhead in communication	High	Low	Medium

**TABLE- 2 - Parametrized comparison**

### **4. CONCLUSION**

In this paper, the researchers has mentioned the assortment of routing protocols in Mobile Ad-Hoc networks and gave comparisons between them. The protocols are sub categorized into 3 main categories: (i) Source-Initiated (Reactive or On-demand), (ii) Table-Driven (Pro-active), (iii) Hybrid protocols. Thus for each of these categories, it tend to have reviewed and compared many protocols. Whereas some measures still poses several challenges facing Mobile Ad-Hoc networks associated with Routing and Security. Every routing protocol has distinctive options supporting the network environments among them we've to decide on the acceptable routing protocol. The analysis of the various proposals has reflected that the inherent characteristics of Mobile Ad-Hoc networks, like lack of infrastructure and speedily dynamic topologies, introduce extra difficulties to the already difficult drawback of secure routing. The most differentiating issue between the protocols is that the ways in which of finding and maintaining the routes between source destination pairs. The comparison shown between these routing protocols indicates that the planning of a secure Mobile routing protocol constitutes a difficult analysis drawback against the present security solutions. We tend to hope that the taxonomy given during this paper is useful and supply researchers a platform for selecting the correct protocol for their work. Finally this paper provides the general characteristic of all routing protocols and represented that protocols which might perform best in giant networks. Almost all the protocols we tend to mentioned during this paper have their own characteristic options and performance parameter wherever they out perform their competitors. Still Mobile Ad-Hoc networks have display a good challenge for the researchers because of dynamic topology and security attacks, and none of the protocols is totally secured and analysis are done all round the globe.

## REFERENCES

- [1] G. S. Aujla and S. S. Kang, "Comprehensive Evaluation of AODV , DSR , GRP , OLSR and TORA Routing Protocols with varying number of nodes and traffic applications over MANETs," *IOSR J. Comput. Eng.*, vol. 9, no. 3, pp. 54–61, 2013.
- [2] R. Devi, B. Sumathi, T. Gandhimathi, and G. Alaiyariasi, "Performance Metrics of MANET in Multi-Hop Wireless Ad-Hoc Network Routing Protocols," *Int. J. Comput. Eng. Res.*, pp. 179–184, 2010.
- [3] Y. C. Huang, S. Bhatti, and D. Parker, "Tuning olsr," *17th Annu. IEEE Int. Symp. Pers. Indoor Mob. Radio Commun.*, vol. 3, pp. 33–38, 2006.
- [4] A. K. Gupta, H. Sadawarti, and A. K. Verma, "Review of Various Routing Protocols for MANETs," *Int. J. Futur. Comput. Commun.*, vol. 1, no. 3, 2011.
- [5] "2012 2nd IEEE International Conference on Parallel , Distributed and Grid Computing Metrics Improvement ofMANET Using Reactive Protocols Approach Metrics Improvement ofMANET Using Reactive Protocols Approach," *2012 2nd IEEE Int. Conf. Parallel, Distrib. Grid Comput. Metrics*, no. March 2016, 2013.
- [6] A. Hinds, M. Ngulube, S. Zhu, and H. Al-aqrabi, "A Review of Routing Protocols for Mobile Ad-Hoc NETworks (MANET )," vol. 3, no. 1, pp. 1–5, 2013.
- [7] S. S. Ali and G. M. Someswar, "Vulnerability Discovery and Mitigation in OLSR of Mobile Ad-hoc networks," *Int. J. Sci. Eng. Res.*, vol. 4, no. 5, pp. 537–545, 2013.
- [8] A. Aneiba and M. Melad, "Performance Evaluation of AODV , DSR , OLSR , and GRP MANET Routing Protocols Using OPNET," *Int. J. Futur. Comput. Commun.*, vol. 5, no. 1, pp. 1–4, 2016.
- [9] A. Prof, M. K. Ibrahim, and A. M. Aboud, "A Secure Routing Protocol for MANET," *Int. J. Comput. Sci. Eng. Technol. IJCSET*, vol. 4, no. 7, pp. 223–230, 2014.
- [10] S. Dhawan, V. Saroha, and K. Kalan, "OPTIMIZE THE ROUTING PROTOCOL ( GRP , OLSR , DSR ) USING OPNET & ITS PERFORMANCE EVALUATION," *Int. J. Adv. Eng. Technol.*, vol. 6, no. 3, pp. 1399–1408, 2013.
- [11] M. Computing, "PERFORMANCE ANALYSIS OF AODV , OLSR , DSR AND GRP ROUTING PROTOCOL OF MOBILE ADHOC NETWORK – A REVIEW," *IOSR J. Comput. Eng.*, vol. 2, no. June, pp. 359–362, 2013.
- [12] M. A. Mehmood, A. M. Buttar, and M. Ashraf, "Experimental based Performance Analysis of Proactive OLSR , Reactive Tora and Hybrid GRP Routing Protocols in MANET," *Int. J. Comput. Appl.*, vol. 89, no. 15, pp. 23–30, 2014.
- [13] W. K. Lai and S. Hsiao, "Adaptive backup routing for ad- hoc networks," *SCI Comput. Commun.* 30, vol. 30, 2007.
- [14] D. Wadbude and V. Richariya, "An Efficient Secure AODV Routing Protocol in," vol. 1, no. 4, pp. 274–279, 2012.
- [15] D. Kaur and N. Kumar, "Comparative Analysis of AODV , OLSR , TORA , DSR and DSDV Routing Protocols in Mobile Ad-Hoc Networks," *I. J. Comput. Netw. Inf. Secur.*, no. March, pp. 39–46, 2013.
- [16] A. Munaretto, H. Badis, K. Al Agha, and G. Pujolle, "A Link-state QoS Routing Protocol for Ad Hoc Networks," *LIP6 Lab. Univ. Paris VI, 8, rue du Capit. Scott, 75015, Paris, Fr.*, pp. 23–26, 2002.
- [17] M. V. Khiavi, S. Jamali, and S. J. Gudakahriz, "Performance Comparison of AODV , DSDV , DSR and TORA Routing Protocols in MANETs," *Int. Res. J. Appl. Basic Sci.*, vol. 3, no. 7, pp. 1429–1436, 2AD.
- [18] M. V. H. B. Murthy and B. P. Rao, "Performance of Multimedia Traffic in OLSR Routing Protocol with Weighted Fair Queuing," *Int. J. Comput. Appl.*, vol. 44, no. April, pp. 6–10, 2012.

# Review of Various Mobility Models and its Applications for MANETS

Heena Rani<sup>[1]</sup>

Jasvir Singh<sup>[2]</sup>

1. Heena Rani, Department of Computer Engineering, Punjabi University, Patiala

Gaba.heena@gmail.com

2. Jasvir Singh, Assistant Professor, Department of Computer Engineering, Punjabi University, Patiala

**Abstract-** MANETs is used for mobile ad hoc networks. In this transmission is done by wireless medium. It have some special characteristics such as open network architecture shared wireless medium and highly dynamic topology which make them much prone to different attacks. If this movement is done between the wireless nodes. The communication between these mobile nodes is carried without any centralized control. In our paper we give information about various mobility models and there pros and cons. There are various types of mobility pattern. Each mobility pattern has their own effect on different networking applications. According to the nature of mobility pattern network performance will be affected. There are various routing protocols are available to check the movement of nodes. There are various types of mobility metric's are available to check the performance of network like how they work, output is according to input etc. In mobile opportunistic network mobility plays a important role to understand the nature with respect to humans, vehicles and wild animals. There are some latest mobility models are come which are on realistic based like vehicular mobility model, human based realistic model.

**Keywords – MANET, MOBILITY MODEL**

## I. INTRODUCTION

In mobility movement is done according to network perspective. If there is a no mobility then mobile user uses same access point. If there is a medium mobility then mobile user connecting or disconnection from network using dynamic host configuration protocol. If there is a high mobility then mobile user passing through different access points while maintaining ongoing connections.

MANET is basically used for mobile adhoc networks. In MANET, combination of wireless nodes are communicating without the use of existing data or any infrastructure. It is not very expensive .It is best suited for wireless connections. Today it is used in may research fields because this is very interesting. Mobility models plays a important role in the area of mobile adhoc network. Models help to determine how the protocols are performed in MANET. In Mobile Adhoc Network ,firstly we have to establish the protocol for deliver a packets. Routing efficiency relies on the nature of node movement and same protocol can perform differently for different type of node movement .

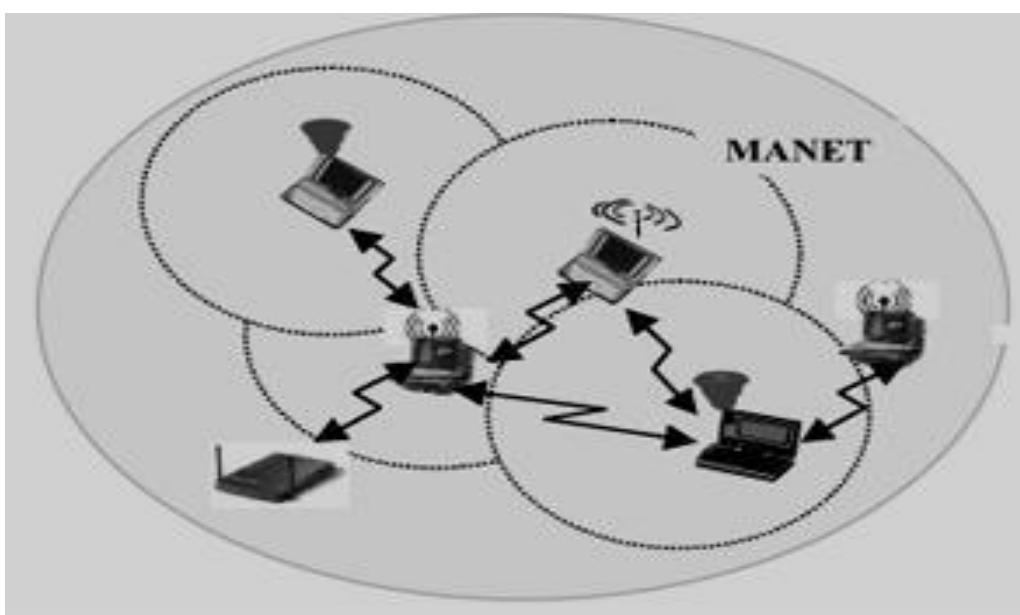


Figure1. Mobile AD HOC NETWORK [1]

Mobility means movement. Mobility Model tells how movement is carried out between the nodes. In this movement of mobile users is done and it also describes how location , velocity and acceleration change with the time. There are various mobility models which are used for the simulation of MANET. Each and every model has its own advantage. Model are used on the basis of requirements of node and network and topology.

There are various type of mobility models used for MANET. Like Random Based Mobility Models in which nodes are move randomly from source to destination, In Models with Temporal Dependency , Limit the movement of nodes, In Models with Spatial Dependency, in this Velocity of different nodes are correlated , Some Latest Mobility Models in which Human based, Vehicular based Mobility models are considered. New and latest mobility models are based on real life movement of humans and vehicles. In mobility for transferring the packet from source to destination we have to establish the path. There are various types of algorithms are available for finding the optimal path[2].

*Advantages:*

- Mobility improves the area of sensor network and help in various security issues.
- It helps in studying the network performance in any type of network.
- It is used in disaster management

## II TYPES OF MOBILITY MODEL

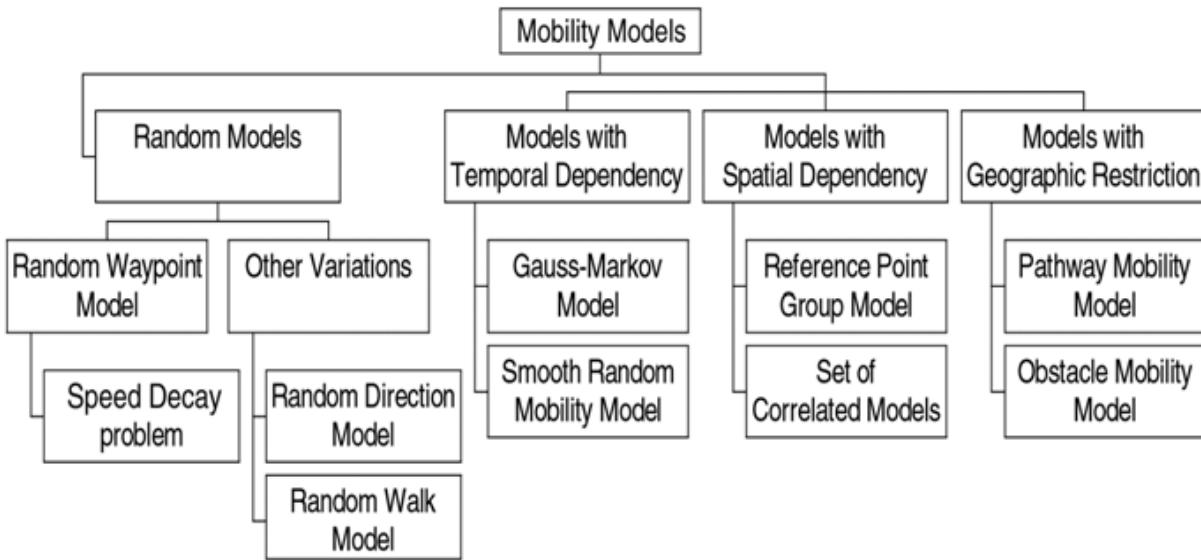


Figure2, Types of Mobility Model

### A. Random-Based Mobility Models :

In random-based mobility models, the mobile nodes move randomly and freely without any restrictions. To be more specific, the destination, speed and direction are all chosen randomly and independently of other nodes. This kind of model has been used in many simulation studies. One frequently used mobility model, the Random Waypoint model. Then, two variants of the Random Waypoint model, namely the Random Walk model and the Random Direction model.

Types are Random Way Point Model, Radom Direction Model, Random Walk Model.

1). *Random Walk Mobility Model*[3]: This model is given by Einstein in 1926[4].It is also called Brownian motion. In this model nodes move freely from its current location to new location by randomly choosing a direction and speed.

*Pros and Cons are:* Generate random samples from large set. Used to balance the load. It is used in typical networks etc. Sometimes it is not suitable for wireless networks. It is a memory less mobility process.

2). *Random Way Point Mobility Model*: It was first proposed by Johnson and Maltz[5]. In this model node choose their own path anywhere in the network area. In this model we use pause time parameter between change in

direction and speed. In this node move towards target after reaching the target node stop for certain amount of time and then again move. This process will be continuing until the simulation task is not complete.[6]

*Pros and Cons are:* It is used in ad hoc networks. It is easy to implement. In this we see fluctuations in density waves. It has high variability that effects result of simulation. When simulation time elapse , then unbalanced spatial distribution becomes worse. It suffers from speed decay problem. It has unrealistic nature of node movement.[7]

3). *Random Direction Mobility Model[8]:* This model was given by Bettstetter. It is designed to minimize the effect of density waves. In this also node reached the target and the stop for a little time and then choose a new target to move. In random direction mobility model slight modifications are done so its also called Modified Random direction mobility model which if given by Royer et al[9]. Simplified version of this model is given in literature[10].

*Pros and Cons are:* less fluctuation in density waves. It overcomes the non uniform distribution.

#### B. Mobility Models with Temporal Dependency

Mobility of a node may be constrained and limited by the physical laws of acceleration, velocity and rate of change of direction. Hence, the current velocity of a mobile node may depend on its previous velocity. Thus the velocities of single node at different time slots are ‘correlated’. We call this mobility characteristic the Temporal Dependency of velocity. However, the memory less nature of Random Walk model, Random Waypoint model and other variants render them inadequate to capture this temporal dependency behavior. As a result, various mobility models considering temporal dependency are proposed.

Types are Gauss Markov Model, Smooth Random Model.

1). *Gauss Markov Mobility Model[11]:* This was introduced by Liang and Haas[11]. It is developed for Personal Communications Services. In this firstly given the current location and speed to the node and then after a interval of time movement is done and speed and direction will be changed for each node. In this new location is computed by current speed, direction and location.

*Pros and Cons:* It removes sudden start and stop turns. In this memory plays a important role to find out the temporal dependency.

2). *Smooth Random Mobility Model[12]:* It is proposed by Bettstetter. Its name describes nodes moves smoothly. In this new value of speed and direction corelate with the values of previous process. It helps to remove sudden start and stop turns. Correlation between speed change and direction has two concepts. Stop turn and go behavior, and second is Slowdown of turning nodes. In this model nodes movement is smooth and realistic then other models.

*Pros and Cons are:* In this nodes movement are behave like a realistic manner. It helps to evaluate the routing performance. In this nodes are move incrementally and smoothly. In this frequency of speed change is assumed to be a poisson process. In this model speed and movement is based on the previous values.

### C. Models With Spatial Dependency

In the Random Waypoint model and other random models, a mobile node moves independently of other nodes, i.e., the location, speed and movement direction of mobile node are not affected by other nodes in the neighborhood. As previously mentioned, these models do not capture many realistic scenarios of mobility. For example, on a freeway to avoid collision, the speed of a vehicle cannot exceed the speed of the vehicle ahead of it. Moreover, in some targeted MANET applications including disaster relief and battlefield, team collaboration among users exists and the users are likely to follow the team leader. Therefore, the mobility of mobile node could be influenced by other neighboring nodes. Since the velocities of different nodes are 'correlated' in space, thus we call this characteristic as the Spatial Dependency of velocity.

1). *Reference Point Group Mobility Model[13]:* In this model nodes are move within a group. Each group having a leader. Group movement is based on the path travelled by a logical centre of the group. If there is a individual node then there is no need to use pause time parameter. Pause time parameter is used when group reference point reaches the target and the all group nodes stop for some duration of time. There are various applications of this model : In place mobility mode, Overlap mobility model, Convention mobility model.

*Pros and Cons are:* This model is used for conferences, for rescue teams and museum visits. In this model distortion of link is less, it has high performance for various routing protocols.

2).*Set of Correlated Mobility Model:* Sanchez and Manzoni[14] propose a set of mobility models in which the mobile nodes travel in a cooperative manner. This set of mobility models, including Column Mobility Model[15], Pursue Mobility Model[15] and Nomadic Mobility Model[15], are expected to exhibit strong spatial dependency between nearby nodes.

Following we describe these mobility models and their applications.

It has further 3 types

- 1) Column Mobility Model
- 2) Pursue mobility Model
- 3) Nomadic Community Mobility Model

### D. Mobility Models with Geographic Restriction

This mobility model is come to improve the drawbacks of random mobility model where nodes are move freely without any restriction, mobility model with temporal dependency and mobility model with spatial dependency. In geographic restriction nodes have given their own predefined environment to move. Each node move in a predefined area by this way chance of overlapping, traffic are less.

Types are pathway mobility model, obstacle mobility model.

*1). Pathway Mobility Model[16]:* In this map is used which is predefined in the simulation area. To create a map of city Tian et al[17], utilize a random graph. In this two models are used to design a real city map and to represent a section of a city in a street network. These models are city section model and city map model.

*Pros and Cons:* In this model , nodes are travelling in a pseudo random way on the pathways.

*2). Obstacle Mobility Model[18]:* Johansson, Larsson and Hedman et al.[19] develop three 'realistic' mobility Scenarios to represent the movement of mobile users in real life, like Conference Scenario, Event Coverage scenario, Disaster Relief Scenario. In these scenarios obstacles are in the form of rectangular boxes. Jardosh et[18] al gives detail about the effect of obstacles on mobility modeling This model tells the positions of problems in the cities or any buildings. Problems effect the movement of nodes. To avoid the problems of area , mobile node is required to change its trajectory.

*Pros and Cons:* It is used in mobility modeling. When obstacles are combined with mobility model , then they both effect the node mobility and an radio propagation. When node transmit, the obstacles obstruct the propagation of the transmission in an area and defined as the obstruction cone of the node.

#### *E. Latest Mobility Models*

There are various new mobility models are available some models are based on real life scenario. So these models are Human or Sociality based mobility model, Vehicular mobility model.

*1) Human Based Mobility Model[20]:* To improve the Manet accuracy and mobility model. In this model movement is not in the random manner it is in regular manner. These parameters are needed for efficient working of human mobility model: Pause time; Return time; Velocity & Acceleration; Direction angle change; Displacement (Flight length); Preferred Area & Boundary.

*Pros and Cons:* It is used in Pocket Switched Networks[21]. This will improve the performance of manet. Disadvantage is that lack of real data of human movement records.

2) *Vehicular Based Mobility Model*[22]: In this model nodes are not move in a random direction. In this Intelligent transport system is a integral part of vehicular networks , which are used for traffic safety . It has two components : vehicle to vehicle and vehicle(V2V) to infrastructure(V2I). These components are used for traffic management and congestion management. V2I is used for internet access. V2V is used for direct communication.

*Pros and Cons:* It is used for both wireless networks and transportation research. It has high speed and degree of freedom of node is limited. But its high mobility affect the network performance and length of Streets and density of building, affects the network performance.

### Reference

- [1] Dr. Ravi Sindal and Nidhi Jaiswal, Performance Analysis of Various applications Protocols in MANET, International Journal of Advance Research in Computer Engineering and Technology, Vol 2, Issue & 7 July,2013.
- [2] Suvadip Batabyal and Parama Bhaumik : Mobility Models, Traces and Impact of Mobility on Opportunistic Routing Algorithms: A Survey, in IEEE Communication Surveys and Tutorials, Vol..17, No. 3, Third Quarter 2015
- [3] \* F. Bai, A. Helmy, "A Survey of Mobility Modeling and Analysis in Wireles Adhoc Networks", Book Chapter in the book "Wireless Ad Hoc and Sensor Networks", Kluwer Academic Publishers, June 2004.
- [4] A. Einstein, Investigations on the Theory of the Brownian Movement. New York, NY, USA: Dover, 1956
- [5] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network routing protocols," in Proc. 4th Annu. ACM/IEEE Int. Conf. MobiCom Netw., Oct. 1998, pp. 85–97.
- [6] Ahmed Helmy, Tutorial Mobility Modeling for Future Mobile Network Design Simulation, Computer and Information Science and Engineering (CISE) College of Engineering University of Florida.
- [7] Suvadip Batabyal and Parama Bhaumik : Mobility Models, Traces and Impact of Mobility on Opportunistic Routing Algorithms: A Survey, in IEEE Communication Surveys and Tutorials, Vol..17, No. 3, Third Quarter 2015
- [8] A Survey of Mobility Models for Ad Hoc Network Research Ha Yoon Song Guest Professor at ICT, TUWien [song@ict.tuwien.ac.at](mailto:song@ict.tuwien.ac.at).
- [9] E. M. Royer, P. M. Melliar-Smith, and L. E. Moser, "An analysis of the optimum node density for ad hoc mobile networks," in Proc. IEEE ICC, 2001, pp. 857
- [10] Z. J. Haas and M. R. Pearlman, "The performance of query control schemes for the zone routing protocol," in Proc. ACM SIGCOMM, Sep. 1998, pp. 167–177
- [11] B. Liang and Z. J. Haas, "Predictive distance-based mobility management for PCS networks," in Proc. IEEE INFOCOM, Apr. 1999, pp. 1377–1384
- [12] C. Bettstetter. Smooth is better than sharp: A random mobility model for simulation of wireless networks. In Proceedings of MSWiM'01. ACM, July 2001.
- [13] X. Hong, M. Gerla, G. Pei, and C. Chiang, "A group mobility model for ad hoc wireless networks," in Proc. ACM Int. Workshop MSWiM, Aug. 1999, pp. 53–60.
- [14] M. Sanchez and P. Manzoni, A Java Based Ad Hoc Network Simulation in Proceedings of the SCS Western Multiconference Web basedSimulation Track, Jan.1999.
- [15] M. Sanchez and P. Manzoni, "ANEJOS: A java based simulator for ad hoc networks," Future Gen. Comput. Syst., vol. 17, no. 5, pp. 573–583, Mar. 2001
- [16] V. Davies, "Evaluating mobility models within an ad hoc network,"M.S. thesis, Dept. Math., Colorado School of Mines, Golden, CO, USA, 2000.
- [17] J. Tian, J. Hahner, C. Becker, I. Stepanov, and K. Rothermel, "Graphbased mobility model for mobile ad hoc network simulation," Proc. 35th Annu. Simul. Symp., San Diego, CA, USA Apr. 2002, pp. 337–344.

- [18] A. Jardosh, E. M. Belding-Royer, K. C. Almeroth, and S. Suri, "Towards realistic mobility models for mobile ad hoc networks," in Proc. 9th Annu. Int. Conf. MobiCom Netw., Sep. 2003, pp. 217–229.
- [19] P. Johansson, T. Larsson, N. Hedman, B. Mielczarek, and M. Degermark, Scenario-based performance analysis of routing protocols for mobile ad-hoc networks, in International Conference on Mobile Computing and Networking (MobiCom'99), 1999, pp. 195—206
- [20] Yukun Zhao, Analyzing the mobility model, Wireless and mobile communication group , Delht university and Technology
- [21] P. Hui *et al.*, "Pocket switched networks and the consequences of human mobility in conference environments," in *Proc. ACM SIGCOMMWDTN*, 2005, pp. 244–251.
- [22] Sanjib Debnath and Abhishek Majumder, A SURVEY ON MOBILITY MODELS FOR VEHICULAR AD HOC NETWORKS, IJARET: Volume: 02 Special Issue: 02 | Dec-2013, Available @ <http://www.ijret.org>

# Compression Using Deflater Algorithm with Encryption using AES-256 of HEVC Videos for More Effective Transmission

Deepak Sharma  
Department of CSE  
SBSSTC Ferozepur, Punjab (India)  
deepdeep2003@gmail.com

Japinder Singh  
Department of CSE  
SBSSTC Ferozepur, Punjab (India)  
japitaneja@gmail.com

**Abstract**— A very high demand of the computer technology and internet technology, multimedia service has become a new interested area. But there are lots of problem like a huge amount of data, high speed play, bandwidth limitation. For enhancing the security and speed, HEVC-high efficiency video coding provides better compression with greater quality and low bandwidth usage. Several algorithms have been proposed for efficient and secure streaming of HEVC/HD video files. These algorithms are only useful for encrypting high resolution videos. In the proposed algorithm, a combination of AES-256 encryption algorithm and Deflator compression algorithm is used for efficient encrypted data delivery with compressed size. Experimental results proves that proposed algorithms gives better encryption and smarter compression of all type of multimedia files as compared to previous algorithms with minimum processing time.

**Keywords-** *transmission; deflater; compression; base 64 encoder; encryption; AES- 256*

## I. INTRODUCTION

In recent years, Internet becomes a backbone of every sector, which provides knowledge of each domain like educational concerns, entertainment etc. Data theft and data stealing is the well-known thing in networks. Cryptography is a core technology of cyberspace.” Cryptography is a technology that allows us to bring security in systems. Therefore, cyberspace without it, there would be no privacy, no e-commerce, and no security of any information [3].

The major goal of cryptography is to enhance communication security by encrypting the original text in the encoded form. The primary goal of cryptography is to provide the confidentiality, data integrity, authentication and non-repudiation. As good security is becoming a necessity for non-governmental, military applications and in other medical, financial fields. Many algorithms have been proposed and developed for efficient and secure streaming of multimedia files. But AES is the strongest algorithm for providing security. The AES encryption algorithm generates cipher text in blocks and algorithm works on individual blocks at a time by using an encryption key and several rounds of encryption. In the case of standard AES encryption the block length is 128 bits. The label ‘Round’ indicates the way in which the encryption algorithm intermix the data encrypting again and again ten to fourteen times based on the key length. It is a mathematical description of a process of obscuring data [2].

AES encryption uses a single key as a part of the encryption process. The key can be 128 bits (16 bytes), 192 bits (24 bytes), or 256 bits (32 bytes) in length. The term 128-bit encryption refers to the encryption key of a 128-bit. In AES the encryption and the decryption are performed using the same type of key. This is called a symmetric encryption algorithm. Encryption algorithms that use two different types of keys, one is a public and second is a private key, is called asymmetric encryption algorithms [2].

The block cipher Rijndael was developed by Joan Daemen and Vincent Rijmen and was based on their previously developed block cipher, Square. The algorithm can be efficiently implemented on a wide range of computer system processors and hardware. The AES development process has determined that the Rijndael algorithm is very secure and has no known weaknesses. In accord with AES requirements, Rijndael's key length can be defined at 128-, 192- or 256-bits. Rijndael has a variable block length that can define as 128-, 192-, or 256- bits. What does this mean? Basically, Rijndael, which will use the AES specified key sizes of 128-, 192- and 256-bits will provide approximately:

- $3.4 \times 10^{38}$  possible 128-bit keys;
- $6.2 \times 10^{57}$  possible 192-bit keys; and
- $1.1 \times 10^{77}$  possible 256-bit keys.

If we compare these key possibilities with that of DES, which has a 56-bit key size and approximately  $7.2 \times 10^{16}$  possible keys, it is evident that it would require much more computing power to crack the key. In their AES Fact Sheet, NIST uses the following hypothetical example to illustrate the AES security in theoretically. If one were to assume that a computing device existed that could retrieve a DES key in some seconds, it would take that same system almost 149 trillion years to crack a 128-bit AES key. They further illustrate the point by reminding us that the universe is believed to be less than 20 billion years old [3] [4]. This algorithm is applied on different-different techniques like as compression (HEVC) for providing better security.

## II. HIGH EFFICIENCY VIDEO CODING

HEVC is the modern video coding standard of the ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group. The main purpose of the HEVC standardization effort is to enable improved compression performance relative to previous standards — in the range of 50% bit-rate reduced for equal perceptual video quality. The standard of HEVC is designed to attain different goals, including ease of transport system integration, coding efficiency and data loss resilience, as well as ease of implementation with the help of parallel processing architectures. HEVC is also known as H.265.[1][5][7]

The major video coding standard directly lead to the HEVC project was H.264/MPEG-4 AVC, which was initially developed in the period between 1999 and 2003, and then was extended in different methods from 2003–2009. [9] [10] H.264/MPEG-4 AVC has been an enabling technology for digital video that was not previously covered by H.262/MPEG-2 Video and has substantially displaced the previous standard within its existing application domains. It is widely used for many applications, including broadcast of high definition (HD) TV signals over cables, satellite and terrestrial transmission systems, video content acquisition and security applications, Internet and mobile

network video, Blu-ray Discs, and real-time conversational applications such as video chat, video conferencing, and telepresence videoconferencing[1][6].

To increase the diversity of services, the increasing popularity of HD video, and the emergence of beyond HD formats are creating stronger needs for efficiency of coding superior to H.264/MPEG-4 AVC's capabilities. Moreover, the traffic caused by video applications targeting mobile devices and the transmission needs for video-on-demand services, are imposing severe challenges on today's networks.

In HEVC only the bit stream structure and syntax is standardized, as well as constraints on the bit stream and its mapping for the origination of decoded images. The leveling is given by defining the semantic meaning of syntax elements and a decoding process such that every decoder belong to a particular standard will generate the same output when given a bit stream that conforms to the constraints of the standard. This limitation of the scope of the standard permits maximal freedom to optimize implementations in a manner appropriate to specific applications (quality and balancing of compression, cost of implementation, time for commercialism, and other applications). However, it provides no guarantees of end-to-end reproduction quality, as it permits even basic encoding techniques to be considered conforming [1].

### III. DEFLATE COMPRESSION

In information technology data compression, source coding, or bit-rate reduction deals with encoding of information using fewer bits than the original representation. In other words the mechanism of reducing the size of a data file is commonly known as data compression or source coding, as coding is done at the source of the data before it is stored or transmitted. Compression can be identified as lossy or lossless. Lossless compression reduces bits by identifying and eliminating redundancy and no information is lost in this compression. The art of reducing the bits by recognizing minimal principal information and removing it refers to lossy compression [8][11].

There are numerous compression algorithms to compress the size of data. In this paper Deflate Compression Algorithm is presented to reduce the text size. In practice, deflate is a data compression algorithm and associated file format that uses a combination of the LZ77 algorithm and Huffman coding. This algorithm was originally proposed by Phil Katz for version 2 of his PKZIP archiving tool.

During the compression stage, it is the encoder that chooses the amount of time spent looking for similar strings. The zlib/gzip reference implementation helps the user to select from a sliding scale of likely resulting compression-level vs. encoding speed. Options range from -0 (do not attempt compression, just store uncompressed) to -9 representing reference implementation maximum capability in zlib/gzip. Other Deflate encoders have been produced, all of which will also produce a compatible bit stream capable of being decompressed by any existing Deflate decoder. Differing implementations will likely give variations on the final encoded bit-stream produced. The focus with non-zlib versions of an encoder is generally to generate a better efficiently compressed and smaller encoded stream.

### IV. PROPOSED METHOD

Proposed work presents 256-bits AES Algorithm for encryption process and Deflater compression algorithm with base 64 bit Encoder for decryption process. Initially 16 bit or 32 bit secret key is generated for selected multimedia

file. Afterword's, the cipher mode in the coding part is activated to read the contents of the file from disk location and these contents are encoded with base 64 Encoder. In the next step the contents size is reduced up to some extent using Deflater compression algorithm. Furthermore, implementation of AES 256 algorithm is done for encrypting the compressed file. As a result a final encrypted file is obtained at one end. Similarly, in decryption process same procedure is followed in reverse direction with bottom up approach.

Figure1 shows the working of proposed algorithm. AES-256 algorithm has generated encrypted files with greater size as compare to our implemented algorithms. The use of AES-256 encryption algorithm and Deflate compression algorithm has created small size encrypted files. Thus, it can be considered that implemented algorithm is of better quality as compare to other algorithms, as it produces better results with less time consumption.

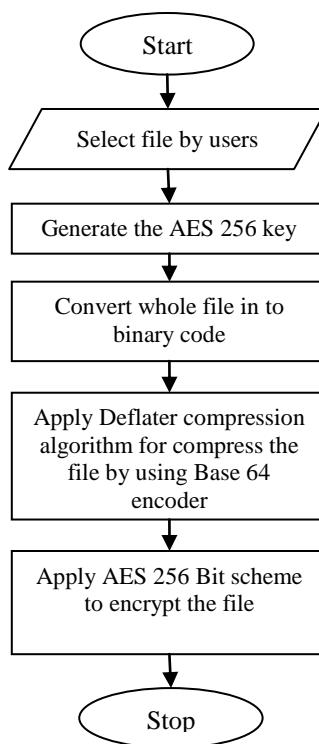


Figure1. Working of proposed Algorithm

The steps of proposed algorithm are given below:

#### *The encrypted process*

*Step 1: First select the file to be encrypted.*

*Step 2: Generate Key For Encrypt the file*

```
Key secretKey = new SecretKeySpec(key.getBytes(), ALGORITHM);
```

*Step 3: Select the cipher mode if Encryption: select*

```
Cipher.ENCRYPT_MODE.
```

```
Cipher cipher = Cipher.getInstance (TRANSFORMATION);
cipher.init(cipherMode, secretKey);
```

*Step 4: Get the file content into Bytes*

```
FileInputStream inputStream = new FileInputStream(inputFile);
byte[] inputBytes = new byte[(int) inputFile.length()];
```

*Step 5: Encode this bytes Using Base 64 Encoder.*

```
String jh=enc.encode(inputBytes);
```

*Step 6: Compress the bytes using Deflater Compression Algorithm.*

```
Deflater compressor = new Deflater();
compressor.setInput(bn);
compressor.finish();
```

*Step 7: Now encrypt the final result with the help of AES algorithm.*

```
byte data[]=cipher.doFinal(output);
```

## V. RESULTS AND ANALYSIS

The Research is carried on different size of multimedia files with proposed algorithm which include compression and cryptography with AES-256 Bit. As shown in table1. Significant results are obtained as compared to Encryption algorithm for efficient transmission of HEVC media.

### A. Performance Parameters

The proposed algorithm is compared with AES-256 bit Encryption algorithm for efficient transmission of HEVC media on the basis of two performance parameters:

*1) Encryption Time:* Encryption time deals with time required for encryption process. AES-256 Bit mechanism is used for encoding and encrypting the files. Table1 shows the encryption time for different set of multimedia files belongs to classes A, B, C and D. Over here, encryption time taken by proposed algorithm for encrypting different set of multimedia files is compared with encryption time of previously implemented AES HEVC algorithm.

TABLE I  
 RESULTS OBTAINED BY PROPOSED ALGORITHM

Class	Name of Multimedia File	Actual size	Resolution of file	Old encryption time duration	New Encryption time duration with proposed Algorithm	Size after encryption with AES-256 HEVC—(H.264) (MB)	Size after Purposed encryption with AES-256 HEVC—(H.264) (MB)	Compression Ratio %
A	A1.MP4	8.75	2560 x1600	5	0.1406	9.60	9.09	94.69
A	A2.MP4	8.75	2560x 1600	5	0.1404	9.60	9.09	94.69

B	B1.MP4	7.49	1920 x1080	5	0.1162	7.68	7.53	98.04
B	B2.MP4	7.49	1920 x1080	5	0.1190	19.20	8.06	44.79
C	C1.MP4	2.50	800 x480	10	0.0501	16.00	2.53	15.81
C	C2.MP4	2.50	800 x 480	10	0.0412	16.00	2.68	16.75
D	D1.MP4	1.88	400 x240	10	0.0352	9.60	1.96	20.42
D	D2.MP4	1.88	400 x 240	10	0.0350	19.20	1.95	10.16

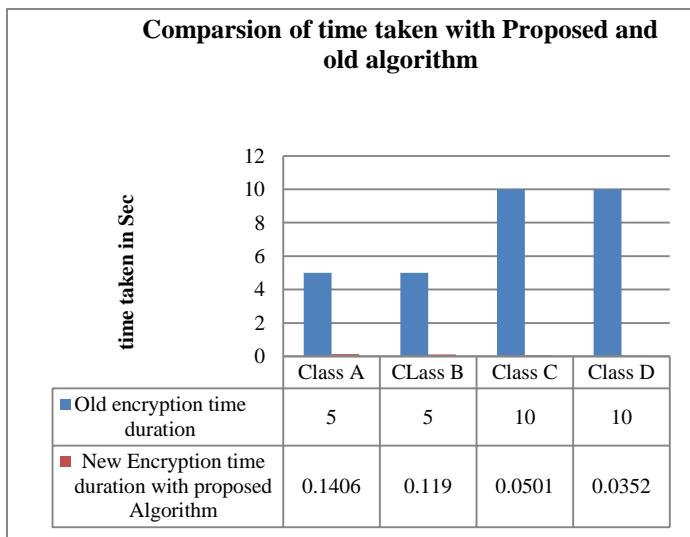


Figure 2. Comparison Encryption Time

Figure 2 depicts the graphical representation of comparison of encryption time of four different classes of multimedia files for both the algorithms. Encryption time duration of proposed algorithm is less as compare to old algorithm.

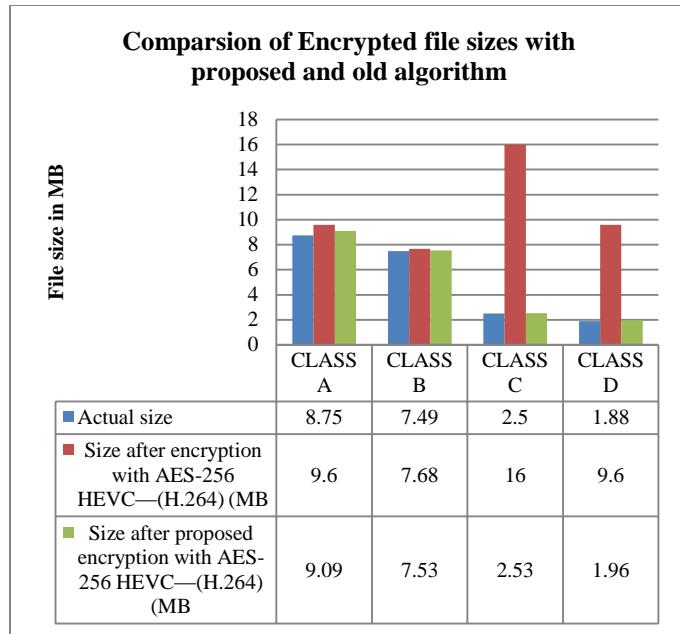


Figure 3. Comparison Encryption Time

Similarly, figure 3 shows that required encryption size of files by proposed algorithm is smaller than encrypted file size of AES-256 HEVC.

## VI. CONCLUSION

Proposed work, an association of compression and encryption algorithm is presented. This implementation of AES-256 in union with Deflater compression algorithm increases the performance of data transformation. The proposed algorithm provides better encryption of all type of multimedia files with minimum encryption time. As a result, it effectively improves the bandwidth utilizations and also decreases the traffic of entire network due to minor size of packet. The entire performance of the network will increase while according to the comparative analysis of the recent encryption standards, HEVC compression standard is shown to be good than H.264 using the same algorithms—DES, AES-128, and AES-96—for intra-frames encryption, because of the fact that it shows bandwidth utilization and need less time to be encrypted with suitable algorithms compared to H.264.

## VII. FUTURE WORK

We have implemented Deflater compression algorithm with encryption which reduces the size of encrypted compressed files. So, better compression algorithm can be defined which efficiently work with cryptography key length algorithm with fast access and secure delivery of packets.

## REFERENCES

- [1] Vasileios A. Memos, Kostas E. Psannis, "Encryption algorithm for efficient transmission of HEVC media," Springer journal of real time Image processing, springer, May 2015.
- [2] Townsend Security, "Introduction to AES Encryption".

- [3] William M. Tatum, "The Advance Encryption System (AES) Development Effort: Overview and Update," SANS Security Essential GSEC Practical Assignment version 1.2e, 2001.
- [4] United States Department of Commerce/National Institute of Standards and Technology. "Advanced Encryption Standard (AES) Fact Sheet."
- [5] F. Bossen, B. Bross, K. Suhring and D. Flynn, "HEVC Complexity and Implementation Analysis," IEEE Trans. Circuits Syst. Video Technol, pp. 1685–1696, vol. 22, no. 12, 2012.
- [6] Frank Bossen, Benjamin Bross, Karsten Suhring, and David Flynn, "HEVC Complexity and Implementation Analysis," IEEE Transactions on Ransactions on Circuits and Systems for Video Technology, vol. 22, no. 12, December 2012.
- [7] Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han and Thomas Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," IEEE Transactions on Ransactions on Circuits and Systems for Video Technology, vol. 22, no. 12, December 2012.
- [8] Jens-Rainer Ohm, Gary J. Sullivan, Heiko Schwarz, Thiw Keng Tan, and Thomas Wiegand, "Comparison of the Coding Efficiency of Video Coding Standards—Including High Efficiency Video Coding (HEVC)," IEEE Transactions on Ransactions on Circuits and Systems for Video Technology, vol. 22, no. 12, December 2012.
- [9] Guo Jie, Qiu Weidong, Du Chao and Chen Kefei, "A Scalable Video Encryption Algorithm for H.264/SVC," Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 513).
- [10] Dan Grois, Detlev Marpe, Amit Mulayoff, Benaya Itzhaky, and Ofer Hadar, "Performance Comparison of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC Encoders," Picture Coding Symposium 513 (PCS 513), San José, CA, USA, Dec 2011.
- [11] P. Hanhart, M. Rerabek, F. De Simone and Ebrahimi T, "Subjective quality evaluation of the upcoming HEVC video compression standard, applications of digital image processing XXXV,"In Proceedings of SPIE, vol. 512.
- [12] Niansheng Mao, Li Zhuo, Xiaoguang Li and Jing Zhang, "A Video Encryption Algorithm in H.264 Compressed Domain for Resource-Limited Systems, " in Proceedings of International Journal of Intelligent Information Processing, vol. 2, no. 1, March 2011.
- [13] Zafar Shahid and Williard Puech, "Visual Protection of HEVC Video by Selective Encryption of CABAC Binstrings," in IEEE Transactions on Multimedia, vol. 16, no. 1, JANUARY 2014.
- [14] C Saranya. P and Varalakshmi. L.M, "H.264 based Selective Video Encryption for Mobile Applications," International Journal of Computer Applications , vol. 17, no.4, pp. 0975 – 8887, March 2011.
- [15] Thomas Schierl, Miska M. Hannuksela, Ye-Kui Wang and Stephan Wenger, "System Layer Integration of High Efficiency Video Coding," IEEE Transactions on Ransactions on Circuits and Systems for Video Technology, vol. 22, no. 12, December 2012.

# Cross layer integrated MAC and Routing Protocol based on ring based structure with adaptive sleeping for Wireless Sensor Networks

<sup>1</sup>Bindiya Jain, <sup>2</sup>Gursewak Brar and <sup>3</sup>Jyoteesh Malhotra

<sup>1</sup>DAV Institute of Engg. And Technology,

<sup>2</sup>BBSBEC.Fatehgarh Sahib

<sup>3</sup>Guru Nanak Dev University, Jalandhar

<sup>1</sup>E-mail:bindiyajain29@gmail.com

<sup>2</sup>E-mail: brargs77@rediffmail.com

<sup>3</sup>E-mail: jyoteesh@gmail.com

## **Abstract:**

Cross-Layer Design have now developing hottest research area in wireless sensor systems as it is pursuing to improve the ability of wireless networks remarkably through the joint optimization of several layers in the network. The design of our protocol is to deal with the overhearing problem and low duty cycle based on adaptive sleeping. The Mac layer is using the information of routing layer to decide the node duty cycle in order to extend the node's sleep time. Thus cross layer protocol design objective is to reduce protocol overhead significantly for route discovery as ring based structure find shortest path for the node in the outer ring to the sink in the network. Extensive simulation results shows that ring based protocol EEIMRP as compared to IMRP can significantly reduce the protocol overhead thus reduces delay in the network, improves energy efficiency ,throughput and prolonged network lifetime.

**Keywords:** Wireless Sensor Networks, Cross Layer protocol, Energy efficiency, Ring based sensor

Network, Adaptive sleeping.

## **I. INTRODUCTION**

Today, in wireless sensor Network bearing in mind energy constraints, different layers of protocol stack have given role to save energy, but designing such cross layer algorithms which conserve energy and maximize network lifetime has been widely used to improve performance of WSNs. One of the major active research areas in the field of wireless sensor networks is the design of cross layer integrated routing and MAC protocols. In cross layer design at routing layer new cross layer routing algorithms are proposed to balance energy and prolong the network lifetime that are different from conventional routing protocols. At Mac layer, many protocols are designed to address the energy issue in wireless sensor networks. In this paper, MAC protocol is proposed in which the idle listening of node is avoided which are in the range of the actual node which are used for routing, thus MAC layer is sharing the routing information from network layer. The authors of [1, 2] indicated that the major energy wastage occurs when the nodes remains

listening during an idle period i.e. idle listening. To deal with this matter, some research works [3, 4] have proposed different Medium Access Control (MAC) protocols to reduce the time of idle listening so as to prolong the lifetime of WSNs. Cross-Layer MAC protocol [5] is advised to handle the latency of the packet because this scheme proposed multi flow data by taking routing layer also in consideration that buffer packets from neighbors which is different from regular MAC layer protocol. Cross layer protocols allow communication between adjacent layers routing and MAC layer to have an integrated protocol for routing of packets from outside the ring system to the sink. The ring system has been developed to ensure energy saving as it ensures that there are no too long hops and no too short hops. Numerous cross-layer protocols have been suggested to address precise optimization variables in wireless sensor networks, such as at Mac layer link scheduling and routing flow at routing layer [13], [14], [15],[16]. Today, most of studies identified that major source of energy consumption in sensor networks is idle listening such as listening for start of neighbor transmission and header processing to determine if it is addressed to node and it is observed that 50% of energy is consumed when no actual transmission is taking place in sensor network. Cross layer integration is used to solve energy wastage due to idle listening and overhearing because cross layer integration protocol effectively reduces the protocol overhead and makes the protocol stack lightweight.

## II. RELATED WORK

In this section very brief review on the some of the existing cross layer MAC and routing layer integrated protocols in wireless sensor networks is presented. Complex network functions are broken down in to independent layers and the function of Mac layer is different from the function of Network layer. In the field of wireless sensors the need of Mac and routing cross layer integrated application specific protocols has been identified [6].The authors in [7] proposed a routing protocol CL-RS describing a joint cost function that balances the energy in the network and that enables routing decision for different network sizes, different connections and different rate of transmission. By using cross layer design in wireless sensor networks we can make these networks more application specific and more energy efficient by reducing protocol overhead as MAC layer is using the information of routing layer. ALPL [8] considers both MAC and routing layer and adaptively selects routing path .In BMAC the node interval is fixed whereas in ALPL the interval of node is based upon the node tree-level and thus avoiding the waste of energy. Hybrid MAC protocol [9] proposed cross layer approach by using the strengths of both MAC protocols which utilizes the routing information of AODV and results in increase in network lifetime and reduces packet latency. RAWMAC [10] an adaption layer which uses routing layer in to management layer for asynchronous duty cycled MAC protocols. For every sensor node, the received control messages of AODV are received to adjust the behavior of our MAC protocol between TDMA and CSMA. Area Cast [11] uses cross layer approach which uses routing layer information to select the explicit relay node and k limited self elected implicit nodes and secondly this protocol is used to apply sleep/listen period which improves energy consumption, end to end delay and relay efficiency. In this protocol communication by area dynamically avoids faulty nodes and unstable links. A typical cross layer approach to

solve such a problem is to use a shortest path algorithm in which the edging cost is the energy consumed to send a packet between the nodes on that boundary. Cross layer energy aware routing scheme for multichannel access WSNs [12] guarantee to meet data rate requirements of end to end flows while maximizing the network lifetime.

#### A. Proposed EEIMRP MAC protocol Earlier

CI –MAC cross layer integrated MAC and Routing layer protocol was proposed .In this case whole sensor field network is divided in to rings with sink at the center as shown in Figure 1.

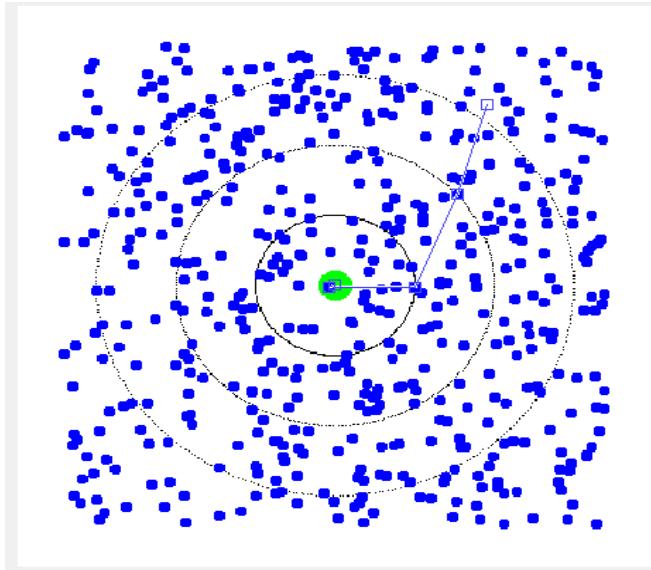
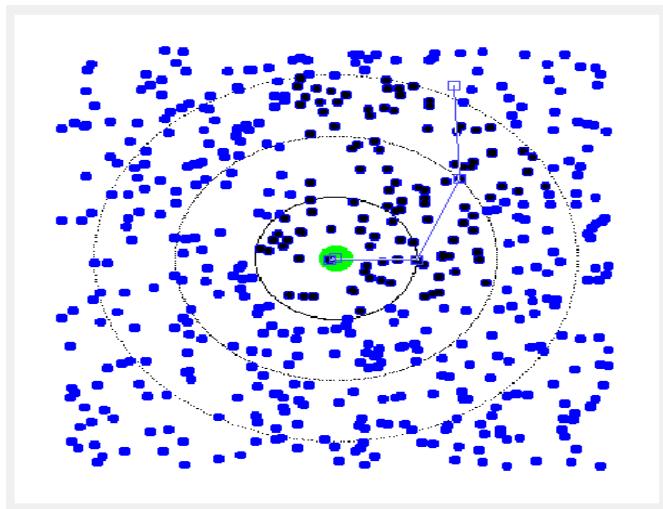


Figure 1: Network Model

In CI-MAC both MAC and routing layer are used for routing of packets from outside the ring system to the sink which is at the center of network. The ring system has been developed to ensure energy saving as it ensures that there are no too long hops and no too short hops. The rings are numbered 1, 2 and 3 so on starting from the innermost ring. It is assumed that each node in our network is GPS enabled and aware of its geographic location. While selecting the next hop candidate the node will always select the node from next ring which is closest to sink. For example, if the current node is in the third ring then the next node will be from next ring that means from second ring. So the node which is in the next ring and closest to sink is selected in its own range. The role of MAC layer is that it always assigns the ring no whenever a node is deployed and while accessing the nodes for consideration of next candidate this information from the MAC layer is used. The role of routing layer is that it uses the range, distance table and information from MAC layer to choose the next hop candidate. In CI-MAC, routing information plays a vital role to send data progressively to the sink, besides decrease in energy consumption.

In this paper, the cross-layer approach is used to design an energy-efficient communication protocol for the wireless sensor networks - EEIMRP (Energy Efficient Integrated MAC and routing layer protocol). EEIMRP uses the routing information from routing layer to find the duty-cycle of each node, and in the meantime this protocol also pays attention to collision and overhearing problem in the MAC layer.



**Figure 2: Illustration of routing in EEIMRP protocol**

### B. EEIMRP Description

In this section working of EEIMRP (Energy Efficient Integrated MAC and routing layer protocol) is described. In this proposed work a new rule is proposed in which idle listening of node is avoided which are in the range of the actual node which are used for routing. In Sensor networks main source of energy consumption is idle listening, overhead and overhearing i.e. listening to the packet not addressed to it. In order to reduce energy consumed due to idle listening, nodes periodically sleep. In the proposed protocol the ring system has been developed in which every node has a ring no. and while selecting the next hop candidate the node will always select the node from next ring. For example, if the current node is in the third ring then the next node will be from next ring that means from second ring as shown in the Figure 2. While selecting the next hop candidate it has been seen that the next candidate is closest to sink. The nodes that are in the transmission range of the actual node and which are used for routing are made to sleep to avoid idle listening to reduce energy wastage. So the node which is in the next ring and closest to sink is selected in its own range and remaining nodes are made to sleep. The role of MAC layer is that it always assigns the ring no whenever a node is deployed and while accessing the nodes for consideration of next candidate this information from the MAC layer is used. The role of routing layer is that it uses the range, distance table and information from MAC layer to choose the next hop candidate. And it also uses the distance table to determine the in range nodes. It is assumed that the sensor nodes collect sensor data and transmit the data in a packet once in every period T. The energy model that the nodes followed contained the following states: transmission power, reception power, idle/listening power, sleep power and state transition power. The following equation governs energy consumption E at a sensor node.

$$E(EEIMRP) = E_t + E_s + E_r$$

$$E(IMRP) = E_t + E_r + E_i$$

$E_r$  = Energy of reception

$E_i$  = Energy of idle listening

$E_s$  = Energy of sleeping nodes

$T_{sa}$  = Survey time for all awake nodes in range without node sleep concept in IMRP.

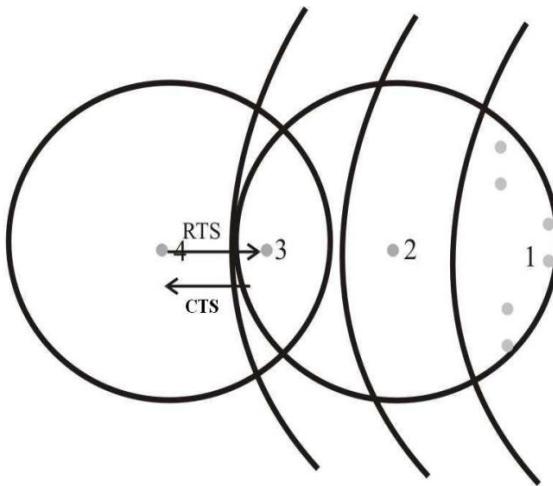
$T_{ss}$  = Survey time for all awake nodes in range with node sleep concept in EEIMRP.

When a source node using routing layer establishes a routing path in its request zone using EEIMRP, MAC layer protocol allows only nodes belonging to routing path to be awake while allowing other nodes to sleep mode in that zone.

### **III. REDUCTION OF ENERGY CONSUMPTION AND END TO END DELAY IN EEIMRP**

The main purpose of using RTS packet is to start data exchange and to silence all nodes around sender and data packet is transmitted from the node in outer ring to sink in innermost ring. The rings are numbered 1, 2, 3..... starting from the innermost ring.

In the listen period the node continuously listen to the medium and exchange SYNC, RTS and CTS packets. In this case synchronization of nodes is carried out using SYNC packets .The node which has data to send first contend for medium for transmission of SYNC packet and after that it follows the RTS/CTS/Data/Ack handshake mechanism. In SMAC idle listening is high because nodes keep on listening even if no packets have been transmitted until the end of RTS period or until they detect an RTS messages addressed to another node. In comparison to earlier protocols our protocol leaves in active mode only nodes belonging to routing path from source to sink and puts the remaining of neighboring nodes into sleep mode thus reducing idle listening by the nodes not belonging to routing path reduces energy consumption. Our protocol puts the ring identification number in the SYNC packet .The node which is in next ring receives RTS packet will send back CTS packet as response as shown in Figure 3. The format of SYNC packet in EEIMRP is shown in Table1. Our protocol leaves in active mode only nodes belonging to routing path from source to sink and puts the remainder of neighboring nodes into sleep mode.



**Figure 3: Idle listening avoidance in EEIMRP protocol.**

**TABLE I**  
**STRUCTURE OF SYNC PACKET**

<i>Field</i>	<i>Comment</i>
Type	SYNC packet
Length	Fixed size with 9 bytes
Source Address	ID of sender
Sync Node	ID of sender's synchronization node
Sleep time	Sender's sleep time from now
RI	Ring Identification number
CRC	Cyclic Redundancy check

**Table II**  
**STRUCTURE OF RTS AND CTS MESSAGES**

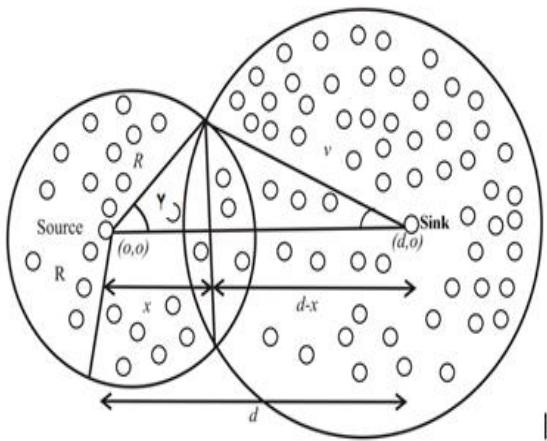
Frame control	Packet type	Previous node Address	Next Node Address	Current ring number
---------------	-------------	-----------------------	-------------------	---------------------

Frame control	Packet type	Previous node Address	Next Node Address	Current ring number
---------------	-------------	-----------------------	-------------------	---------------------

Table 2 defines the packet format for RTS and CTS messages. This packet contains five basic fields. The first field is frame control bits, second field contains the type of packet, and third field is source node address. The fourth field is destination node address .As multi ring topology is used fifth field defines current ring number.

#### **IV.NETWORK CONNECTIVITY IN EEIMRP PROTOCOL**

It is assumed that the communication range of the sensor node is a disk of radius R as our protocol is using ring based routing algorithm shown in Fig. 4. Minimum node density in a region of interest is calculated using 2D Poisson Point process given in equation 1. The probability distribution of a Poisson random variable is a discrete probability distribution for the counts of events that occur randomly in a given space.



**Figure 4: Calculating area of segment and void probability in EEIMRP protocol.**

Let N=Number of nodes in given area

Then  $\lambda$  is the mean number of nodes with in unit area.

The probability of observing n nodes in a given area is given by

$$P(N = n) = \frac{e^{-A\lambda} (A\lambda)^n}{n!} \quad \dots \quad (1)$$

The equation of two circles is

$$x^2 + y^2 = R^2 \quad \dots \quad (2)$$

$$(x-d)^2 + (y)^2 = v^2 \quad \dots \quad (3)$$

Combining equation (2) and (3)

$$(x-d)^2 - x^2 = v^2 - R^2 \quad \dots \quad (4)$$

Rearranging gives

$$x^2 - 2dx + d^2 - x^2 = v^2 - R^2 \quad \dots \quad (5)$$

This equation gives equation of line which is passing through two intersection points.

$$x = \frac{d^2 - v^2 + R^2}{2d} \quad \dots \quad (6)$$

Solving for y results

$$y^2 = R^2 - x^2 \quad \dots \quad (7)$$

Solving for y and putting the value of y in to chord length given by

$$a=2y$$

$$a = \frac{\sqrt{(-d+v-R)(-d-v+R)(-d+v+R)(d+v+R)}}{d} \quad \dots \quad (8)$$

Area of circle's sector which is given by  $A = \frac{1}{2}R^2 \cos^{-1}\left(\frac{d}{R}\right)$  ----- (9)

Area of sector segment=Area of sector-Area of triangle

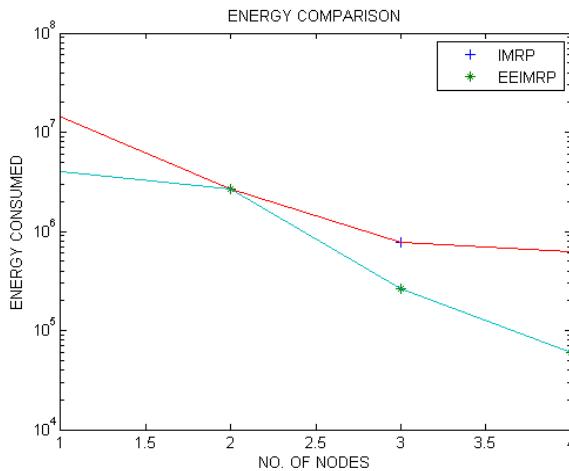
Thus Area of segment =

$$v^2 \cos^{-1}\left(\frac{d^2 + v^2 - R^2}{2dv}\right) + R^2 \cos^{-1}\left(\frac{d^2 + R^2 - v^2}{2dR}\right) - \frac{1}{2} \sqrt{(-d + r + R)(d + r - R)(d - r + R)(d + r + R)}$$

----- (10)

## V. SIMULATION RESULTS AND DISCUSSIONS

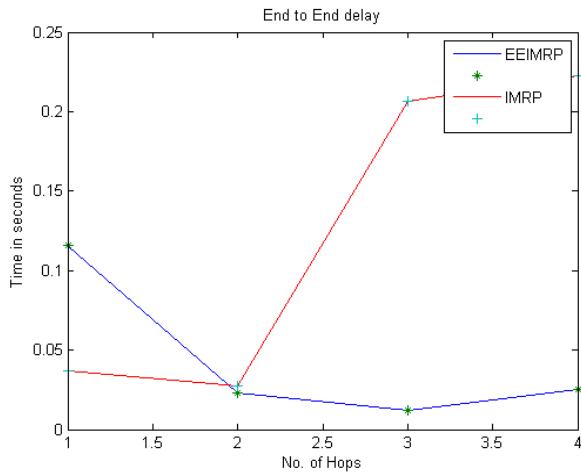
The graphs in Fig 5 shows the energy consumption as the packets move from outside the ring to sink with respect to number of nodes as the packets move to next hop node in next ring the node count increases .The results shows a better performance of EEIMRP as less energy is consumed in this case as compared to IMRP.



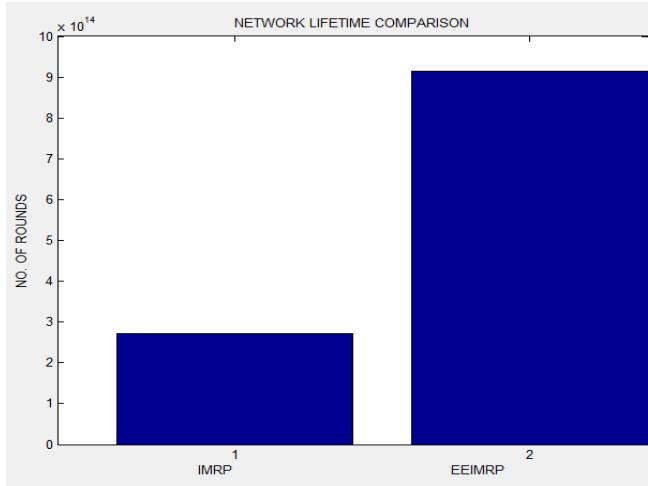
**Figure 5: Comparison of the energy consumed in proposed MAC and Network cross layer Protocol (EEIMRP) with conventional routing protocol (IMRP) based on ring structure**

c

Fig 6 compares average end to end delay of cross layer EEIMRP and IMRP with varying number of nodes .It can be seen from the results that average end to end delay of cross layered EEIMRP protocol is significantly less as compared to IMRP protocol. As shown in Fig 6 delay in IMRP increases sharply at a certain point for passing a data packet across two hops in IMRP as traffic in the network increases whereas in case of EEIMRP delay is very less for passing packets across the 4 hops.



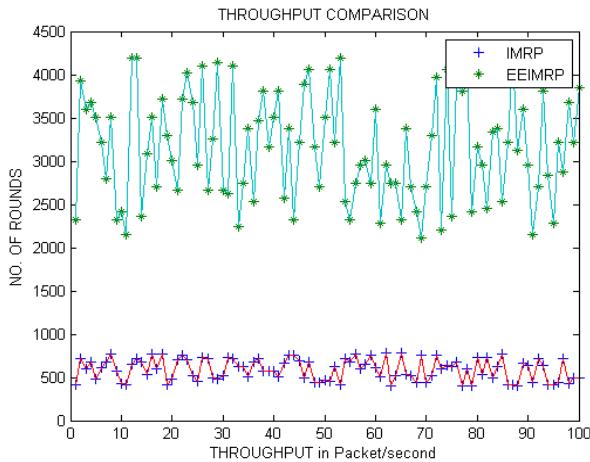
**Figure 6:** Comparison of the time delay in proposed MAC and Network cross layer Protocol (EEIMRP) with Conventional routing protocol (IMRP) based on ring structure.



**Figure 7:** Comparison of the time delay in proposed MAC and Network cross layer Protocol (EEIMRP) with conventional routing protocol (IMRP)

“Fig 7” show the network lifetime comparison and it is clearly shown that network lifetime of EEIMRP is much more than IMRP protocol. Network lifetime of EEIMRP protocol is much better and as shown nodes remains alive for  $9 \times 10^{14}$  rounds but in case of IMRP protocol nodes remain alive only up to  $3 \times 10^{14}$  rounds.

Throughput is defined as the number of packets delivered successfully over a channel to the destination per unit of time. Fig 8 shows the throughput comparison and simulations show that throughput in case of EEIMRP is much better as compare to IMRP protocol. As shown in Fig 8. throughput in case of EEIMRP goes up to 5000 rounds and in case of IMRP throughput goes up to 1000 rounds only.



**Figure 8: Comparison of the throughput in proposed MAC and Network cross layer Protocol (EEIMRP) with conventional routing protocol (IMRP).**

## VI. Conclusion

In this paper, a ring based routing protocol EEIMRP is proposed for wireless sensor networks. Each ring has given ring identification number starting from the innermost ring. Cross layer Integration of MAC and routing EEIMRP protocol is the method to reduce protocol overhead as the routing information from routing layer is used by MAC layer to find the duty-cycle of each node, this protocol also reduces energy wastage due to collision and overhearing problem in the MAC layer based on adaptive sleeping. When a source node finds a shortest routing path based on ring based structure in its request zone using EEIMRP protocol and it allows only nodes belonging to shortest routing path to be awake while permitting other nodes in the zone to sleep mode. Simulation results shows that the performance of EEIMRP protocol improves energy efficiency and prolonged network lifetime as compare to IMRP protocol.

## REFERENCES

1. S. Guo, Y. Gu, B. Jiang, and T. He, "Opportunistic floods in low-duty-cycle Wireless sensor networks with unreliable links," in Proceedings of the 15th annual international conference on Mobile computing and networking ,ACM , 2009, pp.133–144.
2. Jurdak, P. Baldi, and C. V. Lopes, "Adaptive low power listening for Wireless sensor networks," IEEE Trans. Mob.Comput., vol. 6, no. 8, 2007, pp. 988–1004.
3. J. Polastre, J. L. Hill, and D. E. Culler, "Versatile low power media access for Wireless sensor networks," in SenSys, 2004, pp. 95–107.
4. M. Buettner, G. V. Yee, E. Anderson, and R. Han, "XMAC: a short preamble MAC protocol for duty Cycled Wireless sensor networks," in SenSys, 2006,pp. 307–320.
5. Khokhar, M.S. Hefeeda, T. Canli., "CI-MAC: A cross-layer Mac protocol for heterogeneous Wireless sensor networks. Elsevier Ad Hoc Networks," 2012, Vol. 11, pp.213–225.
6. S. Tilak, N. B. A.Ghazaleh, and W. Heinzelman,"A taxonomy of Wireless micro sensor Network models," ACM Mobile Comput. Commun. Rev, vol. 6, 2002, no. 2.
7. S. Kusumamba , S.M. D. Kumar, "A Reliable Cross Layer Routing Scheme (CL-RS) for Wireless Sensor Networks to Prolong Network Lifetime", IEEE International Advance Computing Conference (IACC),2015, pp.1050-1055.
8. R. Jurdak, P. Baldi, and C. V. Lopes, "Adaptive low power listening for Wireless sensor Networks," IEEE Trans. Mob. Computing, vol. 6, no. 8, 2007, pp. 988–1004.
9. T.H.Hsieh, K.Y. Lin, P.C. Wang, "A Hybrid MAC Protocol for Wireless Sensor Networks" Proceedings of 2015 IEEE 12th International Conference on Networking, Sensing and Control Howard Civil Service International House, Taipei, Taiwan, 2015.
10. P. Gonizzi, P.Medagliani, G. Ferrari, J. Leguay , "RAWMAC: A Routing aware Wave-based MAC Protocol for WSNs" IEEE The second international Workshop on Green optimized Wireless Networks ,2014,pp 1-8.
11. K. Heurtefeux, F. Maraninchi, F. Valois, "AreaCast: A Cross-Layer Approach for a Communication by Area in Wireless Sensor Networks" ICON 2011 978-1-4577-1826-11/\$26 © IEEE 2011.

12. S. Ehsan, B. Hamdaoui, and M. Guizani, "Radio and Medium Access Contention Aware Routing for Lifetime Maximization in Multichannel Sensor Networks," IEEE Transactions on Wireless Communications, 2012, Vol. 11, NO. 9.
13. P. Skraba, H. Aghajan, and A. Bahai, "Cross-Layer Optimization for High Density Sensor Networks: Distributed Passive Routing Decisions," Proc. Int Conf. Ad-Hoc, Mobile and Wireless Networks, Nov 2004, Vancouver.
14. R.L. Cruz and A.V. Santhanam , " Optimal Routing, Link Scheduling and Power Control in Multi-Hop Wireless Networks," Proc. 0-7803-7753-2/03 INFOCOM 2003.
15. ElBatt and A. Ephremides, "Joint Scheduling and Power Control for Wireless Ad Hoc Networks," IEEE Trans. WirelessCommunicatons, vol. 1, 2004, pp. 74-85.
16. U.C. Kozat, I. Koutsopoulos , and L. Tassiulas , "A Framework for Cross-Layer Design of Energy-Efficient Communication with QoS Provisioning in Multi-Hop Wireless Networks," Proc. INFOCOM, 2004, Honk Kong, pp. 1446–1456 .

# A Review on security issues on routing protocols in Delay Tolerant networks

## 1. Swati

Research scholar  
Punjabi University  
Patiala

Email:er.swati.jindal87@gmail.com

## 2. Dr.Jagtar Singh

Associate professor,ECE department  
Ycoe talwandi sabo Punjabi university Patiala  
Email: jagtarsivian@yahoo.com

## 3. Dr.Harminder Singh Bindra

Assistant Professor,IT department  
MIMIT,Malout

Email: bindra.harminder@gmail.com

**Abstract:** In modern world, Delay Tolerant Network (DTN) has significant role in communications. DTN's are used in various applications like military wars and conflicts, earthquakes, volcanic eruptions, terrorist attacks etc.DTN provides environment where two nodes can only exchange messages when they move into the transmission range of each other due to Intermittent Connectivity in the network. Security is main challenge in these types of networks. In this paper we have elaborated various DTN routing protocols and its variants along with security issues.

## 1 Introduction

DTN used in various situations like wild life tracking systems, traffic controlling systems, military wars, attacks of terrorist, earthquakes, floods, storms, hurricane, rigorous volcanic eruptions, etc .These types of challenging conditions results unwarranted delays, severe bandwidth limitations, significant node mobility, regular power outages and frequent communication difficulties. Therefore, under such conditions wireless networks connectivity becomes considerably irregular and the existence of simultaneous end-to-end connectivity between any source-destination pair can no longer be assured. So, many researchers are carrying out work in this area by considering these issues.

DTN have properties like High latency, low data rate, disconnection, long queuing delay, restricted longevity, limited resources, irregular Connectivity, Long, Variable Delay, High Error Rates.

Depending on the current conditions of the demanding networks, this paper has discussed the various routing protocols in DTN and their security issues. In section 2 related work of previous ten years has been discussed, in section 3 three types of routing protocols Predicting Good Forwarders, opportunistically forwarding messages and Meeting the destinations by schedule have been described. Then in section 4 security in these routing protocols with various attacks and mechanisms that degrade the performance of a network has been discussed .Section 5

overviewed prophet routing protocol and its variants and the security issues. Section 6 concludes the complete paper.

### 1.1. Architecture of DTN

The RFC4838 (DTN Architecture [3]) expected a general architecture to overcome all the previous challenges for message store and forward switching. This switching is based on asynchronous messaging and uses postal mail as a replica of service classes and delivery semantics. Blocks of user data, called bundles, are forwarded from the source to storage on another node that custodians nodes assumes the responsibility for consistent delivery of the bundle to its destination. The bundle would move along a way of custodians that eventually reaches the destination store.

## 2. Related work

All the presented ad hoc routing protocols have assumed that network is healthy to rapidly changing network topology and also assumed the presence of a connected path from source to destination. A technique developed by A. Vahdat et al. to deliver messages when there is no guarantee of a connected path from source to destination (DTN) or when a network partition exists at the time a message is originated [10]. The author introduced Epidemic Routing protocol, exchanges of messages among moveable nodes ensure final message delivery.

A. Lindgren et al. proposed PRoPHET, a probabilistic routing protocol in which message transferred is based on delivery probability of nodes [8].

T. Spyropoulos et al. introduced a new routing scheme, called Spray and Wait that “sprays” many copies of message into the network, and then “waits” till one of these nodes reach to the destination [18].

J. Burgess et al. proposed MaxProp, a routing protocol of DTN. This protocol is based on prioritizing both the schedule of packets sending out to other nodes and packets to be dropped [16].

A. Balasubramanian et al. presented RAPID, a deliberate DTN routing protocol that can optimize a definite routing metric such as worst-case delivery delay or the part of packets that are delivered within a deadline [22].

Lei Tang proposed a DTN routing protocol SMART.SMART used attendant of the destinations (i.e. nodes that commonly meet the destination) to increase the delivery chances[41].

## 3. Routing Protocols

Routing protocols can be classified into three categories according to the mechanisms used to find path from source to destination: predicting Good Forwarders, opportunistically forwarding messages, meeting the destinations by schedule. All types of protocols are explained in table 1.

**Table 1.** Routing protocols and mechanism

S.NO	YEAR OF PROTOCOL PROPOSED	NAME OF PROTOCOL/Mechanism	WORKING
1	2012	ITRM(iterative reputation management)[48]	Graph based iterative algorithm motivated by the prior success of message passing techniques for decoding low-density parity-check codes over bipartite graphs.
2	2011	SPREAD (countermeasure against	A localised solution that assesses evidence of spoofing and offers

		SPoofing by REplica ADjustment)[47]	countermeasures designed for quota-based multi-copy routing protocols
3	2011	Control message redundancy mechanism [37]	This mechanism is by using counter method, every node adds an encounter counter based on epidemic routing scheme
4	2010	Independent message deletion mechanism[38]	This mechanism is for multicopy routing schemes.
5	2008	Multi-copy spraying algorithm[43]	The number of message copies in the network depends on the urgency of meeting the delivery deadline for that message.
6	2008	Extended Epidemic Routing[39]	It proposed to include immunity based information disseminated in the reverse direction once messages get delivered to their destination.
7	2007	PRioritized EPidemic (PREP)[40]	PREP prioritizes bundles based on costs to destination, source, and expiry time
8	2007	SMART[41]	SMART uses travel companions of the destinations (i.e. nodes that frequently meet the destination) to increase the delivery opportunities.
9	2007	Rapid[22]	An intentional DTN routing protocol that can optimize a specific routing metric such as worst-case delivery delay or the fraction of packets that are delivered within a deadline.
10	2007	Spray And Focus Routing[42]	A scheme that also distributes a small number of copies to few relays alike spray and focus. However, each relay can then forward its copy further using a single-copy utility-based scheme, instead of naively waiting to deliver it to the destination itself
11	2006	Max Prop[16]	MaxProp is based on prioritizing both the schedule of packets transmitted to other peers and the schedule of packets to be dropped.
12	2005	Spray and Wait[18]	“sprays” a number of copies into the network, and then “waits” till one of these nodes meets the destination.
13	2003	PROPHET [8]	A probabilistic routing protocol for such networks
14	2000	Epidemic[10]	Random pair-wise exchanges of messages among mobile hosts ensure eventual message delivery. The goals of Epidemic Routing are to: i) maximize message delivery rate, ii) minimize message latency, and iii) minimize the total resources consumed in message delivery.

**Predicting Good Forwarders:** This scheme try to foretell the nodes which are useful for sending the messages based on node's history-encounter information [8], [15], [16] framework information [19] or location visiting sample [14], [20].

These routing protocols include MobiSpace [14], MV [15], Seek and Focus [17], CAR [19], and MaxProp [16], RAPID [22], PROPHET [8], SMART [41]. All these schemes attempts to predict constructive nodes for delivery on the basis of the past nodes' encountering history or the circumstance information such as remaining battery lifetime.

**a) Meeting the destinations by schedule:** In this W. Zhao purposed Message Ferry (MF) a representative scheme [12]. In MF scheme, there are special types of nodes called ferries which are able to change their routes in advance to help other nodes to send messages. In addition, Tariq et al. regularize ferry go across route to encounter concerned nodes with a certain minimum probability [21].

**b) Opportunistically forwarding messages:** Protocols in this scheme such as Epidemic [10-11], [17-18] opportunistically forward messages to other nodes until the messages reach their destinations.

Another protocol PRioritized EPidemic (PREP) prioritizes packets based on costs to destination, source, and expiry time. Costs are derived from per-link “average availability” information that is distributed in an epidemic manner. PREP sustains a incline of copying density that decreases as the distance from the destination is increasing.

P. Mundur et al. [39] modified and extended epidemic routing in DTNs. They proposed to include immunity based information distributed in the reverse direction once messages get delivered to their destination. This protocol uses an immunity-list for the information regarding delivered messages that will prevent any future exchange of those messages. Therefore, it outperforms the basic epidemic protocols in terms of delivery ratio and delay.

To resolve the issues regarding overhead of message in the epidemic protocol, many protocols are there, such as Spray and Wait [18] protocol and Prophet Protocol [8]. In these protocols, message copies present are limited by some conditions rather than flooding of messages. Prophet, a probabilistic routing protocol for DTN networks in which according to delivery probability of nodes messages are transferred with a lower overhead of communication. Spray routing protocol sends a fixed number of copies of each message in network as epidemic routing protocol. Spray and Wait performs better than all presented routing protocol in case of both average message delivery delay and number of transmissions per message delivered [18]. Erasure-coding Based Routing (EBR) [11] splits a message into a set of code blocks, which are “sprayed” to a set of relays. Any acceptably large subset of the generated code blocks can be used to reconstruct the new message. Data MULE Routing [9] proposed by Rahul et al. develops the randomly-moving mobile nodes (MULEs) to send messages in a sparse sensor network, which accept messages from still sensors when in close range. After that they buffer the received messages and drop them off to wired contact points when in closeness.

A multi-copy routing protocol an independent message deletion mechanism [38] for Delay Tolerant Networks (DTNs) was proposed in 2010. This method can improve the resource utilization and message delivery. In 2011, a new scheme [37] proposed to control message redundancy by encountering a counter(records the number which the node encounters other nodes with the same message copy) based on epidemic routing protocol. Node removes the copy if the counter accomplished the installed threshold.

Therefore various routing protocols used in DTN are different in overhead ratio, delivery probability, average message delay etc. Overhead ratio in Epidemic protocol is more and 100 percent delivery ratio is also provided by it. But other protocols are using some mechanisms to limit message copy forwarding to decrease overhead ratio.

#### 4. Security In Routing Protocols in DTN

Performance of the DTN can also be affected by various attacks from malicious nodes. These attacks need to be identified. A lot of research on security in MANET routing has already been done [23-28] but cannot be used for securing DTN as DTN routing is typically opportunistic with Intermittent Connectivity. Many approaches for security in routing for DTN depend on public key and policy based cryptography to bound participants to a set of allowed nodes. The allotment of space and link capacity has been chosen depending upon the type of service [29-32]. Such mechanisms include validating every routing metadata and packets at every in-between hop, gaining considerable handing out overhead. But key management may not be easy to carry out under certain trust methods and situations, and is problematical by the irregular connectivity of DTN.

##### a) General Attacks

Security is challenging issue in DTN network which is affected by the various possible attacks by the intruders. Therefore various attacks have been identified by various researchers.

Four general attacks ***Drop All***: Drop all the incoming packets, ***Random flooding***: attacker send multiple packets at a time, ***Invert routing metadata***: In this attack data is transferred or dropped in reverse order ***and Acknowledgement counterfeiting***: Attackers send false acknowledgements [33].

Above attacks may be hopeless, many distinctions of these attacks were still possible. Therefore Fai Cheong et.al proposed combination and variant of these attacks like ***Non-Deliverable Packet Flooding***: floods data to nodes who are not in existence and ***Identity Impersonation***: imitates different individualities to act as destinations for packet [34].

In addition to various attacks some other attacks were defined as follow: ***Self-promoting attacks***: it can promote itself by providing good reputation of itself. ***Bad-mouthing attacks***: it can spoil the status of good nodes by bad recommendations against good nodes ***Ballot stuffing***: it can improve the reputation of dreadful nodes so that packet can be sent through it.

So a Trust based mechanism has been proposed to deal with network spoiling nodes [44].

A Reputation based mechanism for forwarding of messages is based on reputation of node against black hole attacks [46].

A two period routing approach which targets at improving the favored delivery ratio by deadline in presence of malevolent nodes was proposed by Bulet et.al [45].

In 2011 a author proposed a mechanism for Making DTNs vigorous Against Spoofing Attacks with Localized Countermeasures using SPREAD (countermeasure against Spoofing by REplica ADjustment), a solution that appraise evidence of spoofing and provides countermeasures considered for quota-based multi-copy routing protocols without using any network wide authentication procedure[46].

ITRM [48] is the planned graph scheme is motivated by the prior success of message passing mechanisms for decoding low-density parity-check codes over bipartite graphs based on iterative algorithm.

In 2012, a trust framework for DTN was purposed in which the trace based mobility model is followed by the nodes in the network. The selection of next hop to forward the data packets is based on the trust value as well as the direction of movement of node towards the destination. From the trust framework of the data forwarding node trust value is called [49].

A Secure User-centric and Social-aware Reputation based Incentive Scheme for DTNs (SUCCESS [50]) in this node can manage its reputation proof by showing its reputation value.

Give2Get [51] Forwarding is helpful in Social Mobile Wireless Network of Selfish Individuals. In this work selfishness is considered as nodes are loyal to only the same community but not to others

## 5. Prophet Routing Protocol and Security:

To overcome the problems of Epidemic protocol, a probabilistic routing protocol (PRoPHET) for DTN was purposed in 2003 [8]. In this protocol, messages are transferred in the network based on delivery predictability (DP) for successful delivery to the destination. DP is calculated as the probability of node using history of participation of

the node. Therefore overhead ratio is less than epidemic protocol. The problem with this protocol was that jitter (fluctuation of DP) arises when DP reduces suddenly due to network fault and increases sharply if two nodes encounter again and again. An advance PRoPHET protocol has been implemented by considering average delivery probability instead of delivery probability to overcome the drawback of jitter [55].

An improved version of PRoPHET was PRoPHET+ which was based on qualitative considerations of weights [58]. Because of this feature PRoPHET+ perform more efficiently in different environments. Many other problems were analyzed with basic PRoPHET and are described as under:

1. It was experienced from N4C deployment that investigated DPs for nodes were very high that did not support network topology the reason behind was the frequency of encounters is not evenly spread over the network which results quick fluctuation of DP
2. Second problem was parking lot problem Reconnection in network is considered as new encountered nodes so DP for these encounters increased too much so it destroy the mobility model.
3. Transitivity property is not considered in the mechanism of PRoPHET, as value of this factor is considered as 0[49].

To deal with the discussed problems the various evolutions of PRoPHET DTN Routing Protocol were proposed:

PRoPHETv2 was implemented in which minor modifications to the routing metric calculation has been done which results better performance than prophet [57].

A distance based PRoPHET routing is implemented. Distance between two neighbor nodes is calculated by drawing the delivery predictability value from fluctuating value and fault forwarding decision [56].

An improved PRoPHET routing protocol was proposed that has improved dissemination speed of Prophet Protocol by employing epidemic protocol for disseminating message m, if the threshold values are larger or equal to forwarding counter and hop counter values [57].

As per security is a concerned malicious node can affect the performance of PRoPHET protocol by increasing its probability anyhow?

3PR Privacy Preserving Prediction based Routing protocol was implemented. In which message is transferred between communities having more probability to send messages [52].

By considering more security issues a probabilistic misbehavior detection scheme was simulated [53], as existing work Consider only either of misbehavior detection or incentives scheme, but this mechanism has considered both jointly.

A risk taking routing algorithm was implemented in which forwarding is done purely on prediction basis rather than presently available nodes if no node having good probability is present then forwarding will be delayed [54].

## 6. Conclusion

In this review paper various routing protocols in DTN have been discussed. It has been analyzed that protocols under the category Predicting Good Forwarders outperform the other protocols like Opportunistically forwarding messages. Various security issues, attacks have been described that may affect the performance of routing protocols.

PRoPHET and its variants have been explored. Various attacks need to be still identified that can degrade the performance of PRoPHET routing protocols and the mechanisms that can make the attacks ineffective.

## 7. References

- [1] Cerf V and Kahn R, “A Protocol for Packet Network Intercommunication”, IEEE Trans. Communications, vol. 22, no. 5, May 1974 pp. 637–648.
- [2] Fall K, “A Delay-Tolerant Network Architecture for Challenged Internets”, SIGCOMM, August 2003.
- [3] Cerf V, “Delay-Tolerant Network Architecture, IETF RFC 4838, Informational”, April 2007.
- [4] Perkins C, Belding-Royer E, and Daas S, “RFC 3561: Ad hoc On-Demand Distance Vector Routing”, Jul 2003.
- [5] Johnson B, Maltz, A and Hu H, “DSR-draft: The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)”, April 2003.
- [6] Johnaon B and Maltz A, “Dynamic Source Routing in Ad Hoc Wireless Networks” in Mobile Computing, pp. 153–181,1996.
- [7] Perkins C and Royer E, “Ad hoc On-demand Distance Vector Routing”, in 2nd IEEE Workshop on Mobile Computing Systems and Applications, 1999.
- [8] Lindgren A,Doria A, and Schelen O, “Probabilistic Routing in Intermittently Connected Networks”, in SIGMOBILE Mobile Computing Communications Review,19–20 july 2003 pp. 7.
- [9] Shah R, Roy S, Jain S, and Brunette W, “Data MULEs: Modeling a Three-tier Architecture for Sparse Sensor Networks”, in IEEE SNPA, 2003.
- [10] Vahdat A and Becker D, “Epidemic Routing for Partially-connected Ad hoc Networks”, in Technical report, Duke University, 2000.
- [11] Wang Y, Jain S, Martonosi M, and Fall K, “Erasure Coding based Routing in Opportunistic Networks”, in ACM SIGCOMM Workshop on Delay Tolerant Networking, 2005.
- [12] Zhao W, Ammar M, and Zegura E, “A Message Ferrying Approach for Data Delivery”, in Sparse Mobile Ad Hoc Networks,” in MobiHoc, 2004.
- [13] Jain S, Fall K, and Patra R, “Routing in a Delay Tolerant Network,” in ACM Sigcomm, 2004.
- [14] Leguay J, Friedman T, and Conan V, “DTN Routing in a Mobility Pattern Space”, in ACM SIGCOMM - Workshop on delay tolerant networking and related topics (WDTN-05), 2005.
- [15] Burns B,Brock O, and Levine B, “MV Routing and Capacity Building in Disruption Tolerant Networks”, in IEEE Infocom, 2005.
- [16] Burgess J, Gallagher B,Jensen D, and Levine B, “Maxprop: Routing for vehicle-based disruption-tolerant networking”, in IEEE Infocom, 2006.
- [17] Spyropoulos T,Psounis K, and Raghavendra A, “Single-copy routing in intermittently connected mobile networks”, in Proc. of IEEE Secon’04, 2004.
- [18] Spyropoulos T, Psounis K, and Raghavendra C, “Spray and wait: an efficient routing scheme for intermittently connected mobile networks”, in WDTN ’05,pp. 252–259,2005.

- [19] Musolesi M, Hailes S, and Mascolo C, “Adaptive Routing For Intermittently Connected Mobile Ad hoc Networks”, in WOWMOM’05, 2005.
- [20] Ghosh J, Ngo H, and Qiao C, “Mobility profile based routing within intermittently connected mobile ad hoc networks (icman)”, in IWCMC ’06,pp. 551–556,2006.
- [21] Tariq M, Ammar M, and Zegura E, “Message ferry route design for sparse ad hoc networks with mobile nodes”, in MobiHoc ’06,2006 pp. 37–48.
- [22] Balasubramanian A, Levine B, and Venkataramani A, “DTN Routing as a Resource Allocation Problem”, in Proc ACM Sigcomm, Kyoto, Japan, [Online] August 2007. Available: <http://www.sigcomm.org/crc/drupal/?q=node/273>
- [23] Djenouri D and Badache N, “Struggling against selfishness and black hole attacks in manets”, Wirel. Commun. Mob. Comput., vol. 8, no. 6, 2008 pp. 689–704.
- [24] D. Hongmei, L. Wei, and A. Dhama P., “Routing security in wireless ad hoc networks”, IEEE Communications magazine, October 2002.
- [25] Hu Y, Perrig A, and Johnson D, “Packet leashes: a defense against wormhole attacks in wireless networks”, vol. 3, pp. 1976–1986 vol.3, 2003.
- [26] Cristina and Rubens H, “An on-demand secure routing protocol resilient to Byzantine failures”, in ACM Workshop on Wireless Security (WiSe), Atlanta, Georgia, September 2002.
- [27] Yi S,Naldurg P, and Kravets R, “A security-aware routing protocol for wireless ad hoc networks”, in in: Proceedings of ACM Symposium on Mobile Ad Hoc Networking and Computing (Mobicoc), 2001 pp. 286–292.
- [28] Hu Y, “Rushing attacks and defense in wireless ad hoc network routing protocols”, in in ACM Workshop on Wireless Security (WiSe),pp. 30–40,2003.
- [29] Fall K, “A delay tolerant network architecture for challenged internets”, in Proc. of Annual Conf. of the Special Interest Group on Data Communication (ACM SIGCOMM’03), August 2003 pp. 27–34.
- [30] Durst R, “A infrastructure security model for delay tolerant networks,” july 2002.
- [31] Seth A and Keshav S, “Practical security for disconnected nodes”, in Proceedings of the 1st IEEE ICNP Workshop on Secure Network Protocols, 2005.
- [32] Kate A, Zaverucha G, and Hengartner U, “Anonymity and security in delay tolerant networks”, in Security and Privacy in Communications Networks and the Workshops, 2007. SecureComm 2007. pp. 504–513,2007.
- [33] Burgess J, Bissias G, Corner M, and Levine B, “Surviving attacks on disruption-tolerant networks without authentication”, in MobiHoc ’07: Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing. Montreal, Quebec, Canada: ACM Press,2007 pp. 61–70.
- [34] Choo F, Chan M,Chang E, "Robustness of DTN against routing attacks", Communication Systems and Networks (COMSNETS), 2010 Second International Conference on , vol., no, 5-9 Jan. 2010 pp.1-10.
- [35] ANEN K, Opportunistic Network Environment Simulator. Special Assignment report, Helsinki University of Technology, Department of Communications and Networking, May 2008.
- [36] TKK/COMNET. Project page of the ONE simulator. <http://www.netlab.tkk.fi/tutkimus/dtn/theone>, 2009.

- [37] YU H, MA J, BIAN H, "Message redundancy removal of multi-copy routing in delay tolerant MANET", The Journal of China Universities of Posts and Telecommunications, Volume 18, Issue 1, ISSN 1005-8885, February 2011 Pages 42-48.
- [38] Thompson N, Nelson S, Bakht M, Abdelzaher T and Kravetsi R "Retiring Replicants Congestion Control for Intermittently-Connected Networks", In main Technical Program at IEEE INFOCOM 2010.
- [39] Mundur P, Seligman M, Lee G , "Epidemic routing with immunity in Delay Tolerant Networks", *IEEE Military Communications Conference, 2008. MILCOM 2008.*, pp.1-7, 16-19 Nov. 2008
- [40] Ramanathan R, Hansen R, Basu P, Rosales-Hain R, and Krishnan R. "Prioritized epidemic routing for opportunistic networks", In *Proceedings of the 1st international MobiSys workshop on Mobile opportunistic networking (MobiOpp '07)*. ACM, New York, NY, USA, 2007, pp 62-66.
- [41] Tang L, Zheng Q, Liu J; Hong X, "SMART: A selective controlled-flooding routing for delay tolerant networks", *Fourth International Conference on Broadband Communications, Networks and Systems, 2007. BROADNETS 2007.*, pp.356-365, 10-14 Sept. 2007
- [42] Spyropoulos T, Psounis K, and Raghavendra S "Spray and Focus: Efficient Mobility-Assisted Routing for Heterogeneous and Correlated Mobility", In Proceedings of the Fifth IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOMW '07). IEEE Computer Society, Washington, DC, USA, 2007, 79-85.
- [43] Bulut E, Wang Z and Szymanski K, "Minimizing Average Spraying Cost for Routing in Delay Tolerant Networks", In Proceedings of 2nd Annual Conference of International Technology Alliance, ACITA 2008, London, UK, September 2008.
- [44] Chen I, Bao F, Chang M, and Cho J, "Dynamic Trust Management for Delay Tolerant Networks and Its Application to Secure Routing", published in ,IEEE transaction on parallel and distributed systems, April 2009, PG-99.
- [45] Bulut E , Szymanski B "Secure Multi-copy Routing in Compromised Delay Tolerant Networks", published in wireless personal communication volume 73(1) November 2013 pp 149-168 .
- [46] Dini G, Duca A," Towards a reputation-based routing protocol to contrast blackholes in a delay tolerant network", 2012.
- [47] Uddin Y, Khurshid A, Jung H, Gunter C, Caesar M, Abdelzaher T, "Making DTNs Robust Against Spoofing Attacks with Localized Countermeasures in sensor mesh adhoc network and communication", 8<sup>th</sup> annual IEEE communication society conference.2011.
- [48] Ayday E , "An Iterative Algorithm for Trust Management and Adversary Detection for Delay Tolerant Networks", published in mobile computing IEEE transaction on vol 11 issue 9 in sept 2012 page 1514-1531.
- [49] Poonguzharselvi B, Vetriselvi V "Trust Framework for Data Forwarding in Opportunistic Networks Using Mobile Traces", International Journal of Wireless & Mobile Networks (IJWMN) Vol. 4, No. 6, December 2012.
- [50] WE L, SHEN X, " SUCCESS: A Secure User- centric and Social-aware Reputation based Incentive Scheme for DTNs" 2012,

- [51] Mei A and Stefa J," Give2Get: Forwarding in Social Mobile Wireless Networks of Selfish Individuals",
- [52] Hasan O, Miao J, Mokhtar S, Brunie L "A Privacy Preserving Prediction-based Routing Protocol for Mobile Delay Tolerant Networks", IEEE 27th International Conference on Advanced Information Networking and Applications 2013.
- [53] Zhu H, Dong M" A Probabilistic Misbehavior Detection Scheme toward Efficient Trust Establishment in Delay-Tolerant Networks" IEEE Transactions On Parallel And Distributed Systems, VOL. 25, NO. 1, 2014.
- [54] Barijough M, Yazdani N, Tavangarian D, Daher R, " A Risk Taking Routing Algorithm for Delay Tolerant Networks" IEEE 27th International Conference on Advanced Information Networking and Applications 2013.
- [55] Xue J, Li J, Cao Y, Fang J "Advanced PROPHET Routing in Delay Tolerant Network" 978-0-7695-3522-7 / 09 IEEE ICCSN.2009.44 413 2009 International Conference on Communication Software and Networks.
- [56] Sok P,Kim K, "Distance-based PRoPHET Routing Protocol in Disruption Tolerant Network" ICTC, 978-1-4799-0698-7/13/\$31.00 ©2013 IEEE.
- [57] Deok H, Chung Y" An Improved PRoPHET Routing Protocol in Delay Tolerant Network" Hindawi Publishing Corporation, e Scientific World Journal Volume 2015, Article ID 623090.
- [58] Huang T, Lee C, Chen L "PRoPHET+: An Adaptive PRoPHET-Based Routing Protocol for Opportunistic Network" 2010 24th IEEE International Conference on Advanced Information Networking and Applications.

# Evaluation of Efficiency and Security of Data using different Algorithms for Encryption and Key Management

Gagandeep

gaganmarken1990@live.com

*Student MTech, Department of Computer Engineering  
at Punjabi University Patiala, Punjab, India*

Gurjit Singh Bhathal

gurjit.bhathal@gmail.com

*Astt. Professor, Department of Computer Engineering  
at Punjabi University Patiala, Punjab, India*

Puneet Kumar

pkumar3397@gmail.com

*Department of Electronic and Communication Engineering  
Guru Nanak Dev University Regional Centre Gurdaspur, Punjab, India*

**Abstract--** In present scenario, security is the most challenging aspects which is used in cryptography algorithms to encrypt data which is transmitted through internet or any network application to secure data from various passive and active attacks comes out from other in-secure medium. In this survey, we have compares the performance and energy consumption by various different symmetric algorithms like AES, 3DES, DES, RC6, blowfish and RC2 on same input file size.

**Keywords**—Security, Encryption, Decryption, Advanced Encryption Standard, Triple Data Encryption Standard, Data Encryption Standard, RC2, Blowfish and RC6.

## I. Introduction

Cryptography defines to be an art or science of designing cryptosystems while Cryptanalysis refers to the science of breaking them. In present scenario, confidentiality of data is a major of all organization requirements and one can have techniques such as digital signatures and secret sharing. Cryptography has number of applications such as online transaction, secured funding and many more. Basically, it is a technique used to avoid unauthorized access of data[3, 13]. The basic cryptography model has been shown in the figure 1. A number of cryptographic algorithms are available such as DES, TDES, AES, Blowfish, RSA and MD5. The strength of these encryption algorithms depends upon their key strength. Solid encryption algorithms and optimized key management techniques always help in achieving authentication and integrity of data and reduce the overheads of the system. The long length key takes more computing time to crack the encryption model and it becomes difficult hacker to detect the cryptographic model [1, 13]. It has two main components, a) Encryption algorithm, and b) Key. Cryptography has two categories a) Symmetric Cryptography and b) Asymmetric

Cryptography [3]. In symmetric cryptography same key is used to encrypt and decrypt the message. Although, in asymmetric cryptography unique keys are used to encrypt and decrypt the message. Asymmetric algorithms are relatively slower than symmetric algorithms but provide a good security level [2, 12].

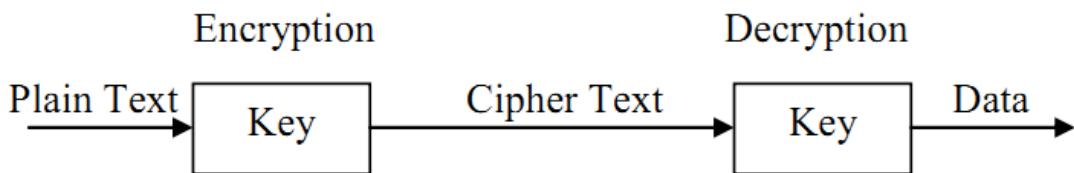


Fig 1. Cryptographic Model

## II. Literature survey

This section involves the work done by the various researchers in the field of cryptographic algorithm for data security. The literature survey has been carried out by dividing into two section have been calculated to formulate the problems and solutions observed in designing of an encryption algorithms.

### A. Symmetric key Encryption Algorithms:

This section describes the various encryption algorithm used for data security. Critical analysis has been done and finally observations have been drawn which are listed at the end of this section.

#### 1. DES (Data Encryption Standard)

It is a symmetric algorithm developed by IBM in 1977. It has key size of 56 bits and block size of 64 bits. It divides 56 bits block size into two 28 bits halves, each half of the key is shifted by one or two bits depending on the round [13]. It has 8 bit parity which has to be removed from the key by subjecting to the key permutation. It takes 16 rounds for the iterations which are based upon same ciphering-deciphering key. It has slow computation time as compared to the AES and RSA [1, 12].

#### 2. Blowfish

It is a symmetric algorithm designed by Bruce Schneier in 1994. It is a symmetric block cipher which could be effectively used for encryption. It takes a variable key length from 32 bits to 448 bits, making it ideal for secured communication. It contains of 14 or less rounds and treated as one of the fastest block ciphers. Blowfish divide the plain text into fixed length blocks during encryption and decryption [1].

#### 3. TDES (Triple Data Encryption Standard)

It is a symmetric algorithm developed in 1998. Triple DES undergoes 3 iterations for effectively encrypting data with 168 bits key size. The data is first encrypted through first 56 bits, decrypted with next 56 bits and finally, again encrypted with the 56 bits. Hence, TDES is a better symmetric algorithm which provides secure information. This algorithm increases the numbers of attempts for the users [2].

#### **4. AES (Advanced Encryption Standard)**

It is a symmetric algorithm introduced in 2001. It has 3 different key size which are as 128, 192, 256 bits and block size of 128 bits. A number of AES parameters depend on its key size. It encrypts the data blocks of 128 bits in 10, 12 and 14 rounds depending on key size. Encryption and decryption speed is faster than DES and RSA [1, 7]. Brute force attack is the only effective attack known against this algorithm. Power consumption is less than RSA.

#### **B. Asymmetric key Encryption Algorithms:**

##### **1. Diffie - Hellman Algorithm:**

Diffie-Hellman key exchange is a first asymmetric encryption algorithm of cryptographic keys which is proposed in 1976. It allows two parties that have no prior knowledge of each other to jointly establish a shared secret key over an insecure communications channel [15]. This key is used to encrypt subsequent communications using a symmetric key cipher.

##### **2. RSA (Rivest Shamir Adleman)**

It was introduced in 1978 by Ronald Rivest, Adi Shamir and Leonard Adleman. RSA algorithm uses the asymmetric key for providing the secure communication. It consists of public and private key for the encryption and decryption process. Stimulation speed in case of RSA is faster, it uses equation editor to write equations. The security of RSA depends upon the product of two prime numbers. It requires keys of at least 1024 bits for secure communication. Block size should be minimum of 512 bits. Ciphering and deciphering key used are different from DES and AES [1, 2].

##### **3. Digital Signature Algorithm (DSA):**

A digital signature algorithm is a public key cryptographic algorithm designed for authenticating digital message. A message is signed by a secret key to produce a signature, and then verified against the message by a public key. Any party can verify the signatures but only one party with the secret key can sign the messages [16]. A valid digital signature gives recipient a reason to believe that the message was created by a known sender.

From the above section few observations have been drawn and are as: a) asymmetric algorithms are much secured against the random attacks generated by the hackers, b) large and variable key size makes the system more secured, c) security of communication model also increases, if more iterations are used, and d) public key encryption is used to solve the problem of key distribution.

#### **C. Key management and Performance measuring parameters:**

**Xiaojiang Du** [4] *et al.* worked on routing driven elliptic curve cryptography based key management scheme in heterogeneous sensor networks. They proposed a heterogeneous sensor network (HSN) model for the better performance and security. Simulation results proved better security, performance, stability and energy

consumption than other management schemes was achieved. The limitation is that the scheme only used to communicate for the small regions.

**Thomas Monoth** [5] *et al.* proposed tamperproof transmission of fingerprints using visual cryptography schemes. They worked on two aspects for the data security, a) Visual cryptography, and b) Biometrics. Visual cryptography scheme (VCS) allows confidential messages to be encrypted into selective secret sharing schemes. The image reconstruction through VCS is still limited; therefore, more research is to be done in depth for the better results. The main drawback of the VCS is the loss in contrast of reconstructed images.

**Lin Cheng** [6] *et al.* discussed Cryptanalysis and improvement of a certificateless encryption scheme in the standard model. They proposed an improved scheme which is really secure against “malicious-but-passive” KGC (Key Generation Centre) attack in the standard model. Their scheme is proved to be insecure even in a weaker security model called “honest-but-curious” KGC at-tack model.

**Ahmet Dogan** [7] *et al.* worked on analyzing and comparing the AES architectures for their power Consumption. They introduced low power AES architectures which have gained importance over an entire area and performance oriented designs. It leads to reduced power consumption in FPGA (Field Programmable Gate Arrays). They also proposed most popular four architectures including; serial, outer pipeline, inner outer pipeline and single s-box only with all the different s-box realizations.

**Jongkil Kim** [8] *et al.* proposed adaptively secure identity based broadcast encryption with a constant sized cipher text. They introduced a system which is fully collusion-resistant and has stateless receivers. They worked on Identity Based Broadcast Encryption (IBBE) scheme which has unique identity and could be used by an authorizer.

**Peng Xu** [9] *et al.* worked on Public-Key encryption with Fuzzy Keyword Search a Provably Secure Scheme under Keyword Guessing Attack. They investigated the insecurity of PEKS under keyword guessing attack. They also proposed the new primitive of PEFKS to resist KGA and formalized the SS-CKA and the IK-NCK-KGA securities of PEFKS, followed with a universal transformation from anonymous IBE to PEFKS. They propose further idea is to reduce the search time in both PEKS and PEFKS schemes.

From the above section following observations have been drawn: a) generation of key from available data which avoid the need of transmitting key over secured channel, b) for secured data dynamic keys are preferred, c) need optimized key management in order to achieve secure cryptographic model, and d) public key encryption is used to solve the problem of key distribution.

From the observation drawn from section 2.1 and 2.2, it has been observed that still there is need to work upon the key size in order to make the model more optimized. The main focus is to improve the encryption and decryption time for the processing of secured data. Small size keys are not capable; therefore, long length keys are used to provide secure cryptographic model.

### III. Comparative Study Analysis on performance of various encryption algorithms

In this section we compared the results of energy consumption by various encryption algorithms performance which was obtained from other resources. There were various symmetric key encryptions consumes energy with different file size. We examined that when data was transmitted through these encryption algorithms then there was insignificant differences found on performance of these symmetric keys.

Table I  
 Comparative Execution Times of various Encryption Algorithms with different packet size

Sr. No.	Input File Size (In kb's)	Encryption Execution Time(msec)				
		AES	3DES	DES	RC6	Blow Fish
1	49	38	48	33	24	36
2	59	56	54	29	41	36
3	100	90	81	49	60	37
4	247	112	111	47	77	45
5	321	164	167	82	109	45
6	694	210	226	144	123	96
7	899	258	299	248	162	64
8	5345.28	1237	1466	1296	695	122
9	7310.336	1366	1786	1695	756	107
10	Time Average	392.33	470.88	402.55	227.44	65.33
11	Throughput (In mb's)	4.174	3.45	4.01	7.19	25.892

Formula's Used :-

$$\begin{aligned} \text{➤ Encryption Time Average} &= \frac{\sum \text{Encryption Time Execution}}{\text{Total no. of Inputs}} \\ \text{➤ Encryption Throughput} &= \frac{\sum \text{Input File Size}}{\sum \text{Encryption Time Execution}} \end{aligned}$$

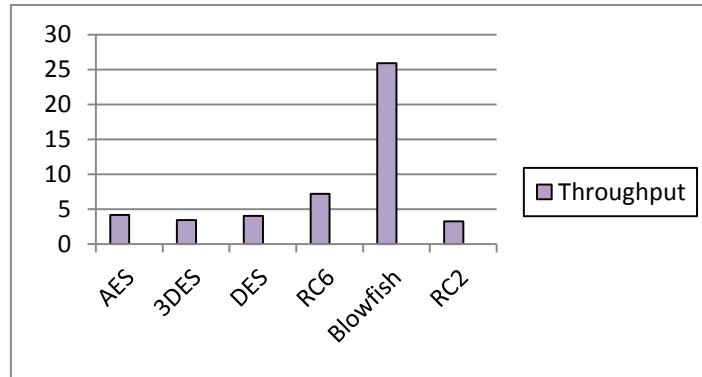


Fig 3.2 : Throughput of each Encryption Algorithm

It was observed that energy consumption [10] shows AES is faster and gives better efficiency form other encryption algorithms. These results shown on two different machines which shows that blowfish had a very good performance compared by other encryption algorithms. It also showed that AES better performance than 3DES, DES and RC2. These results was measured on four different web browsers which shows that with the change in the key size there must changes in battery power and time consumption [14].

#### IV. Conclusion and Future Scope

The study of various encryption algorithms for secured communication has been done. From the recent work, it has been observed that the variations along with the input file size shows that to encrypt the file, average time of blowfish is very much good as compared to other RC6 and AES. But on studying I observed that if we use unstructured data on transmission rather than text file than blowfish algorithm increases its encrypting time. So with the change in circumstances there are up's and down's obtained while encrypting input files using above algorithms. After analyzed in table, the energy consumed by blowfish algorithm was too more as compared by other encryption algorithms.

#### V. References

- [1] Ajay Kakkar, Dr. M. L. Singh, Dr. P. K. Bansal, "Efficient Key Mechanisms in Multinode Network for Secured Data Transmission", International Journal of Engineering Science and Technology, vol. 2, Issue 5, 2010, pp.787-795.
- [2] William Staling, "Network Security Essentials: Applications & Standards", Fourth Edition.
- [3] Bruce Schneier, "Applied Cryptography", John Wiley & Sons, Second Edition, January 1996.
- [4] Xiaojiang Du, "A Routing-Driven Elliptic Curve Cryptography Based Key Management Scheme for Heterogeneous Sensor Networks", IEEE Transactions on wireless communications, vol. 8, no. 3 .pp. 1223-1229 , March 2009.
- [5] Thomas Monoth , Babu Anto P, "Tamperproof Transmission of Fingerprints Using Visual Cryptography Schemes", Science Direct Computer Science, pp.143-148, 2010.
- [6] Lin CHENG, Qiaoyan WEN, Zhengping JIN, Hua ZHANG, "Cryptanalysis and improvement of a certificateless encryption scheme in the standard model", Springer Computer Science, pp.163–173, 2014.
- [7] Ahmet Dogan, S.Berna Ors, Gokay Saldamli, "Analyzing and comparing the AES architectures for their power consumption", Springer, pp.263–271, 2014.
- [8] Jongkil Kim, Willy Susil, Man Ho Au, Jennifer Seberry, "Adaptively Secure Identity-Based Broadcast Encryption with a Constant-sized Ciphertext", IEEE Transactions on Information Forensics and Security, pp. 1-15, 2014.
- [9] Peng Xu, Hai Jin , "Public-Key Encryption with Fuzzy Keyword Search: A Provably Secure Scheme under Keyword Guessing Attack", IEEE Transactions on computers, vol. 62, no. 11, pp. 2266-2278 , Nov 2013.
- [10] Hardjono, Security In Wireless LANS And MANS, Artech House Publishers 2005.
- [11] Chu-Hsing Lin, "Dynamic Key Management Schemes for Access Control in a Hierarchy", ELSEVIER Computer Communications, vol. 20, pp.1381-1385, 1997.
- [12] Sheng Zhong, "A Practical Key Management Scheme for Access Control in a User Hierarchy, Computers & Security, Elsevier Science Ltd., vol. 21, No 8, pp. 750-759, 2002.
- [13] Diffie, W., and Hellman, M., "New Directions in Cryptography", IEEE Transaction Information Theory IT-22, pp. 644-654, Nov 1976.
- [14] Ritika Chehal, kuldeep Singh, " Efficiency and Security of Data with Symmetric Encryption algorithms", IJARCSSE, vol. 2, Issue 8, pp.472-475, August-2012.
- [15] T. Korner, "The Pleasures of Counting", Cambridge University Press, England, 1996.
- [16] R. Rivest, A.Shamir and L. Adleman, "A Method for obtaining Digital Signature and Public key Cryptosystem", Communication of ACM, vol. 21, no.2, pp.120-126, 1978.

# A Survey on Sentimental Analysis using Opinion mining

Rekha<sup>1</sup>, Dr. Williamjeet Singh<sup>2</sup>  
Department of Computer Engineering,  
Punjabi University, Patiala, Punjab, India  
Email-id: [garg15rekha@gmail.com](mailto:garg15rekha@gmail.com)<sup>1</sup>, [williamjeet@gmail.com](mailto:wiliamjeet@gmail.com)<sup>2</sup>

**Abstract-**Opinion mining additionally known as sentiment analysis may be a method of finding users opinion regarding specific topic or a product or drawback. A subject may be a product, movie, news, event, location building etc. Opinion mining may be a field in data processing, natural language process (NLP), and net mining discipline. An outsized volume of data in on-line systems is hold on within the any kind format. This data takes a structured form that can be transmitted on the net, being the foremost common illustration kind and simple to understand by the individuals. In this paper, we have reviewed the mining process for getting customer's review regarding a particular mobile phone. Online Reviews from totally different sites that permit the net users to create their call concerning the merchandise they require buying can be collected from different selling sites and a comparison can be done in the marketing trends of a particular mobile. These reviews can be positive, negative and neutral. It's become quite tough to choose a particular phone as there are numerous available in the market since we have a tendency to unable to choose quickly. So from customer reviews we can compare them and can buy the best match from the information which can be provided by an algorithm on the collected data. Therefore it's obligatory to classify the reviews from structured information sets for analysis and opinion mining of any applications. In future work, we will propose an efficient algorithm which can easily provide the necessary information from collected data. A significant part of our information-gathering is to search out what others suppose. With the growing availability of user's reviews on totally different resources like on-line review sites and private blogs, new opportunities and demands seems as individuals currently will, and do, actively use data technologies to look out and perceive the opinions of others.

**Keywords**— Opinion Mining, Sentiment Analysis

## I. INTRODUCTION

E-commerce has been a primary shift in today's world. A customer who wish to purchase a product or if he/she is interested to know about the product, reviews provided by the web analysts. Reviews are very critical to make decisions which would likely to have a great impact among customers as well as the marketers. With more & more user have become comfortable with web, reviews that a product receives grow rapidly as increasing number of people are writings. Sometimes many reviews are very long & contain only a few sentences that contain review of product. Some popular products can get thousands of reviews at large merchant sites. It makes difficult for a customer to read them for making suitable decision as to purchase product or not. Large number of reviews makes it difficult for companies to keep detail of customer's opinions for their products.

## II. OPINION MINING

It can be defined as a method of computational fundamental that focuses on extracting people's opinion from the web. Recent enlargement of net encourages users to contribute & specific themselves via videos, social

networking sites, etc. of these platforms offer an oversized quantity of useable info that we have a tendency to have an interest to investigate. Given a bit of knowledge contained within the text, opinion-mining systems analyze:

- Which part is opinion expressing?
- Who wrote opinion?
- What is being commented?

Sentiment analysis, on other hand, is about determining subjectivity, polarity (positive or negative) & polarity strength (weakly positive, mildly positive, strongly positive, etc.) of a piece of text.

- What is opinion of writer?

It refers to use of text analysis, language process & machine basic to gather relevant subjective info from source. Sentiment analysis is applied to reviews & social media opinions for a range of applications, from promoting to client service. Perspective could also be his judgment or emotional state, evaluation, or supposed emotional communication. Sentiment analysis targets to see perspective of speaker or author with relevancy some topic or overall polarity of a document.

### III. LITERATURE SURVEY

Shanmuga sundaram, Hariharan, Joan Lu. [1] planned a Classifying product reviews from balanced database for Opinion Mining & Sentiment Analysis They need taken the balance information set to investigate the reviews regarding any explicit application. They conferred a study on classifying the Document victimisation similarity metric from balanced review sets.

Long-Sheng subgenus Chen and Hui-Ju Chiu [2] planned developing a neural network based mostly index for sentiment classification. They used NN based mostly index surpass ancient approaches, as well as Back-Propagation neural network and several other orientation indexes. Disadvantages of NN square are that it measures the dimensional size of date. Which will result into terribly long coaching method to machine learning technique?

F. Zhou, R. J. Chinese monetary unit [3] planned latent client desires input for large - data analysis of their on-line product comments. They need to plan a brand new paradigm of client desires input that supports sentiment analysis for individual products attributes of on-line products. Support vector machine is employed to make the model.

Ashraf Ullah, Khairullah Khan, Aurnagzeb Khan, BaharumBaharudin[4] planned a mining opinion components from a unstructured comments : A review. This paper presents a background study of opinion mining. This study indicates the significance on the opinion mining that is employed from previous couple of years.

DinkarSitaram, Savitha Murthy, Devraj ray, Devansh Sharma, KashyapDhar[5] worked on sentiment analysis of mixed language using Hindi – English code switch. During this paper, they need done the grammatical transition of the reviews of the client. This analysis is just valid on brief and small sentenced opinions which might be obtained from social networking sites.

YuyaSawakoshi, Makoto Okada, KiyotaHashimoto[17] they planned AN Investigation of Effectiveness of “Opinion” and “Fact” sentences for Sentiment Analysis of Customers reviews. They’d used the SVM technique to classify the sentences into positive and negative kind.

Wei Yen Chong, BhawaniSelvaretnam, Lay-Ki Soon [6] planned natural language processing for Sentiment Analysis. They used language process technique to find the sentiment of the sentence. They gift the preliminary

results of their planned system that company natural language processing technique to extract subject from tweets, and differentiate the polarity of the tweets by analyse sentiment lexiconsthat square measure associated to the subject.

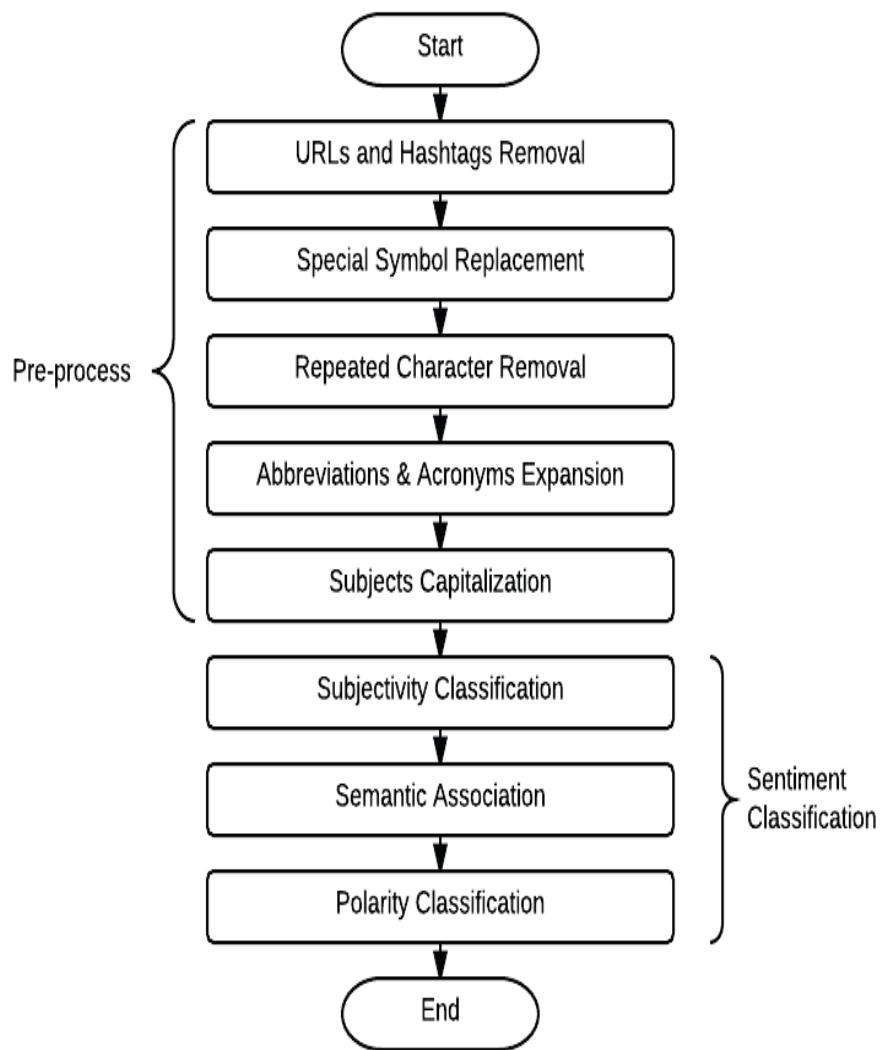


Fig1: Flow Chart of Pre-process and Sentiment Classification

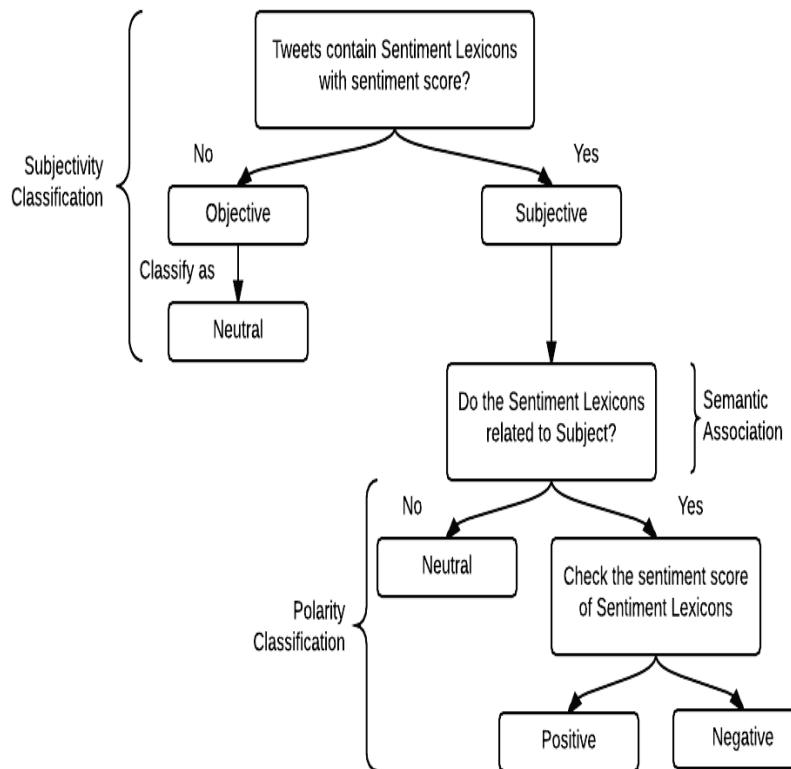


Fig 2: Sentiment Classification Process.

R.Nithish, S.Sabarish[7] An Ontology based Sentiment Analysis for mobile products using tweets. They projected a model as metaphysics to store the merchandise info. This metaphysics brings some relationship within the hold on knowledge. Problem with this method is that not enough knowledge is accessible as tweets for specific option of some mobile.

Lu Lin, Jianxin Li, Richong Zhang, Weiren Yu and ChenggenSun[8] Opinion Mining and Sentiment Analysis in Social Networks: A Retweeting Structure-aware Approach. They planned associate degree opinion descriptive technique associate degreed an opinion mining technique and build a true time analysis system to watch the sentiment analysis.

Elena Lloret, Alexandra Balahur[9], Towards a unified framework of mining , summary , opinion retrieval. They planned a unified framework composed of 3 main elements that enable the retrieval, classification and account of subjective data. They ended that their approach is enough to search out the narrow-minded sentences.

GauravDubey, Ajay Rana, Naveen Kumar Shukla[10], proposed user reviews data analysis using opinion mining on web. They worked on the purchasers reviews within the mobile domain. They collect the opinion words from the reviews and represent those reviews within the graphical type.

Taysir Hassan A. Soliman, Mostafa A. Elmasry[11] proposed utilizing support vector Machines in mining online customer Reviews. They apply associate degree opinion mining approach to compile the unstructured reviews from the net victimization SVM technique.

LilianaFerreira[12] proposed a comparative study of feature extraction algorithms in customer reviews. They compare the 2 algorithms i.e POS pattern and association rule mining. The result they terminated that association mining is healthier than alternative.

Chinsha T C, Shibly Joseph [13] planned a syntactic approach for aspect based opinion mining. They specialise in facet level opinion mining and propose a replacement approach for it that is predicated on syntactic.

Khairullah Khan, BaharumB.Baharudin, Aurangzeb Khan, Fazal-e-Malik[14] planned a mining opinion from text documents: a survey. They'd done a literature survey on completely different machine learning techniques. They conclude that information science techniques would like some improvement within the future work.

Li-Chen Cheng, Zhi-Han Ke, Bang-Min Shiue[15] planned find work changes in opinion from customer reviews. They worked on associative classification methodology to search out the relation between the options. This can be to see the changes within the opinions of users.

David Garcia, Frank physician [16] planned emotions in product reviews – Empirics and models they known completely different time dynamics of the creation of reviews smitten by the presence of selling and word of mouth effects.

MostafaKaramibekr, Ali A. Ghorbani[18], Sentiment Analysis of Social issues. They worked on sentiment analysis of social problems. They conduct an applied math analysis on the distinction between sentiments of the merchandise and social problems.

#### IV. AREA OF WORK DONE

##### A. HOTEL AND HOTEL FEATURES

Farman Ali, Kyung-Sup Kwak, Yong-Gi Kim[19], they proposed the fuzzy domain ontology and Support vector machine for the work on reviews of hotel and hotel features. Their proposed system removes the irrelevant reviews and computes the individual feature polarity. The proposed framework effectively arranges the seriously obscured audits and intelligently computes the individual elements extremity and hotel polarity.

##### B. MOVIES

V.K. Singh, R. Piryani, A. Uddin[20] they proposes a new feature-based heuristic scheme for aspect-level sentiment classification of a movie. They have taken the review of “Guru” movieThe resultant opinion profile is educational, simple to comprehend, and to a great degree valuable for clients. Also, the algorithmic definition utilized for viewpoint level feeling profile rushes, to execute, quick in delivering results and does not require any past preparing.

Abd. Samad Hasan Basari, Burairah Hussin, I. Gede Pramudya Ananta, Junta Zeniarja[21] they used the messages of twitters to review a movie for sentiment and opinion analysis. They have used the Support Vector Machine algorithm to analyze the data and recognized the pattern of data. They improves the accuracy level up to 77%. They have taken the reviews on movies: “Transformers”, “Star Trek”, “The Hangover” and “Angel and Demons”.

##### C. MOBILES

Mrs Vrushali yogesh karkare, Dr. Sunil R Gupta[23] Worked on the reviews of the iphone. They proposed an approach to get the people feedback about the product which is posted on net and create the corpus of the reviews.

The processing of corpus is done and part of speech tagging is performed. And final step to extract the features and rate those features based on their weights and compare the features and presents the data.

Jan Prichystal[22] have used the mobile product for getting the customers reviews to evaluate the hidden feelings of another customer. They used the method to get the bar code of the product and get the product identifier and obtain reviews of the product. The goal of the applied application is to provide the customer opinion to another customer.

Hegder et al. [25] proposed a contextual analysis of Kannada SA for portable item audits as there are numerous client created Kannada item surveys accessible on the web. In this approach a vocabulary based technique for angle extraction has been produced. Moreover, the Naive Bayes characterization model is connected to dissect the extremity of the feeling because of its computational effortlessness and stochastic vigor. This is the main endeavor in Kannada to the best of creator's information. In this manner, a tweaked corpus has been produced. The week by week surveys from the segment 'Device Loka' by U.B Pavanaja are considered to build up this corpus. The hotspot for this is the well known Kannada news paper 'Prajavani'. The preparatory results show this methodology is a proficient procedure for Kannada SA.

## V. DISCUSSION

Below we have given a brief review of some journals which have done similar work like our chosen topic of opinion mining. In this we have briefed the methods, products that are reviewed along with the results given by the proposed algorithms.

TABLE 1  
 PRODUCTS AND METHOD USED FOR CUSTOMER REVIEWS

Title	Review Products	Method	Result/conclusion
A comparative study for feature extraction algorithms	Two digital cameras, a DVD player, MP3 player, a cell phone	Feature extractor algorithm and association mining approach	By comparing these two techniques, an association rule mining is better than other
A syntactic approach for aspect based opinion mining	Restaurant	Syntactic based approach	This approach gives 78.04% accuracy as compared to part-of-speech
Detecting changes of opinion from customer reviews	Ipad, 1st period is from April 2010 to Sept, 2010 2 <sup>nd</sup> period is from Oct 2010 to April 2011.	Opinion change mining approach is proposed and associative classification method is used only for analyzing	Comparison of opinion analysis rule for different period so that customer can get the exact
Developing a neural network based index for sentiment classification	movies	Neural Network (combine machine learning algorithm and info retrieval)	Accuracy is 70.2 %
Natural Language Processing for Sentiment Analysis	tweets	NLP containing 3 steps subjectivity classification, semantic association, and polarity classification.	Accuracy is 59.85% Still need some improvement as SVM having accuracy 64.95

## VI. CONCLUSION

Opinion mining is associate degree emanate field of information mining to excerpt the knowledge from immense volume of knowledge which will be client reviews, and feedback on any product or topic etc. analysis has been organized to order opinions in 3 alternative forms: order opinions document, sentence and have level sentiment analysis it's examined that currently opinion mining progress is moving to the sentimental reviews of social websites that contain the reviews of various folks in line with their interest associated with the merchandise. In this paper we have briefed the basics of opinion mining, its need and the applications which can use this for making a better decision in selecting a particular item or product when there are similar numerous products with almost similar usage available in the market. We have chosen the smart phones for get the information from a collected data containing reviews of customers about different smart phones and future work will to get the performance of individual phones in a structured form. Since OM is associate degree rising and growing field of interest thus during this paper we've primarily centered on the present analysis work to explore the sector so as to search out a transparent direction for future work. In future work we will compare the three mobile sets of same company and find which model is best on the basis of customer reviews about the different features of mobile phones.

## REFERENCES

- [1] Periakaruppan Sudhakaran1, Joan Lu3. Classifying product reviews from balanced datasets of Sentiment Analysis & Opinion Mining 2014 6th International Conference on Multimedia, Computer Graphics and Broadcasting IEEE.
- [2] Long-Sheng Chen and Hui-Ju Chiu ,Developing a Neural Network based Index for Sentiment Classification. International MultiConference of Engineers and Computer Scientists 2009 Vol I
- [3] F. Zhou, R. J. Jiao, Latent Customer Needs Elicitation for Big-Data Analysis of Online Product Reviews. ©2015 IEEE.
- [4] Khairullah Khan , Baharum Baharudin , Aurnagzeb Khan , Ashraf Ullah, 2014. Mining opinion components from unstructured reviews: A review. Computer and Information Sciences Department, University Teknologi.
- [5] Dinkar Sitaram, Savitha Murthy, Devraj Ray, Devansh Sharma, Kashyap Dhar. Sentiment analysis of mixed language employing Hindi – English code switching. 2015 international conference on machine learning and cybernetics.IEEE.
- [6] Wei Yen Chong, Bhawani Selvaretnam, Lay-Ki Soon, Natural Language Processing for Sentiment Analysis. 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology. IEEE
- [7] R. Nithish, S. Sabarish, M. Navaneeth Kishen. An Ontology based Sentiment Analysis for mobile products using tweets ©2013 IEEE Fifth International Conference on Advanced Computing (ICoAC).
- [8] Lu Lin, Jianxin Li, Richong Zhang, Weiren Yu and Chenggen Sun 2014. Opinion Mining and Sentiment Analysis in Social Networks: A Retweeting Structure-aware Approach. 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing.
- [9] Elena Lloret, Alexandra Balahur , José M. Gómez, Andrés Montoyo, Manuel Palomar 2012. Towards a unified framework for opinion retrieval, mining, and summarization, Springer Science+Business Media, LLC 2012.
- [10] Gaurav Dubey, Ajay Rana, Naveen Kumar Shukla, User reviews data analysis using opinion mining on web, 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management, IEEE.
- [11] Taysir Hassan A. Soliman, Mostafa A. Elmasry. Utilizing support vector Machines in mining online customer Reviews. 2012 IEEE.
- [12] Liliana Ferreira, A Comparative Study of Feature Extraction Algorithms in Customer Reviews, The IEEE International Conference on Semantic Computing.
- [13] Chinsha T C, Shibly Joseph. A Syntactic Approach for Aspect Based Opinion Mining. 9<sup>th</sup> international conference on semantic computing (IEEE ICSC 2015).
- [14] Khairullah Khan, Baharum B. Baharudin, Aurangzeb Khan, Fazal-e-Malik. Mining Opinion from Text Documents: A Survey 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies.

- [15] Li-Chen Cheng, Zhi-Han Ke, Bang-Min Shiue. Finding changes in opinion from customer reviews. 2011 IEEE, Eighth International Conference on Fuzzy Systems and Knowledge Discovery.
- [16] David Garcia, Frank Schweitzer. Emotions in product reviews – Empirics and models. 2011 IEEE International Conference on Privacy, Security, Risk, and Trust.
- [17] Yuya Sawakoshi, Makoto Okada, Kiyota Hashimoto. An Investigation of Effectiveness of “Opinion” and “Fact” sentences for Sentiment Analysis of Customer reviews. © 2015 IEEE.
- [18] Mostafa Karamibekr, Ali A. Ghorbani. Sentiment Analysis of Social Issues. 2012 IEEE International Conference on Social Informatics.
- [19] Farman Ali, Kyung-Sup Kwak, Yong-Gi Kim, Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification. Elsevier applied Soft Computing 47 (2016) 235–250.
- [20] V.K. Singh, R. Piryani, A. Uddin. Sentiment Analysis of Movie Reviews ©2013 IEEE.
- [21] Abd. Samad Hasan Basari, Burairah Hussin, I. Gede Pramudya Ananta, Junta Zeniarja. Opinion mining of movie review using hybrid method of Support vector machine and Particle Swarm Optimization. Elsevier, MUCET 2012 Part 4 - Information and Communication Technology.
- [22] Jan prichystal. Mobile application for customer reviews opinion mining. Elsevier on 19<sup>th</sup> international conference enterprise and competitive environment 2016.
- [23] Mrs. Vrushali Yogesh Karkare, Dr. Sunil R. Gupta Product Evaluation using Mining and Rating Opinions of Product Features © 2014 IEEE.
- [24] Gaurav Dubey, Ajay Rana, Naveen Kumar Shukla, User reviews data analysis using opinion mining on web, 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management, IEEE.
- [25] Yashaswini Hegde, S.K. Padma, " Sentiment Analysis for Kannada using Mobile Product Reviews A Case Study" Published in Advance Computing Conference (IACC), 2015 IEEE International Date of Conference: 12-13 June 2015.

## IMPROVED CSNM FOR CONECTION CONTROL AND LOAD BALANCING IN WSN

Harpritpal Singh<sup>1</sup>, Gaganpreet Kaur<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering, SGGSWU, Fatehgarh sahib

<sup>2</sup>Department of Computer Science and Engineering, SGGSWU, Fatehgarh sahib

([bkhaira99@gmail.com](mailto:bkhaira99@gmail.com) , [Gaganpreet\\_cse@sggswu.org](mailto:Gaganpreet_cse@sggswu.org))

**Abstract-**WSN is an emerging area for sensing information from non-approachable fields. In this fields sensor nodes have been deployed for sensing information. Sink node collects information from sensor nodes based on routing strategy.in this paper improved congestion avoidance in WSN approach has been purposed for regular delivery of data from sensing nodes. Due to congestion much amount of meaningful information gets loss and create burden on network for retransmitting. In this paper dynamic cluster based approach has been purposed that works with CSNM for congestion management. Numerical value of various parameters represents best results than previous approaches.

**Keywords:**wireless scenario, cluster head, congestion control algorithm, cycle algorithm.

### I. INTRODUCTION

Wireless sensor networks (WSNs) are generally composed of one or more sinks and tens or thousands of sensor nodes scattered in a physical space. With integration of information sensing, computation, and wireless communication, the sensor nodes can sense physical information, process crude information, and report required information to the sink. These sensors are small, with limited processing and computing resources [1]. These sensor nodes can sense, measure, and gather information from the environment and, based on some local decision process, they can transmit the sensed data to the user. The common task of sensor node is to collect the information from the scene of event and send the data to a sink node. Figure 1 shows the typical wireless sensor network that consist of multiple number of sensor nodes and one sink where data is collected are deployed in the sensing field. WSNs can be used in many applications such as habitat monitoring, security surveillance, target tracking, medical application and etc. Wireless sensor network (WSN) is a high degree of cross-disciplinary, highly integrated knowledge on network communication, and is a forefront research hot spot in the world [2].

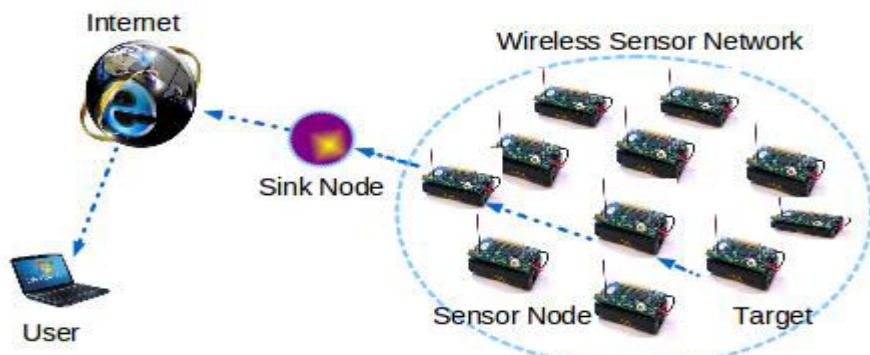


Fig 1 WSN Deployment and architecture

#### 1.1 Leaky Bucket Algorithm

- It is a traffic shaping mechanism that controls the amount and the rate of the traffic sent to the network.
- A leaky bucket algorithm shapes bursty traffic into fixed rate traffic by averaging the data rate.

- Imagine a bucket with a small hole at the bottom.
- The rate at which the water is poured into the bucket is not fixed and can vary but it leaks from the bucket at a constant rate. Thus (as long as water is present in bucket), the rate at which the water leaks does not depend on the rate at which the water is input to the bucket.
- Also, when the bucket is full, any additional water that enters into the bucket spills over the sides and is lost.
- The same concept can be applied to packets in the network. Consider that data is coming from the source at variable speeds. Suppose that a source sends data at 12 Mbps for 4 seconds. Then there is no data for 3 seconds. The source again transmits data at a rate of 10 Mbps for 2 seconds. Thus, in a time span of 9 seconds, 68 Mb data has been transmitted.

### 1.2 Token bucket Algorithm

- The leaky bucket algorithm allows only an average (constant) rate of data flow. Its major problem is that it cannot deal with busty data.
- A leaky bucket algorithm does not consider the idle time of the host. For example, if the host was idle for 10 seconds and now it is willing to send data at a very high speed for another 10 seconds, the total data transmission will be divided into 20 seconds and average data rate will be maintained. The host is having no advantage of sitting idle for 10 seconds.
- To overcome this problem, a token bucket algorithm is used. A token bucket algorithm allows busty data transfers.
- A token bucket algorithm is a modification of leaky bucket in which leaky bucket contains tokens.
- In this algorithm, a token(s) are generated at every clock tick. For a packet to be transmitted, system must remove token(s) from the bucket.
- Thus, a token bucket algorithm allows idle hosts to accumulate credit for the future in form of tokens.

## 2 REVIEW OF LITERATURE

Mukherjee, A. "A low power low noise VCO and a high gain LNA for WSN in 130nm CMOS RF technology" This paper mainly focuses on the performance analysis of two core building blocks of aWSNtransceiver system. A typical WSN transceiver consists of many blocks of which LNA and VCO are of major importance. These two blocks are design and simulated to operate at 2.4GHz ISM band especially for WSN using 130nm CMOS RF technology. Both the LNA and VCO are designed to achieve low power and high efficiency along with less number of components as to take less area as well. The cascaded common source topology is used with inductive source degeneration to design the LNA. The minimum noise figure (NF<sub>min</sub>) of the proposed LNA is nearly 0.472 dB and gain i.e. S<sub>21</sub> is 15.390 dB, which is excellent for WSN application. The designed LNA is unconditionally stable over the frequency range of 1 GHz to 5 GHz. The VCO is designed with current reuse -Gm cross coupled topology to achieve phase noise of as low as -130dB at 1MHz offset frequency with an extremely low power consumption of about 0.023mw. The tuning voltage (V-tune) is set accordingly for 2.4 GHz WSN application. The proposed VCO shows excellent optimization performance for low power and low phase noise along with a FOM of about -214.69 dBc/Hz which proves to be very good for a compact WSN transceiver system.

Sharawi, M. "WSN's energy-aware coverage preserving optimization model based on multi-objective bat algorithm" This research expands the scope of wireless sensor network (WSN) optimization from single objective to multi objective optimization. It introduces a WSN's energy-aware and coverage preserve hierachal clustering and routing model based on multi-objective bat swarm optimization algorithm. Two objectives are taken into consideration; coverage and nodes residual energies. The proposed model optimizes the WSN by selecting the best fitting set of nodes as cluster heads. It works to maximize the WSN's coverage and to minimize the nodes' consumed energy. This minimizes the number of active cluster heads while preserving a higher percentage of the covered nodes in WSN. It extends the longevity of the WSN's lifetime and achieves good functioning reliability. The proposed optimization model overcomes the WSN's coverage and lifetime challenges. The proposed model outperforms the LEACH routing and clustering protocol.

Mathur, A. "Healthcare WSN: Cluster Elections and Selective Forwarding Defense" Wireless sensor networks are an array of different components. Upgrading a single component may not be enough. It is necessary to take each component and apply the best possible techniques to upgrade the system. This paper looks at the configurations

necessary for a medical based WSN with particular focus on clustering and routing. It includes an implementation of WSN architecture in Contiki operating system using Tmote Sky and open mote technologies. The underlying power consumption related to this architecture is analyzed and real world measurements are presented. Finally, the case of a modified secure routing algorithm is introduced with simulation based results showing single and collaborative selective forwarding detection/correction, and network latency. Additionally, a solution to the problem of malicious nodes dropping control messages has been provided. The uniqueness of the paper resides in the implementation of monitoring mechanism without any watchdog nodes. Moreover, to the best of the authors' knowledge, this mechanism has not been implemented on the Contiki OS.

Akele, G. "Virtual group leader election algorithm in distributed WSN" A distributed system is a collection of nodes interconnected by a communication network in which each node can work together and has its own local memory and other peripherals. The communications between the nodes are held by message passing over the communication network. An important challenge confronted in distributed wireless sensor network (WSN) is the adoption of suitable and efficient algorithms for leader election. The goal of a leader election in distributed WSN of autonomous nodes is to select one of the currently alive nodes as a leader so as to manage the coordination activities of the other nodes in the system. To the best of our knowledge most existing leader election algorithm have limitation either fault tolerance, message passing overhead or if the leader fails an election process could be initiated by communicating all the nodes in the distributed WSN. So, these types of algorithms have a great impact on the performance and the energy efficiency of the WSN because of communication overhead. To mitigate these type leader election limitations, in this paper, we propose a virtual group leader election algorithm by combining a group of nodes (including a master leader and multiple backups) on the WSN into a leader group. Our proposed algorithm improves the energy efficiency and the performance in message passing and less time complexity that results in electing a new leader node faster. The proposed algorithm analyzed and validated through extensive mathematical results. And also the simulation result shows that the proposed algorithm can minimize a lot of energy when the number of nodes increases.

### 3 METHODOLOGY

In the proposed work wireless scenario has been designed by using different simulation parameters that have been used for sensing and transmitting information. Antenna, MAC, LL, QUEUE and routing protocol has been defined for generation of wireless sensor network scenario. After generation of wireless sensor network nodes have been deployed at different locations in the sensing area for sensing the information. These nodes have been provided a battery source for lifetime of the nodes so that data can be sensed and transmitted over the network to the base station.

After deployment of the nodes cluster formation has been done using leach protocol that divides nodes into different clusters on the basis of nodes properties. Nodes have been divided into different clusters that include different cluster members with in clusters. These members are sensor nodes that have been used for sensing information and transmitting of this information to base station. Cluster head selection has been done on the basis of different nodes residual energy that has been used for selection of cluster head and sub cluster head with in a cluster. The concept of sub clustering is used for congestion avoidance because data division has been done so that network overhead on a single node does not affect performance of the network. The node having maximum energy is elected as cluster head and node with energy less than maximum node has been elected as sub cluster head. After every rerun selection of the cluster head and sub cluster head has been done on the basis of residual energy.

After selection of cluster head and sub cluster head sensing information has been started for transmission over the network so that data can be transmit from source node to base station. The nodes start communicating with cluster head and sub cluster head for transmits of data so that data can be easily transmitted over the network without extra consumption of energy. In the process of transmission CSNM congestion avoidance approach has been implemented that work using token bucket approach. This approach is useful for congestion avoidance over the network. Token bucket approach is an optimized congestion control approach that avoids congestion by using bursty information over the network. In the process of token bucket approach different clock duration has been assigned to a single data message transmitted over the network.

In Token Bucket Algorithm it is done in a different way. Here you have got some kind of a counter and per each tick say ( $\delta t$ ) a token is added to the bucket. And whenever packets come if you have got enough number of tokens accumulated then it allows the traffic to be sent at the rate in which it has come. After all the tokens are exhausted it will then introduce the packets at the rate of one token per tick. That means it will become the same as the Leaky Bucket Algorithm. So let us consider the same example. Suppose it has received at the rate of 10 Mbps for two second this is one this is two because of the bursty nature of traffic therefore in case of token bucket what will be done is may be for one second it will send at the rate of 10 Mbps and assuming that five tokens were stored and after that for five more seconds 1 2 3 4 5 that means in six seconds all the packets are transmitted. Therefore as you can

see, initially for this part there were accumulated tokens and because of that the data is transmitted at the rate which it has been introduced into the network and after that it is transmitted. Both these Leaky bucket and Token Bucket Algorithms can be implemented by the operating system or by a network interface chord which is connected to the network and as you can see it is implemented with the help of a counter which initializes to 0 in the beginning and per tick the counter is incremented. (Refer Slide Time: 31:20) On the other hand, whenever each packet is sent the counter is decremented. In other words countdown for each packet sent and count of it performs for each tick and in this way the counter is maintained to implement the Token Bucket Algorithm so it can be implemented either by hardware or by the operating system of the host. We have seen that the Token Bucket Algorithm saves up to a maximum of N tokens and allows a burstiness of N packets in the output stream thereby giving faster response. We have already seen that the time taken for introducing the packets in the network is smaller than the Leaky Bucket Algorithm in Token Bucket Algorithm so it increases throughput compared to the Leaky Bucket Algorithm thereby providing you better throughput.

On the basis of these steps congestion controlled wireless sensor network has been deployed and used for sensing information.

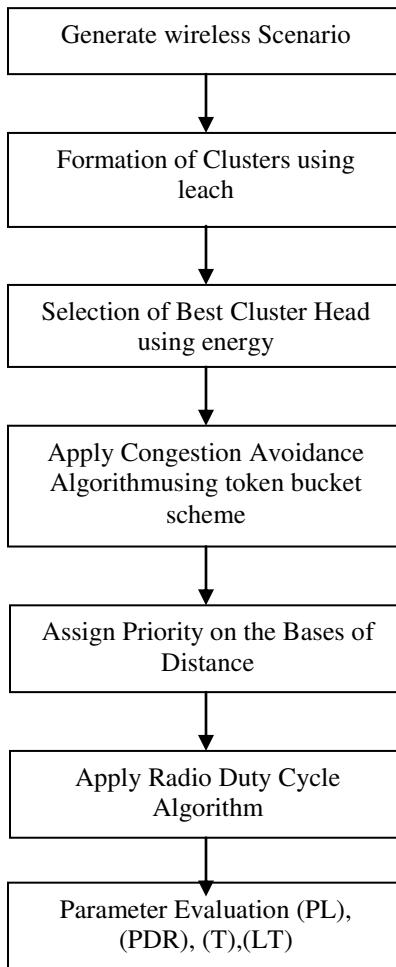


Fig 2 Flow diagram for proposed work

This figure represents flow of the proposed work that must be carried out for achievement of desired objectives. In this flow various steps have been explained that must be followed by the user for development of congestion control wireless sensor network.

#### Energy Calculation of Node

In WSN node have energy to sense information and transmit information from source to destination. The nodes available in network consume sensing, data aggregation, transmission and receiving energy.

$E$  = Energy given to a node,  $E_r$  = Energy Dissipated in receiving data,  $E_t$  = Energy Dissipated in Transmission,  $E_{da}$  = Energy dissipated during data collection,  $E_{res}$  = Remaining Energy.

$$E_{res} = E - (E_r + (E_t + E_{da}) * distance) \quad (1)$$

Distance between two different node having position  $(x_1, y_1)$  and  $(x_2, y_2)$  has been computed using distance formula that works as follow

$$Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (2)$$

On the basis of these energy and distance computation formulas energy of a single node and distance of a node from all the nodes have been computed and cluster head has been selected in the network so that reliable communication can be achieved.

#### 4 RESULTS AND DISCUSSION

In the purposed work WSN nodes have been deployed in the region for sensing information that has been transmitted to base station for decision making process. In this paper various node have been deployed that contain energy for sensing, transmission and receiving data from other sensor nodes. In the purposed work congestion avoided approach that use token bucket algorithm for congestion avoidance. Various parameters have been analyzed for performance evaluation of purposed work.

**Packet Loss:** Packet loss occurs when one or more packets of data travelling across a computer network fail to reach their destination. Packet loss is typically caused by network congestion. Packet loss is measured as a percentage of packets lost with respect to packets sent.

$$P_l = \frac{(T_p - T_d)}{T_p} \quad (3)$$

**Packet Delay:** The sum of store-and-forward delay that a packet experiences in each router gives the transfer or queuing delay of that packet across the network. Packet transfer delay is influenced by the level of network congestion and the number of routers along the way of transmission.

$$D = (T_r - T_s) \quad (4)$$

**Throughput:** It is the number of packets/bytes received by source per unit time. It is an important metric for analyzing network protocols.

$$Th = \frac{T_d * S}{Time} \quad (5)$$

**Packet Delivery Ratio (PDR):** It is the ratio of actual packet delivered to total packets sent. The following table shows the values of the various parameters used during simulation of these protocols.

$$P_d = \frac{T_d}{T_p} \quad (6)$$

Where  $T_d$ , represents total number of packets delivered from source to destination,  $T_p$ , represents total number of packets send from source to destination,  $T_r$ , represents packet receiving time,  $T_s$ , represents packet sending time,  $S$ , represents size of message transmitted.

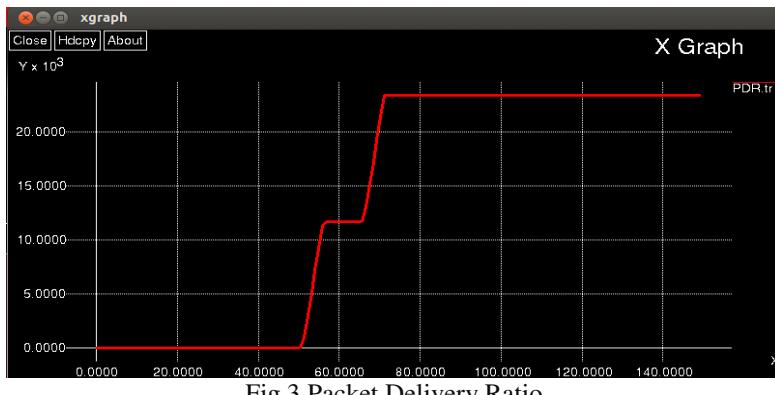


Fig 3 Packet Delivery Ratio

In this X-axis represent the Time and Y-axis represent the Bytes send over the network. This figure is use to represent the Packet Delivery Ratio. Packet Delivery Ratio is defined as the number of packet deliver with respect to time.

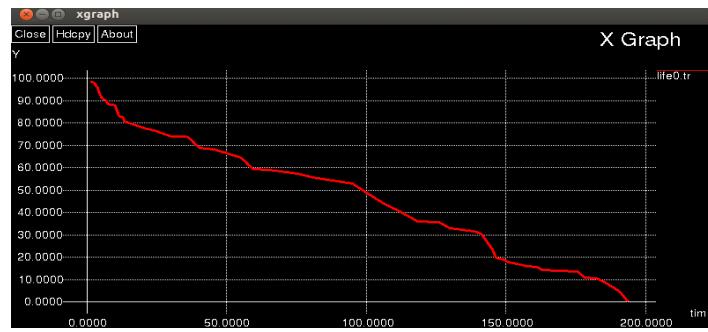


Fig 4 Life time

This figure is use to represent the Lifetime of a node. Lifetime is defined as the total time in which node can survive without any disturbance.

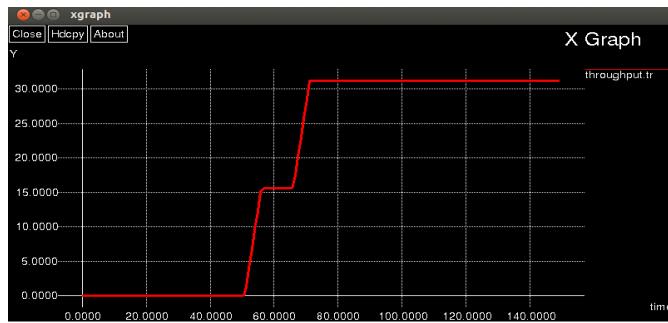


Fig 5 Throughput

This figure is use to represent the Throughput. Throughput is defined as the number of packet delivered successfully over the network.

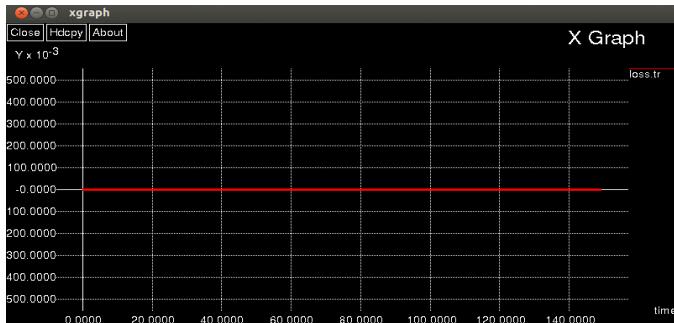


Fig 6 Packet Loss

This figure is use to represent the Loss of packets. Loss is defined as the number of packet loss when we transfer packets over the network

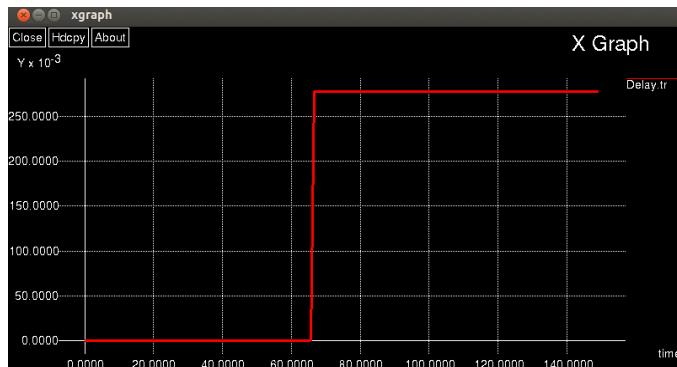


Fig 7 Packet Delay

This figure is use to represent the Packet Delay. Packet Delay is defined as the Delay between packets during transmission.

Table 1 Comparison Table based on performance evaluation parameters

Parameter	proposed	Previous[22]
Throughput	80 %	72 %
Packet Delivery Ratio	79 %	63 %
Packet loss	0	100
Average Delay	0.25 ms	0.5 ms
Energy Consumed	20 J	30 J

## 5 CONCLUSION& FUTURE SCOPE

WSN is the emerging field of communication for transmission of sensing information from the sensing environment using different sensor nodes. Sensor nodes transmit information to the base station so that value able information can be used for decision making process. Due to transmission of the messages over a single node by all the nodes data congestion may be occurred that causes to data loss.

In the proposed work congestion avoidance wireless sensor network has been used that avoid congestion over the network so that data loss to be minimum. To avoid congestion in WSN dynamic cluttering based approach has been utilized that computes cluster head dynamically on the basis of energy available at a particular node. In the proposed work cluster head selection and sub cluster head selection has been done so that data can be divided and avoid the congestion occurred in the network. Token bucket approach avoids multiple identities message by checking token id attached to a single message. This congestion avoidance approach provides much better results than previous approaches. We can easily predicts from the results that proposed approach provides much better results than previous approaches that had been proposed.

### Future Scope

In the future proposed approach can be used for real world applications so that congestion can be controlled for data accusation. In the future reference research can be done to purpose an approach that works to check redundancy of the massages that has been transmitted by the nodes that contain same data. These messages can be automatically detected and discarded by the system so that network overhead will reduces.

## REFRENCES

1. WassimZnaidi,“Hierarchical Node Replication Attacks Detection in Wireless Sensor Networks”, IEEE Conf. on Personal, Indoor and Mobile Radio Communications, Vol. no.10 .Issue no. 32, pp.82 – 86, ISSN:2166-9570, Tokoyo, 2009.
2. JaydeepBarad,“Improvement of deterministic key management scheme for securing cluster-based sensor networks”, IEEE Conf. on Networks & Soft Computing (ICNSC), Vol. no. 10, pp. 55 – 59,ISSN 98714993485-0, 2014.
3. NayyerPanahi,“Adaptation of LEACH Routing Protocol to Cognitive Radio Sensor Networks”, IEEE Conf. on Telecommunications (IST),Vol. no. 45, pp. 541–547,ISSN 978-14673, 2012.
4. Suyogpawar,“Design and evaluation of en-leach routing protocol for wireless sensor network”, IEEE Conf. on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Vol.48, pp. 489–492,ISSN: 978-1-4673-2624,2012.
5. Shin-nosuke, Toyoda, “Dynamic Change Method of Cluster Size in WSN”, IEEE Conf. on Broadband, Wireless Computing,Vol. no. 57, pp. 20–27, ISSN: 978-14799-0, 2014.
6. Muhammad haneef,“Comparative analysis of classical routing protocol leach and its updated variants that improved network life time by addressing shortcomings in wireless sensor network”, IEEE conf. On Mobile Ad-hoc and Sensor Networks (MSN), vol.no. 67, pp. 361–363.ISSN: 978-1-47, 2012.
7. AshrafulAlam, A.S.M,“Helping secure robots in WSN environments by monitoring WSN software updates for intrusions”, IEEE conference on Automation, Robotics and Applications (ICARA),Vol. no. 65, pp. 223–229, ISSN: 876543098, 2013.
8. Rai, Ananad, N.Varma,“Scrutinizing Localized Topology Control in WSN using Rigid Graphs”, IEEE conference on Computing for Sustainable Global Development, Vol. no. 75, pp. 1712 – 1715, ISSN: 978-9-3805-4415-1,2015.
9. Alaiad, LinaZhou, “Patients Behavioral Intentions toward Using WSN Based Smart Home Healthcare Systems: An Empirical Investigation”, IEEE conference on System Sciences (HICSS), Vol. no. 67, pp. 824–833,ISSN 1530-1605,2015.
10. Nagpurkar, A.W,“An overview of WSN and RFID network integration”, IEEE conference on Electronics and Communication Systems (ICECS),VOL no. 60,ISSN 978-1-4799-72, pp. 497–502,2015.
11. Hemalatha, S.Rajamani,“VMIS: An improved security mechanism for WSN applications”, IEEE conference on Science Engineering and Management Research (ICSEMR), Vol. no. 32, ISSN 978-1-4799-7614-0, 2015 pp. 1–3.
12. Mukherjee, Asutkar,“A low power low noise VCO and a high gain LNA for WSN in 130nm CMOS RF technology”, IEEE conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials , Vol. no. 62,ISSN 978-1-4799-9854, pp. 268 –274,2015.
13. Sharawi, Mahdy,“WSN's energy-aware coverage preserving optimization model based on multi-objective bat algorithm”, IEEE conference on Evolutionary Computation (CEC), Vol no. 47, ISSN 1089-778X, pp. 472–479,2015.

14. Mathur, Rao, "Healthcare WSN: Cluster Elections and Selective Forwarding Defense", IEEE conference on Next Generation Mobile Applications, Services and Technologies, Vol. no. 72, ISSN 978-1-47998660-6, pp. 341–346,2015.
15. Akele,Ki-HyungKim,"Virtual group leader election algorithm in distributed WSN", IEEE conference on Ubiquitous and Future Networks (ICUFN),Vol. no. 56, ISSN 2165-8528, pp. 143–148,2014.
16. Horvat,"Power consumption analysis and optimization of ARM based WSN data aggregation node",IEEE conference on Telecommunications and Signal Processing (TSP), Vol. no. 43,ISSN 978-1-4799-72, pp. 1–5,2015.
17. XiueGao ,Keqiu Li, "Congestion Control Algorithm for Data Center", IEEE conference on Services Computing Conference (APSCC), Vol. no. 69,ISSN 978-1-4799-54, pp. 156–161,2014.
18. LiYang, DebinWei,"Congestion control algorithm based on dual model control over satellite network", IEEE conference on Wireless Communications & Signal Processing (WCSP), Vol. no. 54, pp.1 – 6,2015.
19. Kun Wang LeiShu,"An improved congestion control algorithm based on social awareness in Delay Tolerant Networks", IEEE conference on Communications (ICC), Vol. no. 71, ISSN 1550-3607, pp. 1773 – 1777,2014.
20. Watkins, L.Sharmin,"Using network traffic to infer power levels in wireless sensor nodes", Vol. no. 51, pp. 864 – 870,2014.
21. Macaluso, DaSilva,"Fungible Orthogonal Channel Sets for Multi-User Exploitation of Spectrum", IEEE conference on Wireless Communications,Vol. no. 48,ISSN 1536-1276, pp. 2281 – 2293,2015.
22. V. Vaidehi,"Secure Data Aggregation in Wireless Sensor Networks", 2nd International Conference on Computing for Sustainable Global Development,Vol. no. 75,ISSN 1537-8056, pp. 2179 – 2184,2015.

# Image Compression: A Systematic Review and Evaluation

Prabhjot Kaur<sup>a</sup>, Neelofar Sohi<sup>b</sup>

<sup>a</sup>Research Scholar, <sup>b</sup>Assistant Professor

<sup>a</sup>Prabhjot.gne@gmail.com, <sup>b</sup>Sohi\_ce@yahoo.co.in

Department of Computer Engineering, Punjabi University, Patiala, Punjab-147002, India

**Abstract**—In the present scenario, storage, transmission and faster computation are the basic needs of image processing world. Image compression plays an important role in delivering these three features to image processing applications. Huge variety of algorithms and techniques are available for lossless and lossy image compression. This paper presents a systematic literature review of image compression techniques presenting the basic concepts and available methods with their research gaps. An approach is proposed by reinforcing ROI based compression method based on separation of ROI and Non-ROI parts of the image where Huffman Encoding is applied to compress ROI part and Quad Tree Decomposition is applied for Non-ROI part. Performance evaluation of studied image compression techniques is done using parameters like Compression Ratio (CR), Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR) and Bits Per Pixel (BPP). Experimental results of performance evaluation demonstrate that proposed technique outperforms other techniques.

**Keywords**— *Image Compression; Huffman Coding; Arithmetic Coding; Run Length Encoding.*

## 1. INTRODUCTION

Image Compression is the art and science of reducing the amount of data required to represent an image. This is one of the most useful and commercially successful technologies in the field of digital image processing [1]. Huge amount of data are available in today's world. Data may be anything. It may be text, image, audio, video and graphics. These types of data take lots of redundant information that take lots of storage space and transmission time. To solve these problems data compression is very necessary. The compressor is used to reduce the number of bits of original data. Decompressor is used to take back compressed data or reduced data into original form. The basic measures for the performance of image compression are picture quality and the compression ratio. Compression is divided into two categories: Lossless compression and Lossy Compression. Lossless compression is necessary for those applications which require exact recovery of original images. Lossy compression is useful in domain where loss of data is acceptable. Third category which is a combination of lossless and lossy compression is called Hybrid image compression. Hybrid image compression gives best compression ratio and better picture quality. Compression Ratio is usually less in lossless compression but image quality is better in lossless compression as compared to lossy compression. On the other hand, compression Ratio is high in lossy compression but image quality is not good. Image quality improves when the value of PSNR increases. The value of PSNR decreases with increase in CR. Performance parameters used for compression is PSNR (Peak Signal to Noise Ratio), MSE (Mean

Square Error), CR (Compression Ratio), BR (Bit Rate). Compression ratio and image quality is inversely proportional to each other.

#### 1.1 Motivation For Work

- Despite the vast amount of review work exists for compression methods but after assaying the work, lack of systematic literature review and performance evaluation of existing techniques for image compression is realized.
- It will explore the research gaps and statistical knowledge for future researches.

This paper is organized as follows: Section 2 presents the background of Image Compression including basic concepts and basic compression methods. Section 3 reports the review method discussing main research issues, key research areas, research gaps and literature review of few compression techniques. In Section 4, experimental results of performance evaluation of compression techniques are presented and performance metrics used for evaluation are also described. Section 5 discusses the findings based on qualitative and quantitative results and Section 6 presents conclusion of the research and gives recommendations for future research.

## 2. BACKGROUND

Compression can be done on anything like text, images, audio, video, signal and graphics. Storage, Transmission, fast computations are the basic needs of data/image compression. The storage requirements of imaging applications are very high. The basic goal of Image/data Compression is to reduce the memory size by reducing the number of bits, while at the same time maintain the reduced data to reconstruct the image. The reduction of data reduces the memory requirement. The image size is directly proportional to the transmission time of the image. The reduction of data leads to faster and easier transportation of data. Reduced data simplifies the algorithm design and facilitates faster execution of the algorithms.

#### 2.1 Various Compression Methods

Compression is divided into two categories: Lossless compression and Lossy Compression. Lossless compression is necessary for those applications which require exact recovery of original images. Lossy compression is useful in domain where loss of data is acceptable.

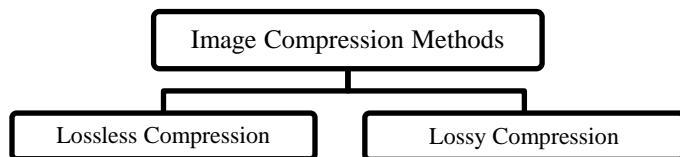


Fig.1. Image Compression Methods

### 2.1.1 *Lossless Compression*

Decorrelation removes inter-pixel redundancy or spatial redundancy [2]. Inter-pixel redundancy appears when pixels are similar to neighbours. It represents the correlation between samples in the form of signals.

Entropy Coding removes coding redundancy. In entropy coding shortest codewords are assigned to frequent symbols and longest codewords to infrequent symbols [2]. Coding redundancy appears due to encoding techniques.

#### (A) *Run Length Coding*

Run length coding is a simplest lossless compression technique. It gives best results for contagious colours and monochrome images [3]. This is more useful when there are more repetitions but increase the file size when repetitions are less [4]. It can be used to compress data made of any combination of symbols.

#### (B) *Huffman Coding*

Huffman coding is entropy coding algorithm used for lossless data compression. The pixels in the image are treated as symbols [3]. The principle of Huffman coding is to use a lower number of bits to encode the frequently occurring data [4]. Huffman algorithms have two ranges static as well as adaptive. Static Huffman algorithm is a technique that encodes the data in two passes. In first pass, it is required to calculate the frequency of each symbol and in second pass it constructs the Huffman tree. Adaptive Huffman algorithm is expanded on Huffman algorithm that constructs the Huffman in one pass, but takes more space than static Huffman algorithm [3].

#### (C) *Arithmetic Coding*

Arithmetic coding is entropy coding technique used in lossless compression. Arithmetic coding needs symbols, probability range and the image sequence for encoding [4]. In arithmetic coding, codewords are constructed by partitioning the range of numbers between zero and one [2]. It uses Binary Fractional numbers [3].

### 2.1.2 *Lossy Compression*

.Lossy compression is useful in domain where loss of data is acceptable.

#### (A) *Chroma Subsampling*

Chroma is the Greek word used for color. It is a lossy technique used for video encoding and also in JPEG encoding. In video, luma represents brightness in an image (black & white or achromatic portion of an image). In other words, Luma represents the achromatic image i.e black and white image while chroma represents the color image or color information. Chroma subsampling is the practice of encoding images by implementing less resolution for chroma information than for luma information.

*(B) Transform Coding*

It is Lossy technique used for natural data like audio signals or photographic images. Output produced by transform techniques are very lower quality than original copy. Variety of methods are used for transform techniques i.e DCT, DFT and discrete Walsh-Hadamard transform (WHT). Discrete Cosine Transform (DCT) has become the most widely used transform coding technique [3].

*(D) Fractal Coding*

It is a Lossy method used for digital images based on fractals. It is a mathematical process which can apply on real world images and describes fractal properties of an image. The fractal compression objects contain redundant information in the form of similar, repeating patterns that is called as fractals. Millions of iterations are required to find the fractal patterns in an image. Fractal decoding is inverse of fractal encoding [3].

*(E) Vector Quantization*

It is a Lossy technique used for signal processing. It is also known as "Block Quantization" or "Pattern Matching Quantization". It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them [3].

*(F) Block Truncation*

It is Lossy technique used for gray scale images. Original image is divided into number of blocks. Reduction of number of gray levels in each block is done by quantizer for maintaining the same mean and standard deviation. It has also been adapted for video coding.

*2.2. Why Compression*

Compression is needed for:

- Reducing storage space.
- Reducing transmission time.
- Faster computation/Increasing speed.
- When high compression ratio is achieved then its quality is affected.
- When image gets high quality then compression ratio is very less.
- PSNR decreases with increase in CR.

### 3. REVIEW METHOD

The pathway taken into account to develop the Review Model for literature modeling is: Conduct the organized review of existing techniques for compression, pinpoint the research gaps of study and key areas of research.

### 3.1. Research Questions

Research questions are the building blocks for the scientists to plan and conduct any research; therefore it is fundamental to frame such questions. Key questions encountered during our study are listed below:

- What are existing techniques, tools, algorithms for compression?
- What are the existing gaps in existing literature?
- Which are the key areas of research in the field of image compression?

Table 1. Reports the answer to the listed question (a),(b).

**Key Research Areas:** Image/ data compression is very useful in many areas like: Geophysics, Telemetry, Non-destructive evaluation, Medical Imaging, Video Coding, Signal Coding and so on.

TABLE1. REVIEW OF IMAGE COMPRESSION TECHNIQUES

S.No	Study	Contribution	Gaps
1.	<b>Paper:</b> An Overview of Lossless Digital Image Compression Techniques  <b>Author:</b> Ming Yang and Nikolaos Bourbakis (2005)	Lossless image compression can be always modeled as two stage procedure.  a. Decorrelation b. Entropy coding  Compression ratio of predictive techniques are better than transform techniques.	<ul style="list-style-type: none"> <li>• Comparing the performance of compression technique is difficult unless identical data sets and performance measures are used.</li> <li>• Some techniques perform well for certain classes of data and poorly for others.</li> <li>• Significant improvements are likely to require much more complex and computationally demanding source models.</li> </ul>
2.	<b>Paper:</b> ROI Based Near Lossless Hybrid Image Compression Technique  <b>Author:</b> Kuldip K. Ade and M. V. Raghunadh (2015)	a. Proposed algorithm is a Combination of lossless and lossy techniques.  b. Huffman coding and snack scanning is used  c. It compresses the image on the basis of priority of regions.  d. Good compression ratio is achieved  e. Less hardware complexity  f. Good picture quality with high compression	—
3.	<b>Paper:</b> A New Image Compression Scheme Using Repeat Reduction and Arithmetic Coding.  <b>Author:</b> Md. Rafiqullslam <i>et al.</i> (2009)	Use of repeat reduction and arithmetic coding  Run-length coding & snack scanning is also used  Better results for face images.	<ul style="list-style-type: none"> <li>• For more redundant data this proposed method shows better results than those of existing methods but for less redundant data in proposed method shows minor improvement of the results of the existing methods.</li> </ul>
4.	<b>Paper:</b> Implementation of Multi wavelet Transform Coding for Lossless Image Compression.	Explains the concept of wavelet and multi wavelet transform.  Good results for images having high frequencies.	—

	<b>Author:</b> K.Rajakumar and T.Arivoli		
5.	<b>Paper:</b> Performance Analysis of Integer Wavelet Transform for Image Compression.  <b>Author:</b> Chesta Jain and Vijay Chaudhary <i>et al.</i> (2011)	Suitable for the applications where speed is a critical factor.	<ul style="list-style-type: none"> <li>• Energy Compaction is desirable.</li> </ul>
6.	<b>Paper:</b> An Efficient Image Compression Technique Using Discrete Wavelet Transform (DWT).  <b>Author:</b> Rajasekhar V. <i>et al.</i> (2014)	Based on JPEG-2000 scheme. Information is divided into subparts.	<ul style="list-style-type: none"> <li>• Not suitable designing embedded hardware architecture for discrete wavelet transform</li> <li>• Storage requirement are the main challenges in the system.</li> </ul>
7.	<b>Paper:</b> IMAGE COMPRESSION USING CALIC  <b>Author:</b> Miss. Rohini N. Shrikhande and Dr.Vinayak K. Bairagi (2014)	Gives high compression ratio.	—
8.	<b>Paper:</b> Implementation of Region based medical image compression for Telemedicine application  <b>Author:</b> P. Eben Sophia and J. Anitha (2014)	Different Algorithms are used for comparison.  ROI based algorithm is used which increase the CR as compared to the traditional block based methods.	<ul style="list-style-type: none"> <li>• Large part of the ROI gives less compression ratio.</li> </ul>
9.	<b>Paper:</b> Region Based Lossless Compression for Digital Images in Telemedicine Application  <b>Author:</b> B. Brindha, and G. Raghuraman (2013)	Segmentation is used for locate the objects and boundaries in image. In this paper a lossless compression based method is proposed.  .	<ul style="list-style-type: none"> <li>• Segmentation is not unique for all images.</li> </ul>
10.	<b>Paper:</b> An Improved Active Contour Medical Image Compression Technique with Lossless Region of Interest  <b>Author:</b> Loganathan R and Y.S.Kumaraswamy (2011)	This paper is based on medical image compression. Important part of the image is marked as ROI.	<ul style="list-style-type: none"> <li>• Further work needs to be done in the area of optimizing the proposed active contour mechanism.</li> </ul>
11.	<b>Paper:</b> A Survey on Image Compression Techniques  <b>Author:</b> Pratishttha Gupta <i>et al.</i> (2014)	This document presents the review of various Lossless and Lossy techniques.	<ul style="list-style-type: none"> <li>• Need an efficient algorithm for reducing the drawbacks of existing algorithms.</li> </ul>

12.	<b>Paper:</b> Quad Tree Decomposition Based Analysis of Compressed Image Data Communication for Lossy and Lossless Using WSN <b>Author:</b> N. Muthukumaran and R. Ravi (2014)	<p>The proposed method is based on Quad tree decomposition for image compression. In this Research analysis compression and restoration of sequentially transmitted images over Wireless Sensor Networks (WSNs) has been presented. The Quad Tree Decomposition based performance analysis of compressed image data communication for Lossy and Lossless through wireless sensor network is presented.</p>	
13.	<b>Paper:</b> A Quantitative Assessment of Image Compression Parameters And Its Algorithm. <b>Author:</b> K Chithra <i>et al.</i> (2015)	<p>Comparisons of different Lossless Image Compression algorithms are performed with performance parameters like CR, PSNR, BPP.</p>	
14.	<b>Paper:</b> Compression Efficiency for Combining Different Embedded Image Compression Techniques with Huffman Encoding <b>Author:</b> Sure Srikanth and Sukadev Meher (2013)	<p>This paper contributes to the implementation of the combining of Lossy image compression techniques(EZW, SPIHT, Modified SPIHT algorithms) with Huffman encoding, and also discussing their advantages and disadvantages of Lossy Compression techniques.</p>	<ul style="list-style-type: none"> <li>• PSNR is low.</li> </ul>

### 3.2. Sources of Information

Recommended and used sources of information during study are:

- IEEE Xplore Digital library (IEEE Geoscience and Remote Sensing Letters)
- Science Direct (Elsevier's Information) [www.sciencedirect.com](http://www.sciencedirect.com), <https://www.elsevier.com>
- CiteseerX, Books and Wikipedia
- Springer([www.springer.com](http://www.springer.com))
- Wiley Online Library (<http://onlinelibrary.wiley.com/>)
- Scholarly Articles(Google Scholar) <http://scholar.google.com>

## 4. PROPOSED APPROACH TO REINFORCE ROI BASED COMPRESSION METHOD

An approach based on separation of ROI and NROI regions for medical images is proposed to reinforce ROI based compression method [5]. Steps involved in the procedure are presented below:

- 1) Initially the image is selected from the given set of images so that the compression is done on the selected image

- 2) The image is divided into two parts i.e. ROI part and the Non-ROI part. Firstly the ROI part of the browsed image is selected and on that selected part the Huffman encoding algorithm is applied.
- 3) After encoding the image, next step is to decode image using Huffman decoding algorithm that was earlier encoded by using Huffman encoding algorithm. The original ROI is obtained.
- 4) On the Non-ROI part of the browsed image applies Quad Tree decomposition [15]; now apply the compression on the image. Final compressed image is obtained.
- 5) In this step both the part of the image after compression are combined. Final compressed image is obtained.
- 6) Now the calculation of Compression Ratio (CR), Peak Signal to Noise Ratio (PSNR), Mean Square Error (MSE) and Bits Per Pixel (BPP) is done of the image by using both techniques.
- 7) Finally on the basis of the parameters calculated, that will show the performance of the proposed method of compression.

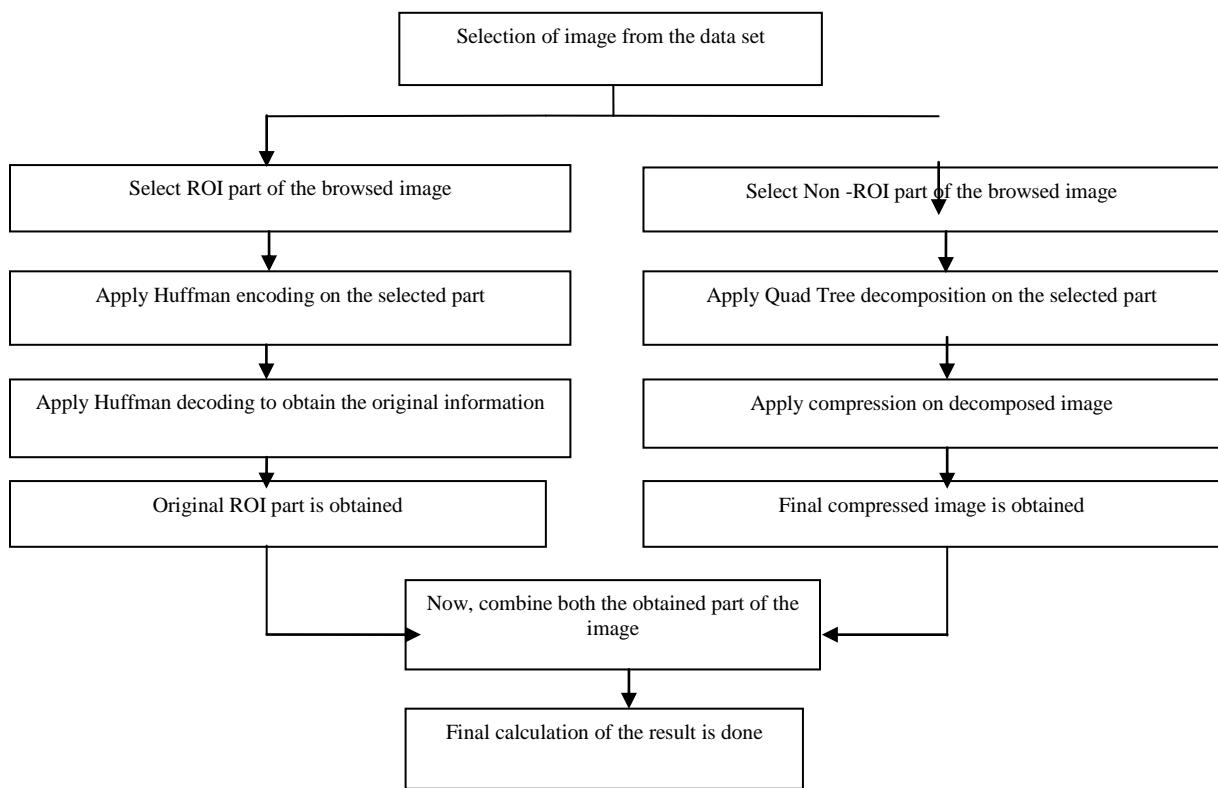


Fig.2. Block Diagram of Proposed Method

#### 4.1 Evaluation Metrics

Traditional framework of evaluation consists of PSNR (Peak Signal To Noise Ratio), MSE (Mean Square Error), CR (Compression Ratio), BPP (Bits Per Pixel).

##### *Peak Signal To Noise Ratio (PSNR)*

- Peak Signal To Noise Ratio avoids the problem by scaling the MSE according to the image.

- Image scaling is the process of resizing a digital image.
- PSNR is measured in decibels (dB).
- PSNR is a good measure for comparing restoration results for the same image.

$$\text{PSNR} = 20 \times \log \frac{255}{\overline{\text{MSE}}} \quad (4.1)$$

#### *Mean Square Error (MSE)*

- MSE is the Mean Square Error between the original and reconstructed data.
- One problem with Mean Square Error is that it depends strongly on the image scaling.

$$\text{MSE} = \frac{m}{i=1} \frac{n}{j=1} \frac{\sum_{i,j} (X_{i,j} - Y_{i,j})^2}{m \times n} \quad (4.2)$$

#### *Compression Ratio (CR)*

- Compression Ratio is ratio between memory space needed to store compressed data. i.e code stream.

$$\text{Compression Ratio} = \frac{\text{Original image size}}{\text{Compressed image size}} \quad (4.3)$$

#### *Bits Per Pixel (BPP)*

- Bits per pixel (BPP) give the number of bits that can be stored in one pixel of the given input image [6].

$$\text{BPP} = \frac{\text{Size Of Compressed File}}{\text{Total No.Of Pixels In The image}} \quad (4.4)$$

#### 4.2 Experimental Results

To perform evaluation and comparison studies of above mentioned methods, experiments are set up in MATLAB7.10.0 (R2010a) on i3 Processor. Compression results of few images from the tested image set are presented for both subjective and objective evaluation.

TABLE II. INPUT IMAGES



TABLE III. DECOMPRESSED IMAGES BY HUFFMAN CODING



TABLE IV. DECOMPRESSED IMAGES BY ARITHMETIC CODING



TABLE V. DECOMPRESSED IMAGES BY RUN LENGTH ENCODING

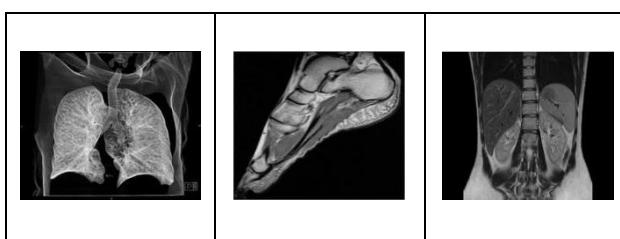


TABLE VI. DECOMPRESSED IMAGES BY PROPOSED METHOD

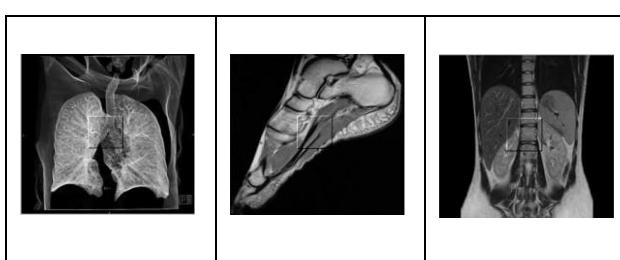


TABLE VII. EXPERIMENTAL RESULTS OF IMAGE COMPRESSION TECHNIQUES

Method	CR	MSE	BPP	PSNR
Huffman Coding	0.9	6.3	1.0	40
Arithmetic Coding	2.3	6.3	0.4	40
Run Length Encoding	2.0	6.5	0.5	39.66
Proposed Method	3.86	4.1	0.2	44.33

## 5. DISCUSSION

Image Compression is done in order to reduce the size of the image so that more of data is stored. Compression Ratio is usually less in lossless compression but image quality is better in lossless compression as compared to lossy compression. On the other hand, compression Ratio is high in lossy compression but image quality is not so good. Image quality improves when the value of PSNR increases. The value of PSNR decreases with increase in CR. Performance evaluation of studied image compression techniques is done using parameters like Compression Ratio (CR), Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR) and Bits Per Pixel (BPP). Experimental results for various methods on medical images are presented in tables II-VII. Huffman Coding gives less Compression Ratio as compared to other methods but it gives good PSNR Value. Value of BPP is more in Huffman Coding. RLE gives more Compression Ratio as compared to Huffman but less by Arithmetic Coding. PSNR value is less as compared to Huffman Coding and Arithmetic Coding. Value of MSE is same in Huffman Coding and Arithmetic Coding. Arithmetic Coding uses less number of bits per pixel as compared to Huffman and RLE. Proposed method gives highest values for Compression ratio and PSNR. Value of MSE and BPP are less as compared to other three techniques.

## 6. CONCLUSION

In this study, techniques for image compression are reviewed presenting their achievements and gaps. Experimental results of performance evaluation of few state-of-the-art compression techniques demonstrate that proposed technique outperforms other techniques. In future the approach can be further enhanced for reducing storage space, transmission time and for enhancement of computational speed.

## REFERENCES

- [1] R. Gonzalez and R.Woods, " Digital Image Processing", 2009.
- [2] M. Yang and N.Bourbakis, "An Overview of Lossless Digital Image Compression Techniques", 2005.
- [3] Pratishtha Gupta *et al.*, "A Survey on Image Compression Techniques", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 3(8) , August 2014.
- [4] K Chithraet *et al.*, "A Quantitative Assesment of Image Compression PaarametersAnd Its Algorithm", *Global Conference on Communication Technologies*, 2015.
- [5] P. Eben Sophia and J. Anitha, "Implementation of Region based medical image compression for Telemedicine application", *IEEE International Conference on Computational Intelligence and Computing Research*, 2014.
- [6] R.Praisline Jasmiet *et al.*, "Comparision of Image Compression Techniques Using Huffman Coding, DWT and Fractal Algorithms", *International Conference on Computer Communication and Informatics(ICCCT)*, 2015.
- [7] Rajasekhar V and Vaishnavi V *et al.*, "An Efficient Image Compression Technique Using Discrete Wavelet Transform (DWT)", *International Conference on Electronics and Communication System (ICECS)*, 2014.
- [8] B. Brindha, and G. Raguraman, "Region Based Lossless Compression for Digital Images in Telemedicine Application", *International conference on Communication and Signal Processing*, 2013.
- [9] R. Shrikhande and V.Bairagi, "Image Compression Using CALIC", *International Conference on Advances in Communication and Computing Technologies*, 2014.
- [10] Kuldip K. Ade and M. V. Raghunadh, "ROI Based Near Lossless Hybrid Image Compression Technique", *International Conference on Electrical, Computer and Communication Technology (ICECCT)*, IEEE, 2015.

- [11] M.Islam *et al.*, “A New Image Compression Scheme Using Repeat Reduction and Arithmetic Coding,” *12<sup>th</sup> International Conference on Computer and Information Technology*, 2009.
- [12] K.Rajakumar and T.Arivoli, “Implementation of Multiwavelet Transform Coding for Lossless Image Compression”, *International Conference and international Communication Embedded System(ICICES)*, 2013.
- [13] C.Jain and V. Chaudhary, “Performance Analysis Of Integer Wavelet Transform for Image Compression”, *International Conference on Electronics Computer Technology*, 2011.
- [14] Loganathan R *et al.*, “An Improved Active Contour Medical Image Compression Technique with Lossless Region of Interest”, 2011.
- [15] N. Muthukumaran and R. Ravi, “Quad Tree Decomposition Based Analysis of Compressed Image Data Communication for Lossy and Lossless Using WSN”, *International Journal of Computer, Electrical, Automation, Control and Information Engineering* Vol. 8(9), 2014.
- [16] Sure Srikanth and Sukadev Meher, “Compression Efficiency for Combining Different Embedded Image Compression Techniques with Huffman Encoding”, *International conference on Communication and Signal Processing*, 2013.

# Comparative Analysis of Phylogenetic Tree Creation Tools

Arshdeep Kaur

Dept. of Information Technology

Guru Nanak Dev Engineering College

Ludhiana. Punjab

arhsamra25@gmail.com

Dr. Manpreet singh

Dept. of Information Technology

Guru Nanak Dev Engineering College

Ludhiana. Punjab

mpreet78@gmail.com

**Abstract-** Bioinformatics is a platform that developing a framework that supports the existing and upcoming technologies for decision making. Bioinformatics main motive is to develop tools that analyzes and execute biological data for better results. Major Focus of research is on the practices that understand and utilize data from innovative technologies and translate these to real application. The branch of science which deals with the study of relationship among different group of organisms and their evolutionary history is called Phylogenetic. These relationships are evaluated by different methods that analyses the observed values of different species or population such as DNA sequences. The result is known as Phylogency. The dataset varies from living to non living population. POWER, Delign, BLAST, HMM (Hidden Markov Model), ClustalW, MEGA6 were analyzed on different data ranges. R and RAXML tools are open source are capable of handling huge data in less space and time. POWER and CVTree3 (Third version of CVTree) tools also provide Web services. ClustalW tool performs well with similar length of sequences.HMM use series of successive instances of data and information of previous state that can be presented sequentially. PHYLIP tool is appropriate for prediction purpose. BLAST is effective in low to medium biological data. Handle Biological data efficiently with in limited resources and time is the main focus of research. R tool is more efficient in handling millions of observations. The comparison of different methods is discussed further in the table.

## 1. Introduction

### A. BIOINFORMATICS

In biomedical domains researches mainly face problem in collecting resources. These resources are available in large datasets. Then second problem arise how to process these datasets effectively and efficiently to avoid failure.[1] An approach is required that access these resources in flexible time with changing requirements. Few years ago, more than 90% of interactions between pattern and bioinformatics are unidirectional.[2] Only PR tools are used to analyze the biological datasets. But now many alternatives approaches are available that

solve the problems occurs while using PR tools. As we know that significant research is still needed for the better applications of bioinformatics.[3] Bioinformatics need further development to improve the storage management of data, integrate databases easily and securely, for better standards of data analyses process. Still bioinformatics have many beneficial applications like data quality, data validation tools. Bioinformatics ensure safe introduction and greater use of these applications.[4] Center of Excellence in Regulatory Science and Innovation (CERSI) and Global Coalition for Regulatory Science Research (GCRSR) are the center that aim is to build knowledge to identify research.[5]

#### A. *PHYLOGENETIC TREE*

The word “Relationship” is the key to understand the tree structure. This also tells the way of using the tree. The phylogenetic tree is about the specification of data in which one species give rise to many other species.[6] This gives relationships to the ancestors which are shared by common species. A path is created between any two species in the way other species ancestor comes that must have to be crossed to reach the located species. This means the branches can be rearranged but relationship should be the same.[7] The data is divided in small chunks that all are related to one another with correct order. Phylogenetic tree are of three types Rooted, Unrooted, Bifurcating Trees. Some rules and regulation should be trails while creating and analyzing the tree. Phylogenetic systematic is the way to understand the inter relationship of the data.[8]

#### B. *PHYLOGENETIC SYSTEMATIC*

Phylogenetic systematic is the way to understand the inter relationship of the data. This will try to interpret the data can be about living or extinct. To understand this large amount of data need classification. The data is classified into groups that have their unique name and properties. The classification should be meaningful.[9] The data is about the evolutionary history of life. New relationships are discovered that have their known organisms. This will elucidate new theories about mechanisms evolution. It deals with the pattern of taxa relationship. History is part that cannot be seen only the clues of actual events are discovered. These clues are used to build hypothesis. While dealing with history one should keep in mind he is he/she is dealing with something fundamentally impossible[10]. Phylogenetic systematic is the name given to the field that reconstruct the history and study the relationship among organisms. Systematic uses the clues of history. By the evolution of the theory one task of biology has to discover the phylogenetic relationships between species.[11]

## 2. Material and Methods

#### A. *R Tool*

On the starting time of R mostly people think that R is slow, use more memory and is unable to handle large sets of data. Now days the system works very fast with large memory space. Many competitions are handling with very small memory. Workstations with large memory can handle millions of observations.[12] Various tools in R are used in connectivity of data in datasets but these are not shown to the users that are hide complexity. R is very powerful system application that distributes the code on back end and front end data significantly.[13] Basic data execution was performed in command line infrastructure with small code. Short

code programs are easily distributed without translation. Large programs sometimes need translation before distribution. Windows, Mac are used for building process of R GUI. Full help information is available in Data Import/Export Manual of R.[14] R is able to catch information from text files, catch from web pages directly and can group the whole database also. R works similarly with the command that do some action and the command that compute data. Every output in R is considered as value. R software can store the whole data frame. One data frame can store more than one datasets. R is user friendly language.[15] Functions in R have some optional arguments. Users are free to choose the value and function arguments that they want to execute. Data frame have their unique names that are used to access the name of variables of data frame while executing the data sets.[16] R can move through the packages. Package library show the entire package used by the programs. Users are free to choose the function from different packages for execution. Foreign package is used to handle export and import of data. R uses string on regular expressions which are like other simple programs of strings. R produces the graphics in various styles like on screen, PDF, JPEG, PNG etc.[17]

**B. POWER (*Phylogenetic Web Repeater*)**

POWER performs multiple phylogenetic analysis, POWER uses open source service like LAMP structure for their analysis as well as use well established algorithms like ClustalW and PHYLIP.[18] According to the calculated result this involves a tree builder to generate a high quality tree topology. Tree builder is based on GD Library. POWER works equally with raw sequences or user uploaded files. Raw sequences are in FASTA format, POWER has user friendly web interface. Users can easily use the interface and sketch the tree effortlessly. Trees can be generated and edited at multiple steps. While creating trees users can easily describe the parameters. Users are free to choose the tree building algorithms. Sequence alignment can be changed at any step or the same for the tree topology (can be edited easily). While working with POWER, it stores every step of processing (Processing History) that can be downloadable. The best part of POWER is its working with iterative trees building. User can add sequences and remove sequences from the previous jobs. POWER estimates the evolutionary distance using genetic information rather than traditional information. Each step is conducted separately so the processes become time consuming. PhyloBLAST provide chained phylogenetic analysis. These programs are not designed for general purposes. These have their working limits and can be only used in special areas like PhyloBLAST analysis only protein sequences. Users can not define their own parameters for the process.[19]

**C. HMM (*Hidden Markov Model*)**

This model analyzes all model parameters like coefficient of diffusion as well as errors of the observation model. This is the big advantage of this model. Observation model is used at every time step for relating the coherent observation and state.[20] HMM as it is a statistical tool considered for execution answer understanding problems with series of successive instances of data and information of previous state that can be presented sequentially.[21] Advantages of HMM model is examined in the robotic, Computer science, finance, speech and natural language processing fields. Other models treat each instance of data as independent but HMM treats each instance of data as dependent on previous instances in time series.[22]

**D. ClustalW**

Several new algorithms were designed to solve the problem with multiple sequence alignment. Analysis of protein families and their evolution is the major concern.[23] According to them the best algorithm provides biologically correct alignment and speed up the alignment procedure. [24] Multiple sequence alignment algorithms can be divided in two categories:

1. Global method aligning over entire length
2. Local method aligning over high similarities only.[25]

ClustalW performs well with similar length of sequences. ClustalW becomes most prominent program. ClustalW is a global progressive method.[26] The algorithm performs action in two steps: Guide tree is constructed based on the similarities of sequence then successive pair wise alignment is done. [27]The algorithms also prove their functioning on BALiBASE test set, where the absence of full length sequence shift the test set to word global methods. ClustalW give top results for global alignment with different percentage identity and orphan sequences.[28]

**E. CVTree3**

It is the latest third released version of CVTree. CVTree is a web server that is completely genome based. Popularity of CVTree3 web server is due to some great properties that are, it easily handle huge data as it reside on 64 cores cluster, have very interactive interface and collapsible and expandable tree display.[29] It compares the tree branching order and produces a print quality sub tree. CVTree3 is an open source server that can be accessed by without login requirements.[30] Whole genomes data is given as input CVTree avoids the ambiguities and do alignments free comparison because genomes differ in their size and gene contents.[31] Two versions of CVTree3 web server are released: version 1 in 2001 and version 2 in 2009. CVTree3 server contains many enhanced features of CVTree[32]. CVTree3 is not just a tool, but also combination of the study pattern of phylogency and taxonomy. CVTree3 provide the printing facility for any sub tree. CVTree is a parameter free method.[33]

**F. RAXML**

It is open source tool for phylogenetic tree construction. The result of RAXML increases speed of the algorithm by 5,5 on a 8-care Nehalem node. In this experiment they use 55000 organism's data for building trees. [34]Multiple consensus tree and plausible trees are constructed. Consensus tree have information agreed upon by plausible trees. The mostly assembled consensus tree is called majority rules extended (MRE) tree sometimes greedy consensus tree. [35]The code in RAxML executes parallel with pithead and MPI.[36] MRE algorithm uses RAxML for fastest exact implementation. Four phases of MRE algorithm are tree parsing, extraction and addition of Bipartitions, selection of candidate Bipartitions, Reconstruction of MRE tree.[37]RAxML tool for Maximum Likelihood methods.[38] He uses data on phylogenetic analyses of gazelles (genus gazelle) based on mitochondrial and nuclear intron markers. This data have 48 tissue samples of nine different species. From this data phylogenetic tree is constructed with Bayesian inference and Maximum likelihood methods. [37]BEAST algorithm is used for range estimation. RAxML 8.0.14 is used for Maximum Likelihood approach. Bayesian approach in BEAST was used for estimation of divergence time.[39]

Table 1.1: RAXML Performance

8-care Nehalem node	Dataset	Increased Speed
	55000	5.5%

#### **G. PHYLIP**

PHYLIP is used by researcher and scientists to epitop search. Focusing on PHYLIP only three are many applications that are upgraded time to time to enhance the ability of this tool. [40] Web PHYLIP is a tool that considers the internet manipulation of PHYLIP. This tool is used by “Viroj Wiwanitkit” to epitop his research article on bioinformatics procedures to calculate annual perspective difference of influenza virus.[41] This tool is appropriate and good for prediction purpose. Several other methods are also there, but any computational tool of bioinformatics has its limitations.[42]

#### **H. DELIGN**

DELIGN tool using the fragment based approach. Nucleic acid and protein sequence alignment are constructed. DELIGN construct pair wise sequence segments.[43] It constructs multiple alignments of nucleic acid sequences and protein sequences. These segments are known as fragments. Chain of values of parts can be explained by pair wise alignments. They may have varying structure and have varying length. Our focus moves around the chain of fragments with maximal overall score.[44]

#### **I. BLAST (Basic Local Alignment Search Tool)**

The working of is explained by “Indra Neil Sarkar” in his research time period. He explained his views in “Appearance of new Tetraspanin Genes during Vertebrate Evolution” named research article. In this paper specific Tetraspanin functions have been defined with the broad evolutionary divergence. Phylogenetic context with large gene families are analyzed by this approach. [45]First of all searching the database and include every part with BLAST statistics or BLAT his statistics. BLAST searches whole Tetraspanin family. We find many other hits with BLAST value. Maximum parsimony and neighbors joining approaches are used for tree building.[46] These approaches are used for the understanding of amino acids characteristics. Initially phylogenetic analyses are establishing the relationship with phylogenetic defined group. In second step phylogenetic tree are constructed. First step is based on topological similarities for example DNA sequence of proteins.[47]

#### **J. PHYTOOL**

Phytool is an R loaded package for Phylogenetic. The library is entirely written in R language with the development of multifunctional ape(Analysis of phylogenetic and evolution) the computing environment R is growing very fast. The library of phytool can be easily from the comprehensive R Archive Network. Phytools are capable to generate, plot, read and write the phylogenetic trees. The outputs are stored within the directory of “ape”, no other space required to be reserved. The current phytool library is not capable to work with

Phylobase Package. The package “ape” is used for the Phylogenetic tree capturing and manipulation only. Phytool Package import several other R libraries. Some packages of R library are animation: (Xie,2011), calibrate: (Graffelman,2010), igraph(Csardi and Nepusz,2006). He implements numerous functions of Phytools. Various methods of comparative biology are also implemented by Phytools like estimation of phylogenetic signal, Phylosig. Some simulation methods of Phytools include fastBM, sim.history(Simulation of discrete character evolution). Various phylogenetic inference procedures are also included by Phytool Library(mrp.supertree. this tools is used for matrix representation of super tree). Graphical method of Phytools are Phylomorphospace, plotsimmap etc.

#### *K. MEGA( Molecular Evolutionary Genetic Analysis)*

MEGA software is designed for comparative analysis of DNA and Protein sequences. It studies the inferring pattern of genes, genomes and species time to time.[48] MEGA is available with two editions GUI (Graphical user interface) and command line interface. In GUI visual tools are available for execution and analysis of data. Command line interface is used for integrated and iterative pipeline analysis.[49]The latest version 6 includes the feature of Time tree (it builds the molecular evolutionary tree scaled to time). X Time trees are highly needed by scientists. Relative time estimation is useful for calculating ordering and spacing of sequences. Due to this purpose RelTime is used for greater number of Sequences. It calculates the time of divergence for all branching points of tree and supply the information with clock calibration and associated distributions. RelTime computation in MEGA6 is good for both performance and storage requirements.[48]

Table 1.2 : MEGA Performance

Number of sequences	Length	Time	Memory
765	2000 bp	43 minutes	1 GB

Table 1.3 : Specification of Different tools with Different Datasets

Method Name	Algorithm	Dataset	Advantages	Drawbacks	Conclusion
Power Tool	ClustalW, PHYLIP	Nucleic acid sequence, Protein sequence	Web Based Service, Estimate distances using genetic information instead of traditional, No Problem of branch crossing and label overlapping[18]	Sometime server error occur	Support for Parallel computation, Good for non experts, Flexible Topology[19]
Delign Tool	Fragment Chaining	Alignment of DNA and Protein Sequence	Construct Pair wise multiple alignments from sequence having continuous pair[44]	Hard to find a chain with maximum overall score[43]	Remove fragment with maximum overall score
Blast Tool	Maximum parsimony, Neighbors joining	Tetraspanin super family [Gene Family][47]	Search whole database with Blast hit Statistics, based on topological similarities[45]	Absence of family members are not recognized	Works for DNA sequence of proteins also[46]

HMM Tools	SEM	Pelagic Fish	Use Temperature – Depth conditions for undersea geolocation, can be used for tracking undersea objects, used to locate migratory patterns of undersea population[21]	Purely based on state space model[20]	Depends upon two key components Dynamic model and Observation model[22]
ClustalW	Multiple sequence algorithm		Best for high sequence similarities[23]	Perform poorly in single domain compared with others[25]	Performs well when all the sequence are of same length[25]
MEGA6 [Molecular evolutional genetics analysis]	RelTime	DNA and Protein sequence	<p>Contain facilities for:</p> <ul style="list-style-type: none"> <li>➤ Building sequence alignments</li> <li>➤ Inferring phylogenetic histories</li> <li>➤ Molecular evolutionary analysis[48]</li> </ul> <p>Efficient for performance and memory, Support both GUI and Command interface</p>		Enhanced algorithms to search for optimal trees[49]
R Tool	R Software R Studio	Web Databases Support old datasets [Vector Format]	<p>Open Source, Use Multiple data frames in single run,[12]</p> <p>Support complex web databases, Handle millions of observations in time limits[13]</p>		Produced graphs in multiple styles, [15]Support PDF, JPEG, PNG etc Styles[14]
PHYLIP		Antimalarials dataset	Good for prediction purposes, [40], WebPHYLIP [allow internet manipulations of PHYLIP], Tool can be assessed by Google to update current situations[41]		Support Dynamic datasets[42]
CVtree			Open Source Server, Parameter free method,[29] Compare tree branching order and produce print quality tree, Easily handle huge data as it reside on 64-cores cluster[31]	Completely genome based[32]	Combination of study pattern of phylogency and taxonomy,[33] provide sub tree printing facility[34]
RAXML	Maximum Likelihood method, Bayesian inference method, BEAST	Gazelles [genus gazelle]	Open Source, Fastest exact implementation[35]	Estimation of divergence time[38]	Multiple consensus tree and plausible trees can be constructed[37]

### 3. Conclusion

Purpose of bioinformatics is to convert the data into real time applications. Understanding of interrelationship of data in must for quality phylogenetic tree. Multiple tools were applied on different data sets having varying data structures and parameter. Depending upon requirements and dataset, particular tool is applied. Better understating of data helps to choose the tool for better results. These tools have their gain and losses. Choice of appropriate tool results in useable real time application.

### References

- [1] J. D. Tenenbaum, "Translational Bioinformatics: Past, Present, and Future," *Genomics, Proteomics Bioinforma.*, vol. 14, no. 1, pp. 31–41, 2016.
- [2] T. Weber and H. U. Kim, "The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production," *Synth. Syst. Biotechnol.*, 2016.
- [3] D. J. Mcmillan, P. A. Drèze, T. Vu, D. E. Bessen, J. Guglielmini, A. C. Steer, J. R. Carapetis, L. Van Melderen, K. S. Sriprakash, P. R. Smeesters, Michael Batzloff, R. Towers, H. Goossens, S. Malhotra-Kumar, L. Guilherme, RosangelaTorres, D. Low, A. Mc Geer, P. Krizova, S. El Tayeb, J. Kado, M. van der Linden, G. rdem, A. Moses, R. Nir-Paz, T. Ikebe, H. Watanabe, S. Sow, B. Tamboura, B. Kittang, J. Melo-Cristino, M. Ramirez, M. Straut, A. Suvorov, A. Totolian, M. Engel, B. Mayosi, A. Whitelaw, J. Darenberg, B. H. Normark, C. Chiang Ni, J. J. Wu, A. De Zoysa, A. Efstratiou, S. Shulman, and R. Tanz, "Updated model of group A Streptococcus M proteins based on a comprehensive worldwide study," *Clin. Microbiol. Infect.*, vol. 19, no. 5, 2013.
- [4] J. Shen, Q. Cong, and N. V. Grishin, "The complete mitochondrial genome of Papilio glaucus and its phylogenetic implications," *Meta Gene*, vol. 5, pp. 68–83, 2015.
- [5] C. a Sarkar, "Notes originally scribed in April 1998, and last modified October 2002. 1," *October*, no. October 2002, pp. 1–22, 1999.
- [6] J. Park, J. Park, W. Song, S. Yoon, B. Burgstaller, and B. Scholz, "Treegraph-based Instruction Scheduling for Stack-based Virtual Machines," *Electron. Notes Theor. Comput. Sci.*, vol. 279, no. 1, pp. 33–45, 2011.
- [7] O. Torres-reyna, "Introduction to RStudio," no. August, pp. 1–16, 2013.
- [8] M. Craven, "Inferring Phylogenetic Trees Phylogenetic Inference : Task Definition."
- [9] W. Hennig, "Phylogenetic Systematics," *Annual Review of Entomology*, vol. 10, no. 1, pp. 97–116, 1965.
- [10] M. Padamsee, R. B. Johansen, S. A. Stuckey, S. E. Williams, J. E. Hooker, B. R. Burns, and S. E. Bellgard, "The arbuscular mycorrhizal fungi colonising roots and root nodules of New Zealand kauri Agathis australis," *Fungal Biol.*, vol. 120, no. 5, pp. 807–817, 2016.
- [11] K. Dowell and K. Dowell, "Molecular Phylogenetics," 2008.
- [12] T. Lumley, "R Fundamentals and Programming Techniques," *R Man.*, pp. 1–225, 2006.
- [13] K. Kourtit, M. M. Marinescu, and P. Nijkamp, "Growth Modelling of Metropolitan Performance Indicators. An Application by Means of R Software," *Procedia Econ. Financ.*, vol. 10, no. 14, pp. 314–323, 2014.
- [14] T. Jombart, D. M. Aanensen, M. Baguelin, P. Birrell, S. Cauchemez, A. Camacho, C. Colijn, C. Collins, A. Cori, X. Didelot, C. Fraser, S. Frost, N. Hens, J. Hugues, M. Höhle, L. Opatowski, A. Rambaut, O. Ratmann, S. Soubeyrand, M. A. Suchard, J. Wallinga, R. Ypma, and N. Ferguson, "OutbreakTools: A new platform for disease outbreak analysis using the R software," *Epidemics*, vol. 7, pp. 28–34, 2014.
- [15] G. R. Guerin and A. J. Lowe, "Mapping phylogenetic endemism in R using georeferenced branch extents," *SoftwareX*, vol. 3–4, pp. 22–26, 2015.
- [16] C. Blume, N. Santhi, and M. Schabus, "MethodsX ' nparACT ' package for R : [ 61 \_ TD \$ IF ] A free software tool for the non-parametric analysis of actigraphy data," *MethodsX*, vol. 3, pp. 430–435, 2016.
- [17] W. Rui, H. Chen, Y. Feng, Z. Shi, and M. Jiang, "Procedia Engineering Classification of Astragalus Membranaceus ( Fisch .) Bge . Var . Mongholicus ( Bge .) Hsiao from Different Areas Based on Chemometric Methods with R Software," vol. 29, pp. 2172–2176, 2012.

- [18] C. Y. Lin, F. K. Lin, C. H. Lin, L. W. Lai, H. J. Hsu, S. H. Chen, and C. A. Hsiung, “POWER: Phylogenetic WEb Repeater - An integrated and user-optimized framework for biomolecular phylogenetic analysis,” *Nucleic Acids Res.*, vol. 33, no. SUPPL. 2, pp. 553–556, 2005.
- [19] *Phylogenetic Networks*. .
- [20] S. Ntalampiras, Y. Soupionis, and G. Giannopoulos, “A fault diagnosis system for interdependent critical infrastructures based on HMMs,” *Reliab. Eng. Syst. Saf.*, vol. 138, pp. 73–81, 2015.
- [21] M. Woillez, R. Fablet, T. T. Ngo, M. Lalire, P. Lazure, and H. de Pontual, “A HMM-based model to geolocate pelagic fish from high-resolution individual temperature and depth histories: European sea bass as a case study,” *Ecol. Modell.*, vol. 321, pp. 10–22, 2016.
- [22] A. Faridi and M. M. H. Rahman, “HMM as an Inference Technique for Context Awareness,” *Procedia Comput. Sci.*, vol. 59, no. Iccsci, pp. 454–458, 2015.
- [23] E. Celli, I. Gabrielli, G. Zehender, M. Giovanetti, A. Lo Presti, A. Lai, G. Dicuonzo, S. Angeletti, M. Salemi, and M. Ciccozzi, “Phylogeny of Murray Valley encephalitis virus in Australia and Papua New Guinea,” *Asian Pac. J. Trop. Med.*, vol. 9, no. 4, pp. 385–389, 2016.
- [24] A. D. Corebima and M. Rengkuan, “The Detection of Superior PO Cattle ( Ongole Cattle Descendant ) Based on The Growth Hormone Gene Sequence,” vol. 1, no. January, pp. 363–368, 2012.
- [25] T. Lassmann and E. L. Sonnhammer, “Quality assessment of multiple alignment programs,” *FEBS Lett.*, vol. 529, no. 1, pp. 126–130, 2002.
- [26] R. Y. Sertoz, S. Erensoy, and O. Tijen, “Hepatitis delta virus ( HDV ) genotypes in patients with chronic hepatitis : molecular epidemiology of HDV in Turkey,” pp. 58–62, 2007.
- [27] E. M. Adriaenssens, R. Edwards, J. H. E. Nash, P. Mahadevan, D. Seto, H. Ackermann, R. Lavigne, and A. M. Kropinski, “Integration of genomic and proteomic analyses in the classification of the Siphoviridae family,” *Virology*, vol. 477, pp. 144–154, 2015.
- [28] J. Arockiaraj, A. J. Gnanam, D. Muthukrishnan, M. Kuppusamy, M. Pasupuleti, J. Milton, and M. Kasi, “Macrobrachium rosenbergii cathepsin L : Molecular characterization and gene expression in response to viral and bacterial infections,” *Microbiol. Res.*, vol. 168, no. 9, pp. 569–579, 2013.
- [29] F. Zhao and V. B. Bajic, “The Value and Significance of Metagenomics of Marine Environments,” *Genomics. Proteomics Bioinformatics*, vol. 13, no. 5, pp. 271–274, 2015.
- [30] G. Zuo, Z. Xu, H. Yu, and B. Hao, “Jackknife and Bootstrap Tests of the Composition Vector Trees,” *Genomics. Proteomics Bioinformatics*, vol. 8, no. 4, pp. 262–267, 2010.
- [31] V. Kubicova and I. Provaznik, “Use of whole genome DNA spectrograms in bacterial classification,” vol. 69, pp. 298–307, 2016.
- [32] G. Zuo, Z. Xu, and B. Hao, “Shigella Strains Are Not Clones of Escherichia coli but Sister Species in the Genus Escherichia,” *Genomics. Proteomics Bioinformatics*, vol. 11, no. 1, pp. 61–65, 2013.
- [33] G. Zuo and B. Hao, “CVTree3 Web Server for Whole-genome-based and Alignment-free Prokaryotic Phylogeny and Taxonomy,” *Genomics, Proteomics Bioinforma.*, vol. 13, no. 5, pp. 321–331, 2015.
- [34] D. Trojan, L. Schreiber, J. T. Bjerg, A. Bøggild, T. Yang, K. U. Kjeldsen, and A. Schramm, “A taxonomic framework for cable bacteria and proposal of the candidate genera *Electrothrix* and *Electronema*,” *Syst. Appl. Microbiol.*, vol. 39, no. 5, pp. 297–306, 2016.
- [35] M. M. Thomson and A. Fernández-garcía, “Phylogenetic structure in African HIV-1 subtype C revealed by selective sequential pruning,” *Virology*, vol. 415, no. 1, pp. 30–38, 2011.
- [36] D. F. Mokodongan and K. Yamahira, “Mitochondrial and nuclear phylogenetic trees and divergence time estimations of Sulawesi endemic Adrianichthyidae,” *Data Br.*, vol. 5, pp. 281–284, 2015.
- [37] A. J. Aberer, N. D. Pattengale, and A. Stamatakis, “Parallel computation of phylogenetic consensus trees,” *Procedia Comput. Sci.*, vol. 1, no. 1, pp. 1065–1073, 2010.
- [38] A. Trimeche, A. Sakly, and A. Mtibaa, “FPGA Implementation of ML, ZF and MMSE Equalizers for MIMO Systems,” *Procedia Comput. Sci.*, vol. 73, no. 1877, pp. 226–233, 2015.
- [39] H. Lerp, S. Klaus, S. Allgöwer, T. Wronski, M. Pfenninger, and M. Plath, “Data on phylogenetic analyses of gazelles (genus *Gazella*) based on mitochondrial and nuclear intron markers,” *Data Br.*, vol. 7, pp. 551–557, 2016.
- [40] P. Chahar, S. S. Gill, and R. Gill, “Molecular cloning and production of type III Hsp40 protein co-chaperone PfZRF1 of human malaria parasite *Plasmodium falciparum*,” *Int. J. Infect. Dis.*, vol. 45, p. 355, 2016.

- [41] J. Hill, L. Tyasa, L. H. Phylic, J. Kay, B. M. Dunnb, and C. Berrya, "High level expression and characterisation of Plasmepsin II , an aspartic proteinase from *Plasmodium falciparum*," vol. 352, pp. 155–158, 1994.
- [42] V. Wiwanitkit, "Bioinformatic methods to analyze annual perspective changes of influenza viruses," *J. Formos. Med. Assoc.*, vol. 115, no. 1, p. 59, 2016.
- [43] P. Sol, "he covering radius 2-designs in 2Ok," vol. 33, pp. 215–224, 1991.
- [44] B. Morgenstern, "A simple and space-efficient fragment-chaining algorithm for alignment of DNA and protein sequences," *Appl. Math. Lett.*, vol. 15, no. 1, pp. 11–16, 2002.
- [45] J. Hill, L. H. Phylic, and T. T. G. T. T. A. A. C, "Bacterial aspartic proteinases," *FEBS Lett.*, vol. 409, no. 3, pp. 357–360, 1997.
- [46] B. M. Dunn, J. Kay, H. Phylic, J. S. Millsb, and B. F. Parten, "Intrinsic activity of precursor," vol. 314, no. 3, pp. 449–454, 1992.
- [47] A. Garcia-España, P. J. Chung, I. N. Sarkar, E. Stiner, T. T. Sun, and R. DeSalle, "Appearance of new tetraspanin genes during vertebrate evolution," *Genomics*, vol. 91, no. 4, pp. 326–334, 2008.
- [48] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "MEGA6: Molecular evolutionary genetics analysis version 6.0," *Mol. Biol. Evol.*, vol. 30, no. 12, pp. 2725–2729, 2013.
- [49] A. Ruan and C. Liu, "Analysis of effect of nicotine on microbial community structure in sediment using PCR-DGGE fingerprinting," *Water Sci. Eng.*, vol. 8, no. 4, pp. 309–314, 2015.

# A Survey on Removal of Gaussian Noise from Face Images

Nirvair Neeru<sup>1</sup>, Dr. Lakhwinder Kaur<sup>2</sup>

<sup>1</sup>Asstt. Prof., DCE, Punjabi University, Patiala.

<sup>2</sup>Prof., DCE, Punjabi University, Patiala

**Abstract—** In digital image processing the face images play important role in many applications like identification or verification of a person, emotion detection and classification and in many of security related applications. This paper includes the survey of well known noise removal filters to remove gaussian noise in spatial domain. The experiments are performed on the face images taken from Female Facial Expression (JAFFE) database. The performance of all filters has been evaluated at various noise densities in terms of SNR, EPI, SSIM and MS-SSIM.

**Keywords**—Denoising, Gaussian Noise, Filters.

## I. INTRODUCTION

A face image is a typical method of biometric identification technology. Face images are used for public security, identity verification, to monitor criminals etc. to implement all these application the initial input is an face image [14 ]. Hence the success of every application depends upon the quality of input i.e. face image. If initial input is computed than the [performance of following application automatically degrades. Hence it is very necessary to remove noise from image before any further processing. this paper deals with the removal of gaussian noise from face images. The number of filters has been discussed in literature. Non linear filters are very effective in removing noise without harming the edges such as median filters or rank order statistical filters and Gaussian filter etc. the Gaussian filter is designed to reduces Gaussian noise [10] Tuan-Anh Nguyen et al. [11] used a spatially additive Gaussian filter to remove the noise. In 2007, V.R. Vijay Kumar et al. [13] removed Gaussian noise using adaptive window based algorithm. In 2011, Yiwen Qiu et al. [15] presented an efficient method to remove mixture gaussian noise from an image.. The result demonstrates that the proposed method efficiently removes the noise from image.. In 2008, Kun He et.al [7] removed the Gaussian noise through proposed new algorithm which is based on the local smooth region of the image .

## II. GAUSSIAN NOISE

It is also known as Gaussian distribution. It has a probability density function (PDF) of the normal distribution. This noise is added to image during image acquisition like sensor noise caused by low light, high temperature, transmission e.g. electronic circuit noise. This noise can be removed by using spatial filtering (mean filtering, median filtering and gaussian smoothing) by smoothing the image but smoothing also blurs the fine-scaled image edges and details.. It can also be removed by applying transformation techniques like wavelet transform on the noisy image The PDF of Gaussian Noise is shown in the following equation and figure [1]:

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(z-\mu)^2/2\sigma^2} \quad (1)$$

PDF of Gaussian noise [1] is shown in Fig. 1

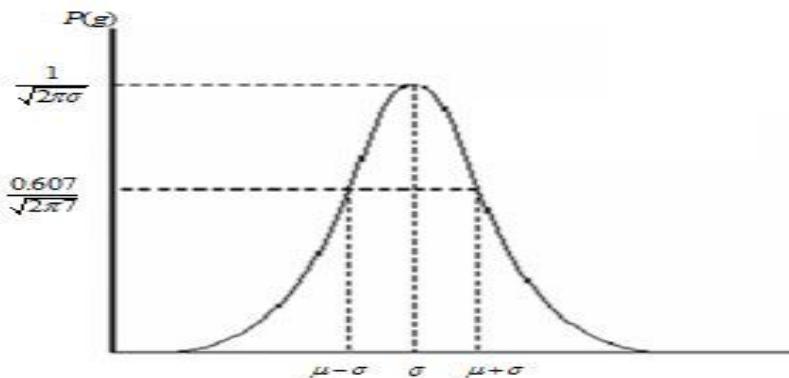


Figure 1 PDF of Gaussian noise

### III. SPATIAL FILTERING

Spatial filters are the traditional way to remove noise from image. Spatial filters can be further classified into non-linear and linear filters. Non-linear filters remove the noise without identify it. In Spatial filters, a low pass filtering is employed on groups of pixels. spatial filters are well capable to remove noise to a reasonable extent but they cause blurriness in the image which leads to loss of edge information. In recent years, a variety of nonlinear median type filters such as weighted median [9], rank conditioned rank selection [8], and relaxed median [2] have been developed to overcome this drawback. Linear filter, like mean filter is the optimal for Gaussian noise in the sense of mean square error. Linear filters also introduce blueness in the image and destroy the edges and fine lines.

**Median filter** is a nonlinear filter. It preserves edges while removing noise. [4] The median filter replaces the pixel with the middle value, which is calculated from sorted list of all pixels of surrounding neighborhood. [3]. The median is more effective than the mean but the main drawback of the median filter is its high computational cost.

**The Gaussian Filter** [6] is similar to the mean filter, but it uses a different kernel that represents the shape of a gaussian hump.

**Disk filter** [5] uses a circular averaging filter (pillbox) within the square matrix of side 2\*radius+1. The circular averaging filter has circular top and straight sides.

**Laplacian operator** is good at finding the fine detail in an image. Laplacian operator enhances the feature by removing discontinuities from image. Laplacian operator restores the fine details of image after removing noise from it [11].

**Log filter** returns a rotationally symmetric Laplacian of Gaussian filter of size hsize with standard deviation sigma (positive). Hsize can be a vector specifying the number of rows and columns in h, or it can be a scalar, in which case h is a square matrix.

**Unsharp filtering** is used commonly used to make sharp edges. A signal obtained by through unsharp or low-pass filtering of the image is subtracted from the image. This is equivalent to adding the gradient or a high-pass

signal to the image [5].

**Wiener filter** is the MSE-optimal stationary linear filter for images degraded by additive noise and blurring. Calculation of the Wiener filter requires the assumption that the signal and noise processes are second-order stationary (in the random process sense). Wiener (1949) proposed the concept of Wiener filtering in two ways; i.e. in spatial domain and in frequency domain [5]. The frequency domain method is used only for denoising and deblurring, whereas the spatial domain method is used for denoising..

**Alpha trimmed mean filter** [5] is an order statistic filter and based on average of the pixels that fall within the window. It is same as average filter, but the difference is that user can clip some of the pixels by specifying an alpha value.

#### IV. IMPLEMENTATION AND RESULT

All the experiments have been performed in Matlab software. The performance has been evaluated on some standard face images taken from JAFEE database, some sample images have shown in fig.1. The quality of recovered images has been evaluated in terms of well known parameters like SNR, EPI, SSIM and MS-SSIM. It has been observed that wiener filter provides better SNR value than other filters which are discussed in this paper. On the other hand the alpha trimmed mean filter retains the original structure of image. The wiener filter also provides good results regarding similarity in structure of original image and denoised image. By analyzing EPI values, it has been observed that wiener filter is also capable to preserve the edges of the image. Median filter also performs well but it damages the edge information from the image. From the visual results it has been seen that it introduces blurriness in the image. Table 1 & 2 shows the results of all parameters calculated with Image 1 & Image 2 respectively. Figure 3, 4, 5 and 6 demonstrates the graphical representation of all parameters (SNR, EPI, SSIM and MS-SSIM respectively) on the noise level with variance ( $\sigma$ ) 0.005.

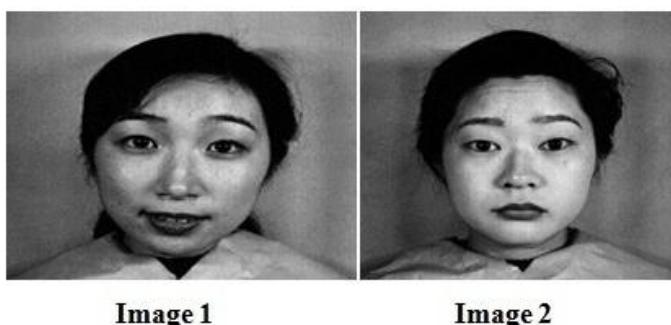


Figure 2: Sample Images of JAFEE database

#### V.CONCLUSIONS:

This paper presents the experimental analysis of various filters for removing the Gaussian noise present in face images. The performance comparison of various filters has been evaluated in terms of quality assessment metrics SNR, EPI, SSIM, and MSSSIM. In case of median filter, the noise has been reduced at the cost of a small

degradation in image quality. The filters disk, average and unsharp show moderate behavior. The qualitative performance of laplacian and log is poorest at all noise levels among all filters under study. From the visual results it has been concluded that wiener filter shows its superiority among all filters discussed in this paper for removing noise from images while preserving its features.

**TABLE 1: SNR, EPI, SSIM AND MS-SSIM VALUES FOR IMAGE 1**

Filter	Variance ( $\sigma$ )	SNR	EPI	SSIM	MS-SSIM
Average	0.005	12.73	0.08	0.64	0.94
	0.01	11.66	0.06	0.5691	0.9194
	0.05	7.74	0.03	0.35	0.80
Median	0.005	13.73	0.21	0.62	0.93
	0.01	12.03	0.15	0.54	0.90
	0.05	6.68	0.06	0.31	0.77
Gaussian	0.005	12.95	0.45	0.60	0.92
	0.01	10.20	0.34	0.47	0.88
	0.05	3.79	0.16	0.21	0.73
Laplacian	0.005	-2.06	-0.34	-0.04	0.05
	0.01	-2.93	-0.25	-0.03	0.08
	0.05	-5.65	0.11	-0.01	0.08
Disk	0.005	8.77	0.04	0.58	0.91
	0.01	8.60	0.03	0.56	0.90
	0.05	7.70	0.01	0.52	0.87
Log	0.005	-3.52	-0.35	-0.06	0.01
	0.01	-4.55	-0.25	-0.04	0.02
	0.05	-6.72	-0.11	-0.01	0.06
Unsharp	0.005	-2.35	0.35	0.13	0.70
	0.01	-3.87	0.26	0.09	0.63
	0.05	-6.16	0.12	0.04	0.46
Weiner	0.005	14.67	0.44	0.95	0.68
	0.01	13.23	0.36	0.94	0.64
	0.05	8.85	0.16	0.44	0.85
Alpha Trimmed Mean Filter	0.005	5.43	0.01	0.99	0.99
	0.01	4.53	-0.03	0.99	0.99
	0.05	2.09	-0.07	0.99	0.99

TABLE 2: SNR, EPI, SSIM AND MS-SSIM VALUES OF FOR IMAGE 2

Filter	Variance ( $\sigma$ )	SNR	EPI	SSIM	MS-SSIM
<b>Average</b>	0.005	13.41	0.09	0.64	0.94
	0.01	12.37	0.07	0.57	0.92
	0.05	11.34	0.07	0.35	0.80
<b>Median</b>	0.005	14.15	0.09	0.64	0.94
	0.01	12.36	0.07	0.57	0.92
	0.05	11.12	0.07	0.35	0.80
<b>Gaussian</b>	0.005	-1.93	-0.33	-0.04	0.09
	0.01	-2.85	-0.24	-0.03	0.06
	0.05	-3.40	-0.20	-0.01	0.06
<b>Laplacian</b>	0.005	-3.31	-0.35	-0.06	0.01
	0.01	-4.32	-0.25	-0.04	0.02
	0.05	-4.97	-0.29	-0.01	-0.01
<b>Disk</b>	0.005	9.47	0.03	0.58	0.91
	0.01	9.14	0.02	0.56	0.90
	0.05	9.33	0.01	0.36	0.80
<b>Log</b>	0.005	13.36	<b>0.45</b>	0.60	0.93
	0.01	10.50	<b>0.37</b>	0.47	0.88
	0.05	9.65	<b>0.34</b>	0.21	0.73
<b>Unsharp</b>	0.005	-1.98	0.35	0.13	0.70
	0.01	-3.58	0.26	0.09	0.62
	0.05	-2.35	0.19	0.04	0.45
<b>Weiner</b>	0.005	<b>15.12</b>	<b>0.44</b>	0.68	0.95
	0.01	<b>13.55</b>	0.36	0.64	0.94
	0.05	<b>12.78</b>	<b>0.34</b>	0.45	0.85
<b>Alpha Trimmed Mean Filter</b>	0.005	5.75	-0.01	<b>0.99</b>	<b>0.99</b>
	0.01	4.87	-0.03	<b>0.99</b>	<b>0.99</b>
	0.05	2.25	0.08	<b>0.99</b>	<b>0.99</b>

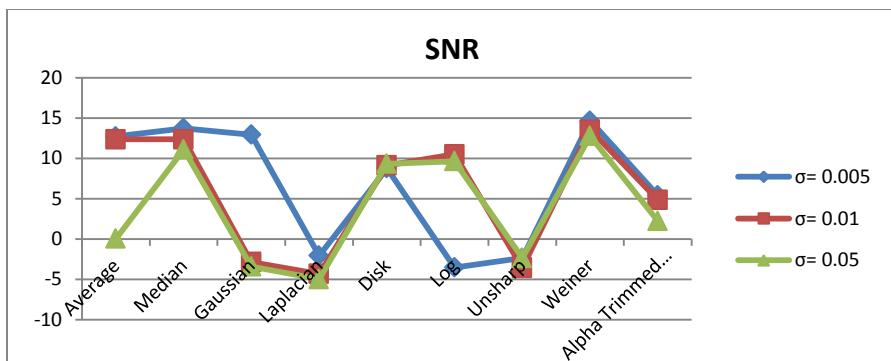


Figure 3: SNR value of all Filters at various noise levels.

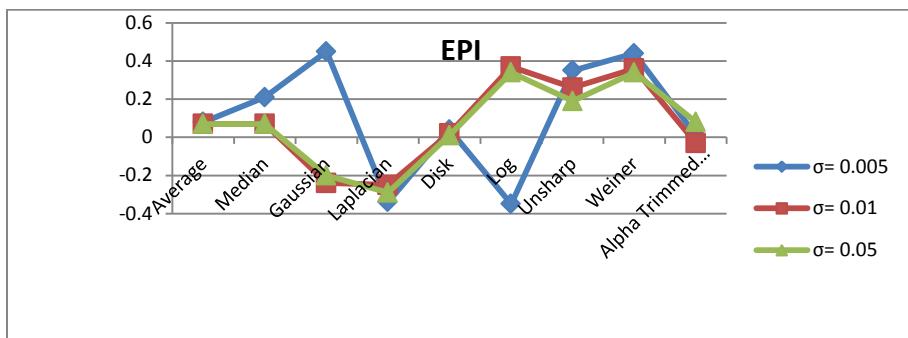


Figure 4: EPI value of all Filters at various noise levels.

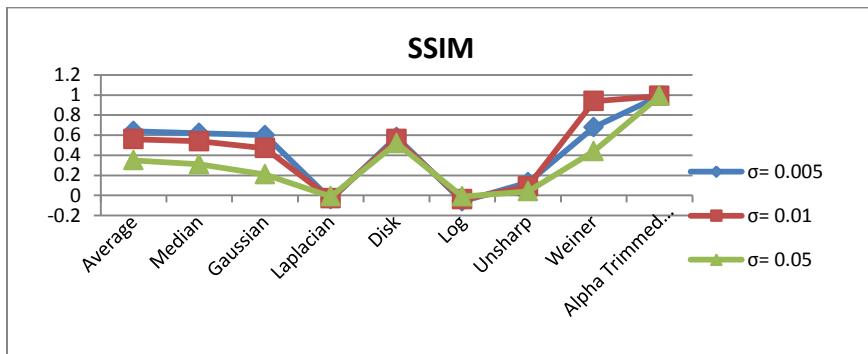


Figure 5: SSIM value of all Filters at various noise levels

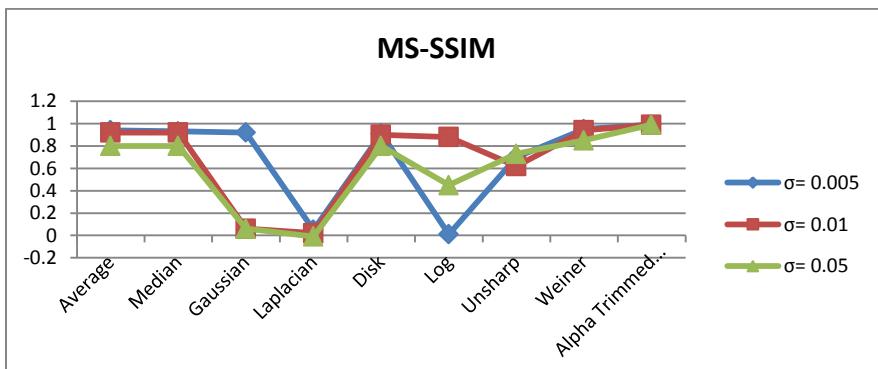


Figure 6: MS-SSIM value of all Filters at various noise levels.

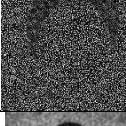
Image 2 at various Noise levels	at $\sigma=0.005$	at $\sigma=0.01$	$\sigma=0.05$
			
Average Filter			
Median			
Gaussian			
Laplacian			
Disk			
Log			
Unsharp Filter			
Weiner			
Alpha Trimmed Mean Filter			

Figure 7: Visual performance of all filters at various noise levels.

REFERENCES:

- [1] Ajay Kumar Boyat and Brijendra Kumar Joshi, "A REVIEW PAPER: NOISE MODELS IN DIGITAL IMAGE PROCESSING", Signal & Image Processing : An International Journal (SIPIJ) Vol.6, No.2, April 2015.
- [2] Ben Hamza, P. Luque, J. Martinez, and R. Roman, "Removing noise and preserving details with relaxed median filters," J. Math. Imag. Vision, vol. 11, no. 2, pp. 161–177, Oct. 1999.
- [3] Chan H, Chung-wa H and Mikolova M., "Salt and Pepper Noise Removal by Median Type Noise Detectors and Detail-Preserving Regularization", IEEE Transactions on Image Processing, 14(10):1479-1485, (2005).
- [4] Eng, H. L., Ma K.. K., "Noise Adaptive Soft-Switching Median Filter,|| IEEE Transactions on Image Processing", 10(2): 242–251, (2001).
- [5] Gonzalez R. C., Woods R. E. "Digital Image Processing," second edition, Prentice Hall, Englewood, Cliffs, NJ, (2002).
- [6] Krystek M., "A Fast Gauss Filtering Algorithm for Roughness Measurements", Precision Engineering 19(2-3):198-200, (1996).
- [7] Kun He, Xin-Cheng Luan, Chun-Hua Li, Ran Liu," Gaussian Noise Removal of Image on the Local Feature", IEEE Computer Society, 2nd International Symposium on Intelligent Information Technology Application, 2008, IEEE, pp 867-871
- [8] R. C. Hardie, K. E. Barner, "Rank conditioned rank selection filters for signal restoration," IEEE Trans. Image Processing, vol. 3, pp.192–206, Mar. 1994.
- [9] R. Yang, L. Yin, M. Gabbouj, J. Astola, and Y. Neuvo, "Optimal weighted median filters under structural constraints," IEEE Trans. Signal Processing, vol. 43, pp. 591–604, Mar. 1995.
- [10] S. Suryanarayana, Dr. B.L. Deekshatulu, Dr. K. Lal Kishore and Y. Rakesh Kumar, "Estimation and Removal of Gaussian Noise in Digital Images", International Journal of Electronics and Communication Engineering, Volume 5, Number 1 (2012), pp. 23-33.
- [11] Taubin G. , "A Signal Processing Approach to Fair Surface Design", Proceedings of SIGGRAPH, 351-358, (1995).
- [12] Tuan-Anh Nguyen, Won-Seon Song, Min-Cheol Hong," Spatially Adaptive Denoising algorithm for a single image corrupted by Gaussian noise", IEEE Transaction on Consumer Electronics, Vol. 56, No. 3, Aug 2010, pp 1610-1615.
- [13] V.R. Vijay Kumar, P.T. Vanathi, P. Kanagasabapathy," Adaptive Window Based efficient Algorithm for removing Gaussian noise in gray scale and color images", International conference on Computational and Multimedia Application", IEEE Computer Society, 2007, pp 319-323
- [14] Yang Jie, Wan Li, and Qu Changqing, "Illumination Processing Recognition of Face Images Based on Improved Retinex Algorithm", JOURNAL OF MULTIMEDIA, VOL. 8, NO. 5, OCTOBER 2013.
- [15] Yiwen Qin, Zongliang Gan, Yaqiong Fan, Xiuchang Zhu, "An Adaptive Image Denoising method for Mixture Gaussian Noise", 2011, IEEE.

# SEGMENTATION OF NATURAL IMAGES USING HSL COLOR SPACE BASED ON K- MEAN CLUSTERING

Gagan Jindal

Student, Master of Technology,  
Dept. of Computer Engineering,  
Punjabi University, Patiala.

Sikander Singh Cheema

Assistant Professor,  
Dept. of Computer Engineering,  
Punjabi University, Patiala

**Abstract** *Image segmentation is one of the important part in image processing. Image segmentation is a process that divides the image into several parts according to its shape, pixel intensity, region formation or by other features. Image segmentation is the process which comes after the image compression and is followed by the various description parameters. In this paper we discuss algorithm to segment natural image using Hsl color space based on the k-mean clustering method. The basic goal of image segmentation is to the convert the given sample image into somewhat more meaningful and understandable, As well as in understanding high level process such as robotics, face recognition, leaf structure study etc.*

**Keyword** K-Mean, Hsl, Rgb, Hue, Saturation

## INTRODUCTION

Segmentation is a basic process which involves converting the image into various segments to have isolate area of interest. Segmentation enable us to have high level of knowledge from given set of image. Image segmentation is basically used to locate objects, boundaries, curve and many more in images. More simply, image segmentation is the process of giving a label to every pixel in an image such that pixels with the same set of properties have same properties.

An RGB color model has three basic colors Red, Green and Blue which mix with each other in various proportion to produce various range of colors. Zero intensity for each component gives the darkest color which means black color, and full intensity of every part gives a sign for white, the *quality* of white color majorly depends on the nature of the primary light sources, but if they are balanced in a fixes manner, the result is a neutral white matching the system's white shade point. When the intensities for all the present parts are the same, the result is a tone of gray, dark or light depending on the intensity. When the intensities components are different, the result is a colorized touch of **hue**,

more or less **saturated** which in turn depends upon the difference of the intensity between strongest and weakest part. of the primary colors employed

## COLOR MAP

Color map is in form of matrix which consists of 0 and 1 to represent the various color bodies. MATLAB draws the objects by assigning data values to each and every colors in the given color map. Color maps length can be according to requirement but it will be three Column wide. Every row defines a specific color using RGB triplet. RGB is a basically row vector whose values tells about the intensity of the Green red and Blue parts in the color. The intensities value should be in the range [0,1]. From the zero value it is estimated that there is no color whereas one represent full intensity color map. color map name defines the color map for the current figure for a built-in color map specified by name. The new color map uses the same number of colors as the current color map. The figure color map cause suffering of all axis in the image unless you define all axis separately. Color map sets the color map for the current image to the color map mentioned by map. Use given syntax if you want build in color map with given no of colors. Build in color map provide a specific range of color available to users which are selected by user randomly.

## COLOR GAMUT

Color gamut is complete subset of colors . It may define as colors which are find in image for a given time. It is basically used in the hue-saturation plane where system can produce large variety of intensity range within its color gamut. When particular colors are not defined in given color model then it is said to be out of gamut. pure red color is example of it. Pure red color can be applied in RGB model whereas it is not available in CMYK model.

## K MEAN CLUSTERING METHOD.

K mean clustering is vector quantization method which partition image into different cluster on the basis of mean of similar pixel set whereas no of cluster are pre defined. In this each cluster have a centriod which represent mean of the cluster. The particular group is obtain is by subtracting the square distance between items called Euclidean distance and corresponding centriod. commonly used initialization method are Forgy and random partition. Forgy method selects k observations from the given pixel set and uses it as initial mean. Forgy method tries to spread initial mean out whereas random partition put them close of the pixel

### Advantage of k means clustering

1. When no of variables are huge then it computational times is much faster than any other provided value of k is small.

2 k means clustering produces tighter cluster than hierarchical clustering even the cluster are globular.

3. k mean clustering use unsupervised learning to sort out clusters.
4. k-means did a good job in case of pre-clustering, which reduce the space into disjoint smaller sub-spaces ,in which place others clustering algorithm can be applied.

## HSL COLOR SPACE

HSL stands for Hue, Saturation, and Lightness. This is one of most accepted cylindrical coordinate system in RGB color model. In each cylinder the angle surrounding the central axis is Hue, the distance from the axis is shown by saturation and the distance along the axis described by lightness. HSL color is simply device transformation of the RGB color model so the physical color they define depend on the amount of Red Green Blue color described by the RGB color space and gamma correction used to represent those primaries. In this model hue component having angular momentum, starting at the red primary at 0, passing through the green primary at 120 and blue at 240, and then wrapping back to red at 360. In each geometry, the central vertical axis comprises the *neutral, achromatic, or gray* colors, ranging from black at lightness 0 or value 0, the bottom, to white at lightness 1 or value 1, the top

## Proposed Methodology

Many image segmentation techniques used hsl color space to segment the natural image into useful constituents. HSL color space is one of most approachable technique used for segmentation as it differs in many forms from other color models. Unlike RGB, HSL separates luma, or the image intensity, from light or the color information. Like if you want to do histogram equalization of a color image then you must want to work on intensity component and leave the color component alone. If we don't do so we will get very strange colors. Also the availability of code for converting RGB to HSL color space is very wide with benefit that this model can separate intensity from color. The basic aim of segmentation here is to segment given natural image into multiple segments based on k-mean clustering. The basic steps followed are.

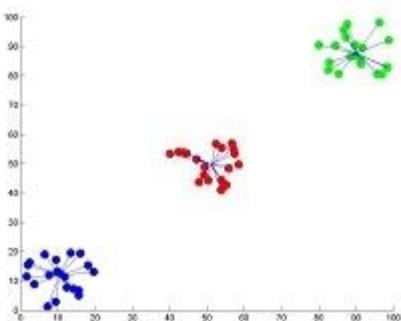
1. First of all the given sample image to segmented is selected.
2. Then the given image is converted to HSL color space using algorithm in matlab.
3. The image is converted from the RGB color space to the HSL which give image representation in form of hue, saturation and intensity.
4. Using k-mean clustering we then segment the image into multiple clusters on basic of different factors.
  - (a) A suitable no cluster to be formed are chosen at runtime.
  - (b) After choosing no of cluster to be formed then we will choose the factors on basis of which segmentation is performed.

- \* In given approach we can segment image on basis of h-s (hue-saturation) components or h-l (hue-lightness) components or h-s-l (hue-saturation-lightness) component.
- \* If we choose hue-saturation component as a standard for segmentation then image is segmented on basis of hue and saturation factors of image without caring about the lightness component.
- \* If we choose hue-lightness component as a standard for the segmentation of the image then image is segmented on the basis of hue and saturation property of image. Hue is always taken as key factor as this is the basic for the natural image.
- \* If we choose Hue-Saturation-Lightness component as a standard for the segmentation of image then image is segmented on the basis of the all these properties.

5. After the cluster formation on basis of the given factors the image is purely segmented. We can see different cluster separately in segmented image.

### **K-mean clustering algorithm**

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function known as squared error function given by:



Graph showing cluster for k mean clustering having n no of pixel and no of clusters c =3.

### Algorithmic steps for k-means clustering

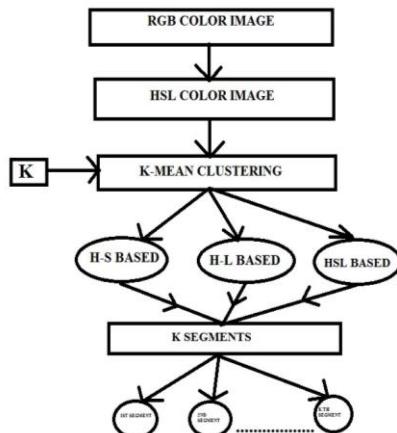
Let  $X = \{x_1, x_2, x_3, \dots, x_n\}$  be the set of data points and  $V = \{v_1, v_2, \dots, v_c\}$  be the set of centers.

- 1) Randomly select 'c' cluster centers.
- 2) Calculate the distance between each data point and cluster centers.
- 3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
- 4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, ' $c_i$ ' represents the number of data points in  $i^{th}$  cluster.

- 5) Recalculate the distance between each data point and new obtained cluster centers.
- 6) If no data point was reassigned then stop, otherwise repeat from step 3)



### **Advantages of using hsl color model**

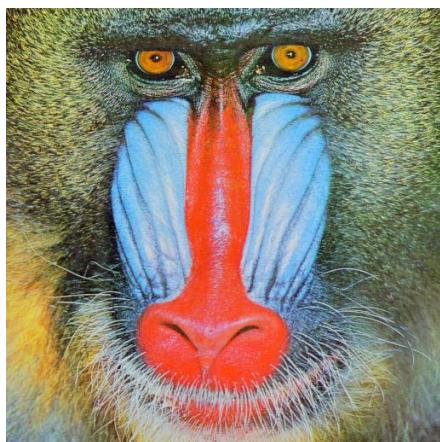
1. This model separates out luma, chroma and color information from the natural image.
2. Code conversion from RGB to HSL is easily available and easy.
3. Image having shadow falling on it could be recognized as separate regions having different characteristics in RGB model whereas HSL model differ only in luminance component of the image whereas Hue component is same for shadow region or without shadow region.
4. HSL model point out colors which are perceived by human whereas RGB give color which is treated by computers.
5. HUE component of model is quite useful as it is mostly same for all parts. Like our hand has many parts palm, back palm, fingers etc. All other components may be different for all these but HUE component is same for all these.

### **Advantages of k-mean clustering**

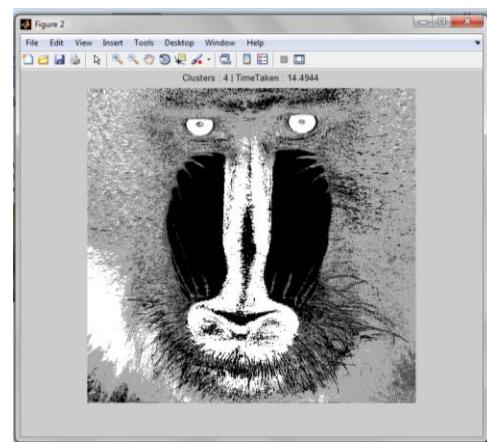
- 1) If variables are large, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls.
- 2) K-Means produce closer clusters than hierarchical clustering, especially if the clusters are globular
- 3) This is one of most efficient algorithm if the value of k is already known.
- 4) Gives best result when data set are distinct or well separated from each other
- 5) Give high performance in terms of computational time when no of pixel are huge.

### **Results**

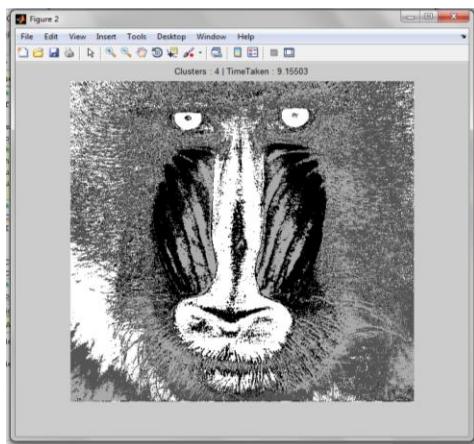
Result obtain are having no of cluster which are defined in predefined manner given by user. By applying K mean clustering algorithm we obtain k no of cluster. The clustering is performed on the basis of the hsl color space



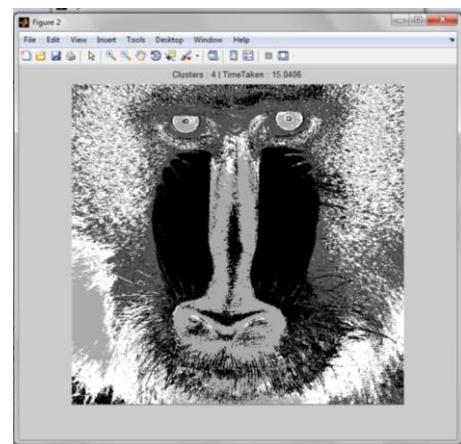
(a) Original Image



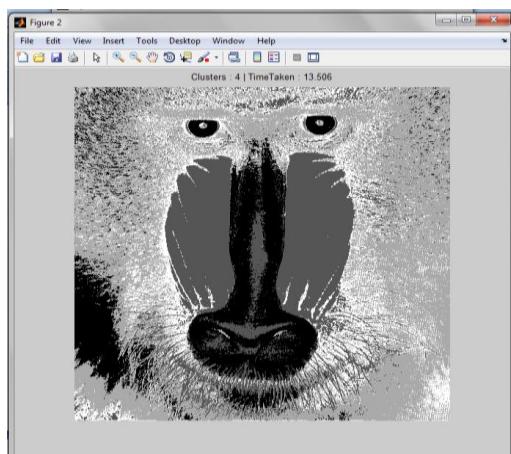
(b) Clustered image with 4 cluster



(c) Clustered Image with H-S color space



(d) clustered Image with H-L color sapce



(e) Clustered Image with H-S-L color space

**Conclusion** in this paper for segmentation of natural image using HSL color space by k mean clustering is proposed. Also number of images is tested using the given method, the result of one image is given and illustrated in this article. Natural images are segmented using RGB and HSL color space. The result obtain is compared using no of factors like execution time. Moreover, if there is a lot of fluctuation in the value of the color information (hue and saturation), pixels with small and large intensities are not considered. The experimental results shows that the proposed algorithms yields segmented gray scale image of perfect accuracy and the required compute time reasonable and also reveal the improved fuzzy c means achieve better segmentation compare to others. The advantages of the new method are the following: (1) it yields regions more homogeneous than those of other methods, (2) it removes noisy spots, and (3) it is less sensitive to noise than other techniques. This technique is a powerful method for noisy image segmentation and works for both single and multiple-feature data with spatial information. For the applications like video conferencing, real-time tracking the proposed algorithm can be used. Our experimental results show that the proposed algorithm can be used in real-time applications.

## REFERENCES

- [1] Yas A. Alsultanny, “Color Image Segmentation to the RGB and HSI Model Based on Region Growing Alrorthim”, International journal of computer applications, Vol. 1 No.-14, 2014, pp. 63-68
- [2] Amit Kumar, Vandana Thakur “Improved Color Image Segmentation Based On RGB And HSI ”, International Journal Of Engineering Devolpment And Research Issues, Vol. 3, Issue 2, No 3, 2015, pp. 969-987.
- [3] Nagasudha D, Madhaveelatha “Telgu Document Image Segmentation Methods”, International Journal Of Research And Application, july-sept,2014, pp. 76-79.
- [4] Ganesan P,Priya Chakarvarty,Shweta Verma, “Segmentation Of Natural Color Images In HIS Color Space Based On FCM Clustering”, ”, International Journal Of Advanced Research In Computer Engineering & Technology,Volume 3 Issue 3,March 2014
- [5] Manpreet Kaur, Chirag Sharma, “Improved Method for Segmentation of Real-time Image of Printed Documents”, International Journal of Soft Computing and Engineering, Volume-4, Issue-2, May 2014, pg. 136-138
- [6] Akira Taguchi, Naoki Nakajima, and Yoshikatsu Hoshi, “Improved HSI Color Space without Gamut Problem”, International Journal of Information Technology and Knowledge Management, Vol. 2, No. 2, , Dec. 2010, pp. 545-548
- [7] K.N. Plataniotis and A.N. Venetsanopoulos, *Color Image Processingand Applications*, Springer, 2000.
- [8] Cheng C., Region Growing Approach to Color segmentation using 3-D Clustering and Relaxation labeling, IEEE Proceedings of Visual Image Signal Processing, pp233-239, 2003.

# Authorization And Multifactor Authentication Using Dynamic Security Feature On Cloud

Preetinder Singh

Student, M.Tech (CE) Department  
UCOE,Punjabi University  
Patiala, Punjab, India  
Preet.dhindsa@hotmail.com

Mr. Gurjot Singh Bhathal

Assistant Professor, M.Tech (CE) Department  
UCOE,Punjabi University  
Patiala, Punjab, India  
gurjot.bhathal@gmail.com

**Abstract—**This paper describes multifactor authentication (more than one security parameters) and authorization resolutions for using online resources. Digital Markets are challenged with allows informal access to online information for authorized users. This paper presents the drawbacks of some present authentication and authorization systems used by industry and proposed new security framework based on multifactor authentication means by including extra security parameters dynamically.

**Keywords**— Multifactor authentication, XACML, Dynamic Security, IDMS (Identity Management Server)

## I. INTRODUCTION

Cloud computing, as emerging computing model of information technology, has been developed very quickly in recent years .The huge spread of Internet resources on the web and fast progress of service providers enabled cloud computing systems to become a big IT service model for distributed network environments. It has established from distributed computing, grown incrementally from conceptual grid computing and has been shaped by various other concepts. Organizations such as e-governments, online education, hospitals and health care are benefiting from this technology by trimming setup costs, incorporating multilateral network effects and increasing performance and storage. Moreover ,it is a field of computer science in which consumer can use resources remotely through web browser .Cloud Computing increases the speed of accessing the services in very much less cost without actually deploy them. It decreases the time from implementing the software to actually deploy it. Cloud Computing users can access resources on demand. Cloud provides the on demand services, virtualization and open source. The Cloud Computing Architecture which contains on-premise and cloud resources, middleware, services, and software components, geo-location, the externally visible properties of those and the relationships between them this is also refers as documentation of a system's cloud computing architecture

Cloud is accessed through the internet and internet is available all over. So anyone can access their information on cloud at any time and from anywhere. In traditional approach all the data of user would be stored at user's end but in cloud whole scenario is different. Data on cloud is stored on virtual servers and user does not know where the virtual servers exist. So the users don't need to be at the location where the data is stored. Cloud provider uses the in-house or external resources for providing services to users.

*Authentication*

Authentication is the procedure of defining whether someone or something is, in fact, who or what it is declared to be. It is any procedure by which a web server validates the identity of a User who wants to use it. Since Access Control is usually based on the identity of the User who requests access to a resource, Authentication is necessary part of effective Security.

Authentication may be provided by using Credentials, each of credentials which is collection of a User ID and Password. Consecutively, Authentication may be provided with many other ways like Smart Cards systems, an Authentication Server or by Public Key Infrastructure.

#### *Authorization*

User authentication is the confirmation of an active human to machine transfer of credentials required for confirmation of a user's authenticity .the term compares with machine authentication, which involves automated processes that do not require user input.

User authentication is achieved in almost all human to computer interactions other than guest and routinely logged in accounts. Authentication authorizes human to machine interactions on both wired and wireless networks to enable access to Internet connected systems and resources.

#### *Multi Factor Authentication*

It is a process of calculating access control in which a web user is only allowed for access after fruitfully presenting some distinct pieces of identity proof to an authentication server.

Two-factor authentication (also known as 2FA) is a technique of approving a user's requested identity by matching a combination of two separate identity components. These identity components may be somewhat that the consumer knows, somewhat that the consumer holds or somewhat that is involved with the consumer. A practical example from our daily life is the drawing of money from a ATM machine. Only the correct arrangement of a bank card (somewhat that the consumer have) and a ATM PIN (personal identification number, somewhat that the consumer knows) or sometimes the user thumb impression allows the transaction to be accepted. 2FA is also failed against modern threats, like ATM skimming, phishing, and malware etc. 2FA is best example of multi-factor authentication

## II. SECURITY CHALAGES ON CLOUD COMPUTING

Privacy and Security is the main issue in cloud Security. To deal with these issues, the cloud provider must build up adequate controls to offer such level of security. The major issue is that the owner of the data has no control on their data processing. There are two types of attacks for cloud data storage those are: Internal Attacks and External Attacks.

#### *Internal Attacks:*

These are attacks which initiated by malicious users or Cloud Service Provider (CSP). They purposefully corrupt the information of user which is inside the cloud by modifying or deleting the data. They can obtain the whole data and they can give the data to that user who is allowed to access that data.

#### *External Attacks :*

These attacks are initiated by unauthorized users those are from the outside of the cloud. Any attacker from outside the cloud who has capability of comprising cloud servers that can access the information of users as long as they are internally consistent i.e. they may modify or delete the information of customer and may disclose the private information of users.

Some of the potential attacks vectors criminal's attackers may attempt include (Neela):

### III. BACKGROUND STUDY

Authentication and authorization is a main part of every secure cloud communication system. It has changed dramatically as users of new-world access technologies look for a way to authenticate, authorize, and start accounting records for billing user time on their networks. Cloud wishes unobtrusive and safe authentication and authorization processes. Many of the online access methods presently exist, even if comprehensive, easy-to-use, virtualized and scalable solutions are still developing. Many third party Access management systems implement authentication and authorization for online resources. Although authentication and authorization are performed at the same time in many schemes, they are actually separate methods. We need Authorization and authentications do not have to be implemented at the same time, or even by the same systems or in the same locations. The first step in achieving privacy and secrecy would require trust development as well as efficient management of user uniqueness and user keys. Keeping this in mind we have taken into account some studies related to trust development and key management. We have divided our writings review into two tracks including the modern authentication and authorization procedures such as OAuth, OpenID and Shibboleth along with the traditional anonymous authentication/authorization and trust development mechanisms.

#### A. Existing Solutions

Firstly IP address validation was main, it but does not provide access to users at remote sites. After IP validation Web proxy service was presented to addition IP address validation by permitting off-campus users to log in to a representation server. After that public key Encryption comes into market.

##### 1) IP Address Validation

Internet protocol (IP) address validation is a comparatively easy access management system. According to this system, resource providers relate the IP address of a user to a list of authorized IP addresses maintained at his side. If there is a successful match, the resource providers allow access. This validation does not need usernames or passwords, building it simple and transparent for customers. The major problem is that it does not provide access to users at remote sites unless individual IP addresses are moved into the database, which is a lengthy process. IP address validation also pretenses a probable threat to the secrecy of the consumer, since all data exchanges are tangled openly to specific IP addresses

##### 2) Web Proxy Service

After IP validation Web proxy service was presented to addition IP address validation by permitting off-campus users to log in to a representation server, whose address would then be accepted and agreed by the resource provider. The main advantage of web proxy service is that it can be merged with an existing IP address validation system. The Web proxy validation also provides excellent privacy for off-campus users.

##### 3) Public Key Encryption Schema

Typical or symmetric encryption requires some kind of shared secret. In the computer world, this is called a key. The problem is, in order to communicate privately using symmetric encryption, we both have to know the same key. We have to communicate this key to each other by some safe means before we can use it. It's a chicken-and-egg situation, we need to send the key privately in order to have a confidential channel. Amongst these protocols, the public key based encryption for secrecy enjoys lesser attention.[8] Suggests construction of an unspecified authentication protocol which is verifiably unspecified. The server encrypts the same challenge along with different random numbers and sends them to the user; the user does decryption and sends the challenge back to the server. The server upon receiving the value of the challenge from the client relates it to the originally sent value of the challenge and then sends back all the random numbers used to encrypt the challenge. If anyone of the random numbers mismatch the authentication process is unsuccessful.

## **B. Industrial Solutions**

After Reviewing traditional authentication and authorization protocols. We have review some Industrial Solutions

### *1) openID*

In OpenID, verification is delegated: server A wants to authenticate user U, but U's credentials (e.g. U's name and password) are sent to another server, B, that A have trusts (at least, trusts for authenticating users). Indeed, server B makes sure that U is indeed U, and then tell A: "ok, that's the genuine U".

### *2) OAuth*

In OAuth, authorization is joined: entity A obtains from entity B an "access right" which A can show to server S to be provided access; B can thus deliver temporary, specific access keys to A without giving them too much power. You can imagine an OAuth server as the key manager in a big hotel; he gives to employees keys which open the doors of the rooms that they are want to enter, but each key is limited (it does not give access to all rooms), furthermore, the keys self-destruct after a few time. To some extent, authorization can be abused into some pseudo-authentication, on the basis that if entity A obtains from B an key through OAuth, and shows it to server S, then server S mayinfer that B authenticated A before granting the key. So some people use OAuth where they should be using OpenID. This schema may or may not be educational but I think this pseudo-authentication is more confusing than anything. OpenID Attach does just that: it misuses OAuth into an authentication protocol. In the hotel analogy: if I encounter a purported employee and that person displays me that he has a key which opens my room, then I assume that this is a true employee, on the basis of that the key master would not have given him a key which opens my room if he was not.

## IV. METHODOLOGY

After complete review of existing protocols and industrial solutions .We design an architecture containing some of essential components to provide security on cloud. Like other traditional systems we have used XACML (eXtensible Access Control Markup Language) based server for authorization and IDMS as key server

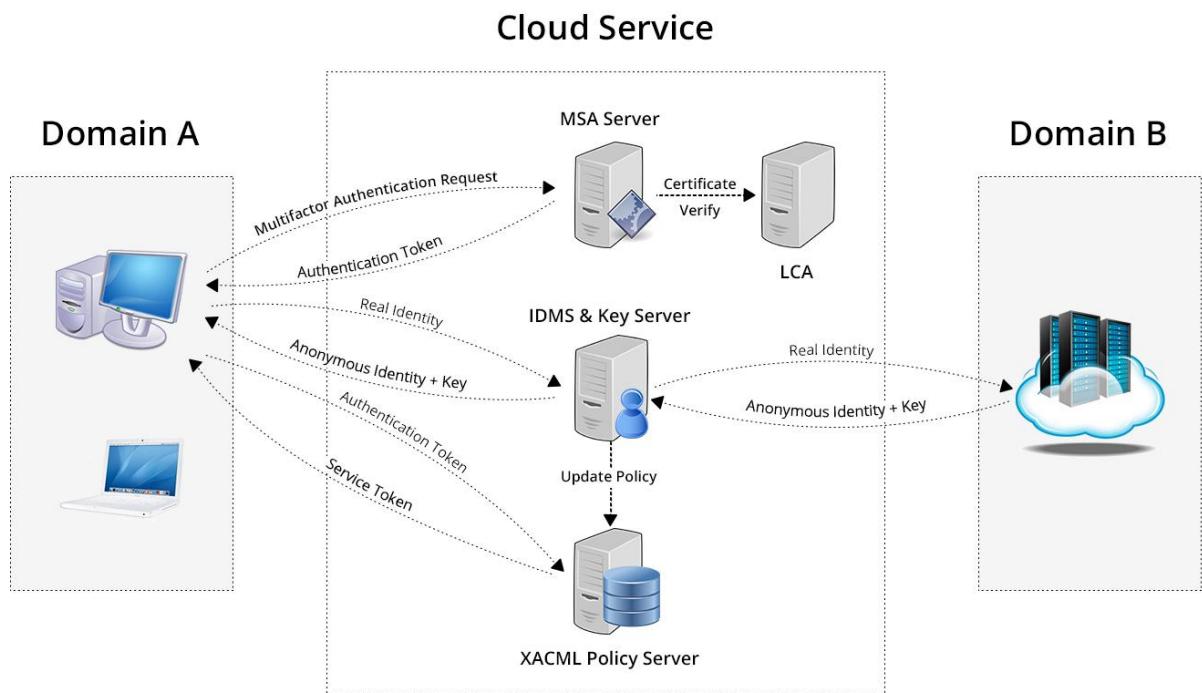


Figure 1. Purposed multifactor authentication framework

#### A. XACML Server{ eXtensible Access Control Markup Language server }

XACML defines the policy language, the request / response scheme, and the architecture .we can store all our registered users and there complete access list on this server.

#### B. IDMS{Identity Management System}

This is a server which stores all the information of identity. In terms of security. In order to grant safe access to sensitive information and resources to all those who need it, It carefully monitor which users are accessing what resources to ensure that they are using the resources that they need in an appropriate manner. It also take care of new users and take responsibility for providing new identity and distribution center take responsibility to distribute all the keys to users.

#### C. SMA Server{Strong Multifactor Authentication server}

An authentication server is an application that enables authentication of an entity that attempts to access a network. Such an object may be a human user or another server. An authentication server can reside in a dedicated computer, an Ethernet switch or a network access server.

Authentication is the process of determining whether someone is actually who it declares itself to be. When a potential subscriber accesses an authentication server, username and password may be the only detecting data required, but strong multifactor authentication server identify according to multiple random security parameters.

#### **D. CA { Certificate Authority }**

The local certificate authority creates and verifies anonymous certificates for the entities wishing to gain access to a assured service. This certificate authority may or may not be verifiable by some other certificate authority in the chain of faith and is dependent on the deployment model. It should be kept in mind that even though the local certificate authority is able to validate the user identity based on the certificate, this identity itself is just a pseudonym and not the real identity of the consumer. The number of anonymous certificate issuing authorities/sub authorities may be further decreased or increased depending on the level of privacy needed.

#### **E. Authorization Server**

The authorization server is responsible for getting and validating the user access request to some particular service. It maintains a list of all the policies associated to the users in the policy engine and updates them when required. If a request is validated successfully, the authorization server issues an access token to the user granting him/her access for a particular amount of time.

#### **F. Web Application**

A web application presented to the user through the web browser is responsible for all the communications of the user with any of above discussed servers. This web application passes on the requests received from the user to the servers as well as the replies from the server back to the user. In short this web application acts as on demand software granting access to any service.

#### **G. Protocols**

Before we start defining the Multifactor authentication and authorization protocol there are four things that must be considered in mind. Every Client is already registered to the IDMS and the necessary identity attributes are stored in the ID Store and every user have multiple security parameters like biometric, passwords etc. All the system components exist on some cloud infrastructure and the client interacts with the web application through the web browser, Mobile browser etc. The authentication and the authorization server share a pre-shared master key in order to check the authenticity of the tokens which are transmitted and the received. The described architecture and the solution are given as a cloud service.

##### *1) Multifactor Authentication Phase*

The Client A begins by selecting a Cloud with which it wants to communicate and so builds an authentication request for the MSA server (Multifactor Secure Authentication Server). Since Client A cannot directly access Client B without authenticating itself to the MSA server first, this request also contains the Client A anonymous certificate [6].

The MSA Server receives the authentication request which is the security parameter selected dynamically, then the verification of the certificate take place with the local certificate authority and then it chooses which security

parameter is selected for this user dynamically and then compare it at IDMS. This is done by sending the ID dynamic to the IDMS here the list of identities is then maintained in its database. The LCA server also sends Certificate for the verification of its validity.

If there is success in result of both of the verification requests the MSA server calculates a challenge for the Client A which is a random number in our case. The SA server also attaches an identifier called the Token ID which is a Session Identifier. The server also stores the generated random numbers and session for future use. If one or both either the verification requests times out or results in a null, the SA server terminates.

Client A stores Session Number on receiving the message and calculates another random number. Client A concatenates received random number with the newly generated random number, the token id and signs the whole with his/her private key

On receiving this message MSA Server uses the Client A Public Key to decrypt the signed part of the message and relates it to the unsigned part as well as the retained value of the random number it sent. The user gets authenticated, if the result is a matched. Having authenticated the Client A, the SA Server generates an ID request. For assigning a new ID, this is a request for the client which will be used as the Anonymous ID in the future communication.

## 2) Authorization Phase

The identities of both the communicating users i.e. Client As anonymous identity and Client B's access token along with the real identity received is sent by Client A from the authentication server to the authorization server. After receiving encrypted access token and identities, the authorization server stores the IDs and decrypts encrypted access token with the symmetric key to get ID. This proves that the access token was indeed meant for Client A. Note that the token is encrypted using the pre-shared symmetric key shared between the authentication and the authorization server [6].

The Authorization Server checks for policies against both the anonymous IDs in its policy engine. It makes a decision based on these policies using the policy combiner algorithm which results in {permit, deny or undefined}.

## V. EVALUATION

From our finding [7][10] there are main some Attacks categories like Reply Attack , Man in Middle Attack ,Sql Injection, phishing , cross site Scripting, cookies tempering , session hijack server hijack. We consider all these attacks in our protocol and many more parameters like how these attacks are taken place and how our proposed solution work against these attacks. Detail is listed in table 1.

Attack Type	Launching Conditions	Defensive Measures
Tampering	Attacker may changes information on local file database and sent to network	Use of session identifiers and sequence number instead of transmitting cookies with user credentials.
Man In Middle Attack	Attacker infiltrate the communication channel to monitor the communication and modify messages	The use of anonymous certificates protects the user identity to be revealed even if the adversary gets hold of a certificate.
Information Discloser Attack	When Attacker gain access to data Path	The use of anonymous certificates protects the user identity to be revealed even if the adversary gets hold of a certificate
SQL Injection	Attacker input validation vulnerabilities to send corrupt commands back to the database.	The use of No-SQL databases such as MongoDB .
Cross-Site Scripting	adversary inserts a malicious script to a dynamic form presented by the user	Input sanitization is the answer to this attack, ensuring that the website only returns the input after validating it and filter meta characters while doing so.
Session Hijacking	Attacker makes use of the improper implementation session ID number and assumes user identity	Using anonymous identities after initial authentication and encrypting traffic using symmetric key prevents such sort of attacks

Table 1 Attacks and There Solutions [10]

## VI. CONCLUSION

In this paper, we have proposed an Anonymous Multifactor authentication and authorization protocol using dynamic multifactor key and public key certificate with Multifactor strong authentication and XACML servers. This Protocol Promises full secrecy and prevent identity Stealing .The protocol has been designed such that same security parameter is never used again it is generated dynamically , No one Knows which would be the next security parameter, it should be biometric, thumb impression etc , So the whole process transparent to user

Our framework is more flexible security levels will be increased by and more security parameters and more certification Authorities. As Work we plan to deploy our protocol in scenario online hotels Management .where information of customers is important and it is highly effective to install security hardware for multifactor authentication.

## REFERENCES

- [1] R. Velumadhava Raoa,\* , K. Selvamanib,, “Data Security Challenges and Its Solutions in Cloud Computing” ICCC-2015 .
- [2] Dimitrios Zisis \*, Dimitrios Lekkas, “Addressing cloud computing security issues ” Journal of Emerging Trends in Computing and Information Sciences, 2012.
- [3] Ahmed E. Youssef and Manal Alageel, “A Framework for Secure Cloud Computing” IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012.
- [4] Ayesha Malik, Muhammad Mohsin Nazir “Security Framework for Cloud Computing Environment: A Review” Journal of Emerging Trends in Computing and Information Sciences, 2012.
- [5] Maneesha Sharma, Himani Bansal, Amit Kumar Sharma,“ Cloud Computing: Different Approach & Security Challenge” , Journal of Emerging Trends in Computing and Information Sciences, 2012.
- [6] Umer Khalida, Abdul Ghafoor, Misbah Irum, Muhammad Awais Shibli “Cloud based Secure and Privacy Enhanced Authentication & Authorization Protocol”, 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013
- [7] Pankaj Arora, Rubal Chaudhry Wadhawan, Er. Satinder Pal Ahuja. “Cloud Computing Security Issues in Infrastructure as a Service” International Journal of Advanced Research in Computer Science and Software Engineering 2012.
- [8] <http://security.stackexchange.com>.
- [9] M. Afzal, M. Hussain, M. Ahmad, Z. Anwar, Trusted framework for health information exchange, in: Frontiers of Information Technology (FIT), 2011, IEEE, 2011, pp. 308–313.
- [10] N. Zhang, Q. Shi, M. Merabti, Anonymous public-key certificates for anonymous and fair document exchange, IEE Proceedings- Communications 147 (6) (2000) 345–350.
- [11] OpenID, It's easy to begin accepting openid on your website (2013). URL <http://openid.net/add-openid/> S. Cantor, Understanding shibboleth (2010). URL <https://wiki.shibboleth.net/confluence/display/SHIB/UnderstandingShibboleth> .
- [12] E. Hammer-Lahav, Beginners guide to oauth (2010). URL <http://hueniverse.com/oauth/guide/intro/>
- [13] A. G. Abbasi, S. Muftic, Cryptonet: security management protocols, in: Proceedings of the 9th WSEAS international conference on Data networks, communications, computers, World Scientific and Engineering Academy and Society (WSEAS), 2010, pp. 15–20.
- [14] Imperva, The top-5 identity theft attacks. URL <http://www.imperva.com/docs/WP Top5 Online Identity Thefts.pdf>
- [15] <https://cloudsecurityalliance.org/>
- [16] <http://www.infoworld.com/article/3041078/security/the-dirty-dozen-12-cloud-security-threats.html>

# Development of Disease Detection System using Leaf Image Analysis

Ankurdeep Kaur	Dr. Manpreet Singh	Er. Priyanka Arora
Research Scholar	Professor	Associate Professor
CSE Department	IT Department	CSE Department
GNDEC, Ludhiana	GNDEC, Ludhiana	GNDEC, Ludhiana
ankurdeepkaur@yahoo.com	mpreet78@gmail.com	arorapriyanka29@gmail.com

**Abstract -** In an agricultural country like India, farmers have a wide range of diversity to select the appropriate fruit or vegetable for cultivation. However, then cultivation of crops for optimum yield and quality produce needs the aid of technological support. The most challenging task for a farmer is to differentiate between a diseased and non-diseased plant and to cure the diseased plant well in time. Thus, there are different techniques which can identify the diseased leaf on the basis of parameters like color, shape and texture. Pixel-to-pixel and similarity measure methods are used to distinguish between a healthy and an unhealthy leaf. These techniques when applied extensively to agricultural science proves to be significantly effective in the field of plant protection. The comparison of these techniques has been done further in the paper, which accounts for disease identification.

## 1. INTRODUCTION

Agriculture has an indispensable role in the socio-economic fabric of India. About 70% of the Indian population depends on agriculture. After China, India is the second largest producer of fruits and vegetables in the world with an annual production of about 94 million tones. Because of the short life shelf of the crops, about 30-35% of the crops are damaged during harvest, storage, transportation, packaging and distribution. Hence, there is an urgent need for the maintenance of the nutritive value of the processed food products as only 2% of these crops are processed into value-added products.

Research in agriculture is designed in order to increase the efficiency and quality of agricultural product. In the field of agriculture, applications of expert system or a computer system are mainly found in the area of disease diagnosis. The implementation of disease's detection application updates the agricultural sector. Appropriate treatment advices are made by disease management.

Diseases in plants lead to major losses in production and economy in the agricultural industry worldwide. A disease can be defined as any impairment of normal physiological function of plants which produces characteristic symptoms and the branch of science which provides skills in order to diagnose and control the disease is called plant pathology.

The four major disease causing agents are fungi, bacteria, viruses and nematodes. It is estimated that pathogens, weeds and animals constitute about 40% loss of the global crop production. About 26% of the yield loss is estimated in India due to plant diseases. Plant diseases lead to period outbreak of diseases that cause death and famine on large scale. Not only to plants, these diseases also affect human health. As diseases in plants are inevitable, thus assessment of health and detection of diseases is necessary for the sustainance of agriculture. Hence, detection of plant disease plays a significant role in the prevention of serious outbreak.

Farmers, no doubt have enough knowledge and experience about the crops and its diseases but when it comes to large scale cultivation of crops and its protection from diseases, there is a need of an automatic system that detects the disease and provides suitable management techniques. Such automatic systems that were built and their efficiency will be discussed further.

## 2. LITERATURE REVIEW

A proliferation of literature is available in the plant leaf disease detection. Some of the key contributions are highlighted. Tables show the comparison between the reviewed papers.

### *2.1 COMPARISON OF DETECTION TECHNIQUES*

P. Revathi et al. presented a paper, "Classification of Cotton Leaf Spot Diseases Using Image Processing Edge Detection Techniques" in which the work proposes strategies for the categorisation of diseases using HPCCDD proposed algorithm. The following steps are considered- 1. The captured image is enhanced. 2. Color Image Segmentation is then processed in order to get the disease. 3. Filters like Sobel and Canny are used to identify the edges; which further helps in classification and identification of disease spots. 4. Suitable pesticide knowledge is provided to the farmer to ensure that precautions are taken timely.[1]

Al Bashish et al. proposed a work in their paper, "Detection and classification of leaf diseases using K-means-based segmentation and neural networks based classification" in which the input of a grape leaf was taken whose background was complex. Green pixels were masked by thresholding and the use of antistropic diffusion was made to remove noise. K-means clustering algorithm was used for segmentation which helps in the detection of the diseased portion. Best results could be observed when for the purpose of classification, Feed Forward Back Propagation Neural Network was trained.[2]

Rakesh Kundal et al. in their paper, "Machine learning technique in disease forecasting: a case study on rice blast prediction" presented an approach which was based on support vector machine s for the development of weather based prediction models of plant diseases. Comparison of the performance of conventional multiple regression, artificially derived neural network (generalised regression and back propagation neural network) and support vector machine was done.[12]

#### *2.1.1 COMPARITIVE STUDY OF IDENTIFICATION TECHNIQUES*

Bed Prakash et al. in their survey paper, "A Survey on identification techniques" compared different identification techniques like Genetic Algorithm, Probabilistic Neural Network, Principal Component Analysis and Back Propagation Neural Network. The survey provides an overview of these techniques along with their merits and demerits.[3]

TABLE 1  
 COMPARITATIVE STUDY OF IDENTIFICATION TECHNIQUES

Sr. No.	Identification Technique	Merits	Demerits
1.	Genetic Algorithm	Can handle complex non-differentiable, large and multimodel spaces  For a complex problem space, it is an efficient search method.	Complications are involved in representing output or training data.  To find some optimal, it is not the most efficient method.
2.	Probabilistic Neural Network	It is adaptive to frequently changing data.  It is tolerant to noisy input.	Consumes more training  Network structure has large complexity.
3.	Principal Component Analysis	It chooses weights depending upon the frequency in the frequency domain.  It is used for variable reductions.	It cannot perform linear separation of classes.  The largest variances do not correspond to meaningful axes.
4.	Feed Forward Back Propagation Neural Network	This model is really easy to understand.  It can be implemented easily as a software simulation.	This method is complex.  This method is time consuming.

Sanjeev S Sannaki et al. in their paper, "Diagnosis and Classification of Grape Leaf Diseases using Neural Networks" presents an approach to grade the disease on leaves of plants, automatically. The system finally contributes to the Precision Agriculture as it inculcates Information and Communication Technology in agriculture. Detection and grading of disease with naked eyes is not a feasible method. Hence, the research proposes an innovative technique by using Fuzzy Logic to grade the disease spread on plant leaves.[4]

Kholis Majid et al. presented a paper, "I-PEDIA : Mobile Application for paddy Disease Identification using Fuzzy Entropy and Probabilistic Neural Network" in which the research developed a mobile application that runs on an Android system. Extraction of paddy diseases was done from digital paddy leaf images by the use of fuzzy entropy and the classification of diseases was done by using PNN. Cross validation can be used for the assessment of the results of a statistical analysis and how they will generalize to an independent set.[5]

Sandeep B Patil et al. in their paper, "An Improved Leaf Disease Detection using collection of Features and SVM classifiers" propose an algorithm in order to preprocess, segment and extract some information from the image. K-means algorithm is used for segmentation so as to achieve various clusters. The extracted features like color and shape from the affected regions are further sent to the SVM classifier.[6]

Percentage accuracy of various techniques can be calculated by the given formula-

$$\text{Percentage Accuracy} = \frac{\text{Number of samples recognised correctly}}{\text{Total number of samples}} \times 100\%$$

TABLE 2  
 COMPARISON OF DETECTION TECHNIQUES

Sr. No.	Author	Detection Technique	Result
1.	P. Revathi et al	Texture Statistics Computation	Less than 92%
2.	Sanjeev S Sannaki et al	Neural Network	For drowny and powdery affected region, its accuracy is 100%.
3	Kholis Majid et al	Fuzzy Logic and PNN	91.46%
4.	Rakesh Kundal et al	SVM Method	97.2%
5.	Sandeep B. Patil et al	Modified SVM Method	99.2%

#### *2.2 CLASSIFICATION ACCURACY OF DIFFERENT IMAGE PROCESSING TECHNIQUES*

S. Arivazhagan et al. in their paper, "Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features" proposes a software solution for the detection of plant diseases and their classification. First of all, a color transformation structure for the RGB image which acts as an input is created. Further, the green pixels are masked and then its removal is done using specific threshold values. Then comes the segmentation process. Further, the texture statistics are computed for segments and the extracted features are sent through the classifier. The classification accuracy gained by Co-occurrence matrix was 86.77% whereas that by Minimum Distance Criterion along with SVM classifier was found out to be 94.74%.[7]

H. Anand et al. presented a paper, "Applying image processing technique to detect plant diseases" in which the research proposes the use of image processing techniques and artificial neural network to help farmers classify diseases rather than depending on pure naked eye observation. The captured images are filtered and segmented by the use of Gabor filter. By the result of segmentation, texture and color features are extracted and the artificial neural network is trained for the purpose of choosing the feature values that distinguish between the healthy and unhealthy parts of a leaf. It was found that the accuracy of the work was 91%.[8]

H. Al-Hiary presented a paper, "Fast and Accurate Detection and Classification of Plant Diseases" that aims at proposing two important steps after the segmentation phase. The first step identifies the green colored pixels. Then, the pixels are masked on the basis of the specific threshold values which have been computed using the OTSU's method. Further, the mostly green pixels are masked. The second step is the removal of pixels on the boundary of the infected object and the pixels with zeros red, green and blue values. The results show that the algorithm can detect the diseases and classify them with the accuracy of about 83% to 94%.[9]

Qing Yao et al in their paper, "Application of Support Vector Machine for Detecting Rice Diseases using shape and color texture features" presented an application of image processing techniques using support vector machine for the detection of rice diseases. The disease spots of rice were segmented and then the shape and texture features were extracted. Finally, in order to classify the disease, methods such as K-nearest neighbor, neural network, support vector machine etc. are used. The research found that the efficiency of the method using SVM classifier was 97.2%.[10]

V. Khanaa et al. proposed a method for detection of weed, a normal leaf and a diseased leaf, and after the detection the various remedial measures are discussed. The algorithm uses leaf comparison method by dividing the leaf into various regions and then finding the Euclidean distance between the sample and the leaf stored in the database. The similarity distance measure is obtained using feature vectors for better performance to find

decision of leaf recognition. The overall accuracy reached upto 92%. [11]

TABLE 3  
CLASSIFICATION ACCURACY OF DIFFERENT IMAGE PROCESSING TECHNIQUES

Sr. No.	Author	Image Processing Techniques Used	Accuracy
1.	Kholis Majid	Feature extraction by the use of Fuzzy Entropy  PNN as paddy disease classifier	91.46%
2.	S. Arivazhagan	Co-occurrence matrix  Minimum Distance Criterion with SVM Classifier	86.77% (Without MDC)  94.74%
3	Qing Yao	SVM Classifier	97.2%
4.	H. Anand	Gabor filter for feature extraction  ANN based classifier	91%
5.	Al. Hiary	Color Co-occurrence  K-means Clustering  Neural Network	94%

### 3. CONCLUSION

The analysis of table 1 depict that PNN is accurate and much faster than any other identification technique and is insensitive to outliers. Comparison of methods in table 2 shows that performance of SVM classifier is better than others. If more features are added with the SVM classifier, the accuracy of detection is enhanced. The analysis of table 3 depicts that image processing techniques play a crucial role in plants pathology. Various feature extraction techniques, segmentation techniques and classifiers have been proposed in order to improve the identification and classification with respect to their speed and accuracy.

### 4. FUTURE SCOPE

There is a scope for researchers to develop a novel hybrid algorithm by the use of various image processing techniques so as to improvise the performance of the system.

REFERENCES

- [1] P. Revathi and M. Hemalatha, "Classification of cotton leaf spot diseases using image processing edge detection techniques," *2012 Int. Conf. Emerg. Trends Sci. Eng. Technol.*, pp. 169–173, 2012.
- [2] D. Al Bashish, M. Braik, and S. Bani-Ahmad, "Detection and classification of leaf diseases using K-means-based segmentation and neural-networks-based classification," *Inf. Technol. J.*, vol. 10, no. 2, pp. 267–275, 2011.
- [3] B. Prakash and A. Yerpude, "A Survey on Plant Leaf Disease Identification," vol. 5, no. 3, pp. 313–317, 2015.
- [4] S. S. Sannakki, V. S. Rajpurohit, V. B. Nargund, and P. Kulkarni, "Diseases using Neural Networks ", pp. 3–7, 2013.
- [5] K. Majid, Y. Herdiyeni, and A. Rauf, "I-PEDIA: Mobile application for paddy disease identification using fuzzy entropy and probabilistic neural network," *2013 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2013*, no. c, pp. 403–406, 2013.
- [6] S. B. Patil and S. K. Sao, "An Improved Leaf Disease Detection Using Collection Of Features And SVM Classifiers," vol. 3, no. Vi, pp. 539–544, 2015.
- [7] S. Arivazhagan, R. N. Shebiah, S. Ananthi, and S. Vishnu Varthini, "Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features," *Agric. Eng. Int. CIGR J.*, vol. 15, no. 1, pp. 211–217, 2013.
- [8] A. H. Kulkarni and A. P. R. K, "Applying image processing technique to detect plant diseases," vol. 2, no. 5, pp. 3661–3664, 2012.
- [9] H. Al Hiary, S. Bani Ahmad, M. Reyalat, M. Braik, and Z. ALRahamneh, "Fast and Accurate Detection and Classification of Plant Diseases," *Int. J. Comput. Appl.*, vol. 17, no. 1, pp. 31–38, 2011.
- [10] Q. Y. Q. Yao, Z. G. Z. Guan, Y. Z. Y. Zhou, J. T. J. Tang, Y. H. Y. Hu, and B. Y. B. Yang, "Application of Support Vector Machine for Detecting Rice Diseases Using Shape and Color Texture Features," *2009 Int. Conf. Eng. Comput.*, pp. 79–83, 2009.
- [11] V. Khanaa and K. P. Thooyamani, "An Efficient Weed and Pest Detection System," *Indian J. Sci. Technol.*, vol. 8, no. 32, pp. 2–8, 2015.
- [12] Rakesh Kaundal, Amar S Kapoor and Gajendra PS Raghava "Machine learning technique in disease forecasting: a case study on rice blast prediction," *BMC Bioinformatics*, 2006.

## LEACH DUTY – CYCLE to PROLONGING THE NETWORK LIFETIME IN WIRELESS SENSOR NETWORK

*Meenakshi*

*Assistant Professor & Head in Computer Science*

SBSS Memorial Khalsa (Girls) College, Guru Ka Khuh, Munne. Nurpur Bedi (Ropar).

E-mail: manuom1574@gmail.com

Mobile No. 9646669707

**Abstract-** In recent era of advance technology, in Communications & Computations have enabled the development of low-cost, low-power, small in size & multifunctional sensor nodes in a Wireless Sensor Network. The lifetime of an energy constrained sensor is determined by how fast the sensor consumes energy. A node in the network is no longer useful when its battery dies. The wireless sensor networks consists of many battery powered sensor nodes that monitor their physical surroundings and send the resulting data to a sink node. Much research has been done in recent years, investigating different aspects like low power protocols, Network establishments, routing protocols & coverage problems of Wireless Sensor Networks. The design of Energy – Efficient protocol i.e. Leach Duty Cycle is key for prolonging the lifetime of a wireless sensor network. Here it, routing protocols play an important role to forward the information from sensor nodes to Base Station regularly or on demanded in wireless sensor network. There are three types of routing, classified as flat, location based and hierarchical routing. A clustering approach is a hierarchical routing technique. In wireless sensor networks, clustering is effectively used for many applications, including environment monitoring.

The clustering associated with data aggregation improves network performance by decreasing the amount of data to be delivered and the number of hops from sensors to the Base Station. LEACH is the most famous clustering protocol that resolves the energy unbalancing problem among nodes. Here, nodes are classified into two groups: Cluster Heads and Sensor Nodes. The main idea of LEACH is to reform clusters once every period of time, called a round, in order to rotate the role of the Cluster Head among members in a cluster. In LEACH, the repetitive election of Cluster Heads based on the changed probability of the total number of live nodes decreased the total network lifetime and a great decrease causes unbalanced energy consumption between nodes in the entire network. To maximizing the lifetime of sensor nodes, it is preferable to distribute the energy dissipated throughout the Wireless Sensor Network in order to maximize overall performance. In this paper, I proposed Leach Duty-Cycle to save energy of nodes which recently performed as Cluster Heads in the last round. To minimize power consumed during idle listening, Cluster Head nodes, which can be considered redundant, can be put to sleep, until it is time to transmit data. Therefore, the energy of such nodes & the energy of the network are conserved. During the sleep mode, when nodes go to sleep, it cannot affect the working of network. The simulation results showing in MATLAB. So, this protocol as a new energy conservation paradigm with the prolonging the network lifetime.

**Keywords:** Wireless Sensor Networks, Energy Efficiency, Clustering Hierarchy, LEACH Protocol, Energy Consumption, Cluster – Head, Base Station, Duty – Cycle, Network Lifetime.

## I. INTRODUCTION

Wireless sensor networks have gained a world-wide attention in recent years due to the advances made in wireless communication, information technologies and electronics field [1]. The concept of wireless sensor network is based on Sensing + CPU + Radio = Thousands of potential applications [2]. A collection of mobile or static nodes are able to communicate with each other for transferring data more efficiently and autonomously can be defined as wireless sensor network[3]. A lot of applications of wireless sensor network can be found in different field such as events, battlefield surveillance, recognition security, drug identification and automatic security etc [4]. Ongoing research on wireless network is very active at present including numerous workshops and conferences arranged each year [5].

In wireless sensor network, large number of tiny, battery powered sensor nodes having limited on – board storage, processing, and radio capabilities. These sensor nodes sense the information and transmit it to the base station [6]. Base station analyzes the received data and computation is performed, which gives the human understandable result. The Base Station is having unlimited energy power. So, it should implement the algorithm and protocols by which it can enhance the life time of the sensor node as well as save the energy power.

In the current body of research done in the area of wireless sensor networks, A key issue in wireless sensor network is to transmit data by increasing the lifetime of the network various routing protocols are used. There is not any protocol that overcomes the problem of excessive energy drain during the last round for Cluster Head. I feel that if I am using the cluster – based routing protocol, it reduce the network traffic toward the Base Station. Clustering techniques used in Hierarchical Routing Protocol & Energy Efficient Leach Protocol based on Duty – Cycle is the best technique with energy efficient, because it selects CHs randomly, which may result in faster death of some nodes & by using this duty – cycle leach, sleep – awake pattern I also improve the performance of the protocol when the sensor nodes (CHs last round) have the problem of the shortage of energy in the current round. So, in my research paper it not only maintains the required energy, it also reduces the global communication & stressed on local comparisons.

## II. RELATED WORK

In this section, we provide a brief overview of some related research work.

Balakrishnan, Chandrakanan, and Heinzelman [7] proposed a An Application –Specific Protocol Architecture for Wireless Micro sensor Networks. In this paper, the author develops and analyzes Low-energy adaptive clustering hierarchy, which is protocol architecture for micro sensor networks. Leach includes a distributed cluster formation technique that enables self configuration of large number of nodes and rotating the cluster head position to evenly distribute the energy load among all nodes.

LEACH [8] protocols highly affect the performance of wireless sensor networks by an even distribution of energy load and decreasing their energy consumption and prolonging their lifetime. Thus, designing energy efficient protocols is important for prolonging the lifetime of WSNs.

Heinzelman et.al [9] had proposed the first clustering protocol LEACH, based on single hop communication model. This protocol resulted in improved energy efficiency of the network. But this protocol had many disadvantages too, like Selection of cluster head was done by random strategy, so if any low energy node was elected as cluster head then it would be overloaded, thus decreasing the network lifetime.

Energy – Efficient Adaptive Protocol for Clustered Wireless Sensor Networks (EEAP)[10] is used to increase the lifetime of the sensor networks by balancing the energy consumption of the nodes. EEAP makes the high residual energy node to become a cluster – head. The elector nodes are used to collect the energy information of the neighbour sensor nodes and select the cluster – heads and increase the energy efficiency.

Xiang – Yang Li et al., prove that the energy consumption by the authors scheduling for homogeneous network is at most twice of the optimum and the time span of this scheduling is at most a constant times of the optimum.

### III. LEACH PROTOCOL

In wireless sensor network, LEACH (Low – Energy Adaptive Clustering Hierarchy)[9] is a cluster based routing protocol and one of the hierarchical based routing protocols. Hierarchical based routing is to efficiently maintain the energy consumption of sensor nodes and communication between a number of nodes within a particular time and by performing data aggregation. Leach is a clustering protocol where cluster heads are randomly rotated to balance energy of network. The principle of LEACH is how to determine cluster head. The cluster head accepts data from other sensors within the same cluster, aggregate data and finally sends data to the Base Station (BS). Here all cluster heads are directly communicate to the Bs. Since cluster heads are randomly choosing in LEACH algorithm so it has some probability to form a low – energy normal node as a cluster head. As a result, this cluster in not able to transfer data to base station for long time. Therefore, network performance and lifetime will decrease. This problem was solved by Energy Efficient leach protocol based on duty – cycle in Wireless sensor network. Duty – Cycle is set a fraction of time to become the nodes is on active/sleep/awake state. By using sleep awake pattern we also improve the performance of the protocol when the sensor nodes (which perform as CHs in the last round) have the problem of the shortage of energy in the current round. To maintain their energy required to act as active/awake sensor nodes, such nodes are sent to sleep mode for a prescribed TDMA schedule. When their turn comes to transmit data they will automatically awake to act as active node in the network. This new approach can achieve energy efficiency, reduces energy consumption and increasing the number of dead nodes in every round than existing algorithms.

In proposed LEACH [11] has two phase's setup phase and steady phase same like LEACH protocol. In this protocol, to save energy of nodes this recently performed as CHs last round. This is an effort to ease excessive energy drain suffered by those nodes during last round as CHs. The main concept of this protocol is cluster heads opportunistically transmit the data to save its energy.

#### A. Set-up Phase

During the setup phase, a predetermined fraction of nodes,  $p$ , elect themselves as CHs as follows. A sensor node chooses a random number, between 0 and 1. If this random number is less than a threshold value,  $T(n)$ , the node becomes a cluster-head for the current round. The threshold value is calculated based on an equation that incorporates the desired percentage to become a cluster-head in the current round from the set of nodes that have not been selected as a cluster-head in the last  $(1/p)$  rounds. The threshold value is given by:

$$T(n) = \begin{cases} \frac{P}{1 - p * (r \bmod 1 / p)} & \text{if } n \in G \\ 0 & \text{otherwise} \end{cases}$$

- Where,
- G - It is the set of nodes that are involved in the CH election.
  - T (n) - a threshold value
  - p- Predetermined fraction of nodes
  - r- Current round

Each elected CH broadcasts an advertisement message to the rest of the nodes in the network that they are the new cluster-heads. All the non-cluster head nodes, after receiving this advertisement, decide on the cluster to which they want to belong to. This decision is taken based on the signal strength of the advertisement. The non-cluster-head nodes inform the appropriate cluster-heads that they will be a member of the cluster. After receiving all the messages from the nodes that would like to be included in the cluster and based on the number of nodes in the cluster, the cluster-head node creates a TDMA (i.e., Time Division Multiple Access) schedule and assigns each node a time slot when it can transmit. This schedule is broadcast to all the nodes in the cluster.

#### *B. Steady-State Phase*

In steady state phase, data transmission takes place. Nodes send their data to the cluster heads at most once per frame during their allocated time slot. The cluster head must keep its receiver on to receive the data from cluster members. Once the cluster-head has all the data from the nodes in its cluster, the cluster-head node aggregates the data and then transmits to the base station. In transmitting data, the nodes as cluster heads suffered from excessive energy drain during last round. So, these nodes determine its TDMA schedule for data transmission and goes to sleep mode until it is time to transmit data. It means these nodes will awake when they are required to transmit data in the prescribed TDMA schedule. By applying the above duty cycle, which reduce the global communication and stressed for local compression and also reduce the energy consumption.

### **IV. PERFORMANCE EVALUATION BY SIMULATION**

#### *A. Simulation*

To evaluate the performance of this proposed protocol [12], it implemented in MATLAB. Its Goals in conducting the simulation are as follows:

- Study the effect of the energy balanced & data delivery id the efficient clustering & CH selection.
- Compare the performance of the LEACH and Efficient Leach Protocol based on Duty – Cycle in Wireless Sensor Network on the basis of energy dissipation and the longevity of the network.
- This scheme improves the lifetime of the network about 10% than the existing LEACH protocol.

The simulation has been performed on a network of 100 nodes and a fixed base station. The nodes are placed randomly in the network. All the nodes start with an initial energy of 0.5J. Cluster formation is done as in the leach protocol. There is radio energy model used to modify to include the energy for transmitter & Receiver dissipates the energy. Here, to transmit an l-bit message to distance d, the radio expends:

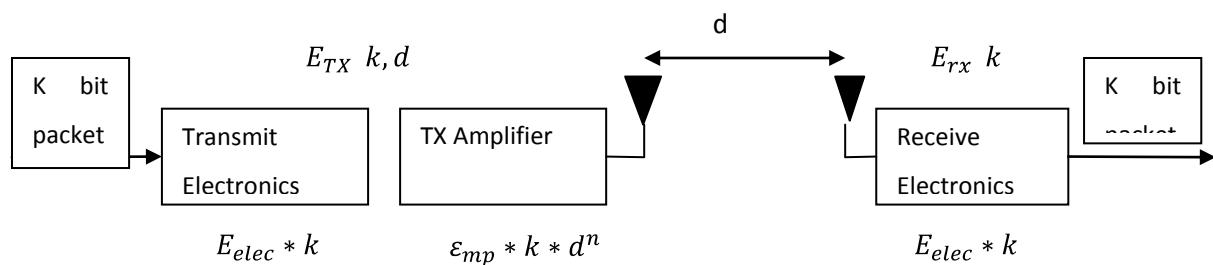
$$E_{TX} \ l, d = E_{TX} \ l + E_{TX-amp} \ l, d$$

$$E_{TX} \ l, d = \begin{cases} lE_{elec} + l\varepsilon_{fs-amp} d^2, & d < d_0 \\ lE_{elec} + l\varepsilon_{mp} d^4, & d \geq d_0 \end{cases}$$

Where  $d_0$  is threshold and is given by:

$$d_0 = \frac{\varepsilon_{fs}}{\varepsilon_{mp}}$$

This radio energy dissipation model is shown in Figure-1:



**Fig - 1** Radio Energy Dissipation Models

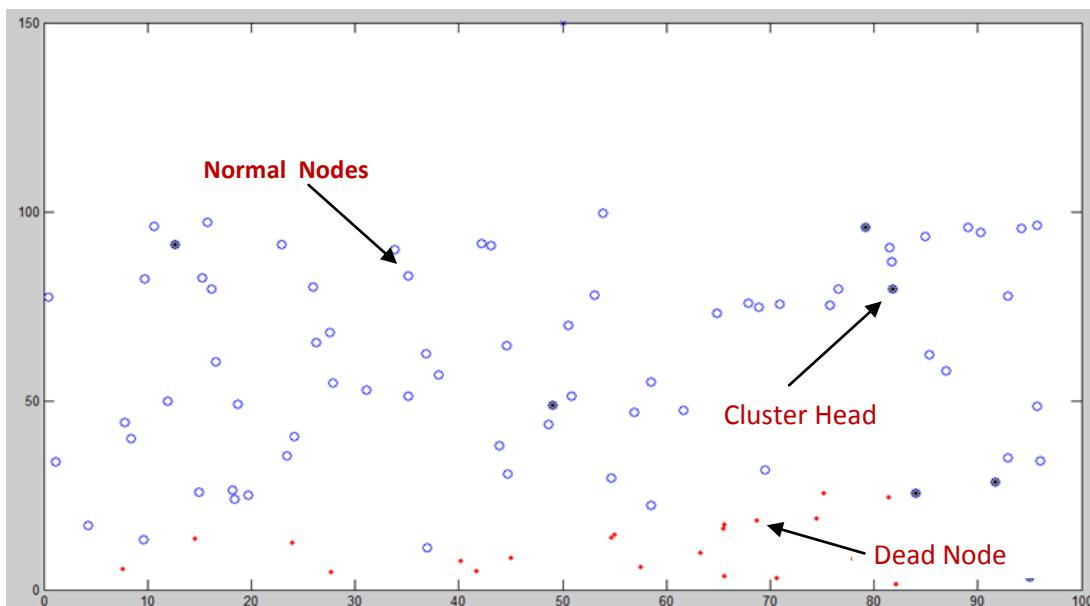
The terms represent the energy consumption of radio dissipation, while another represents the energy consumption for amplifying radio. The use of free space  $E_{fs}$  and the multi – path fading  $E_{mp}$  channel models depends upon the transmission distance. When receiving this data, the radio expands:

$$E_{RX} l = E_{RX-elec} l$$

Additionally, data aggregation operation will consume the energy  $E_{DA}$ .

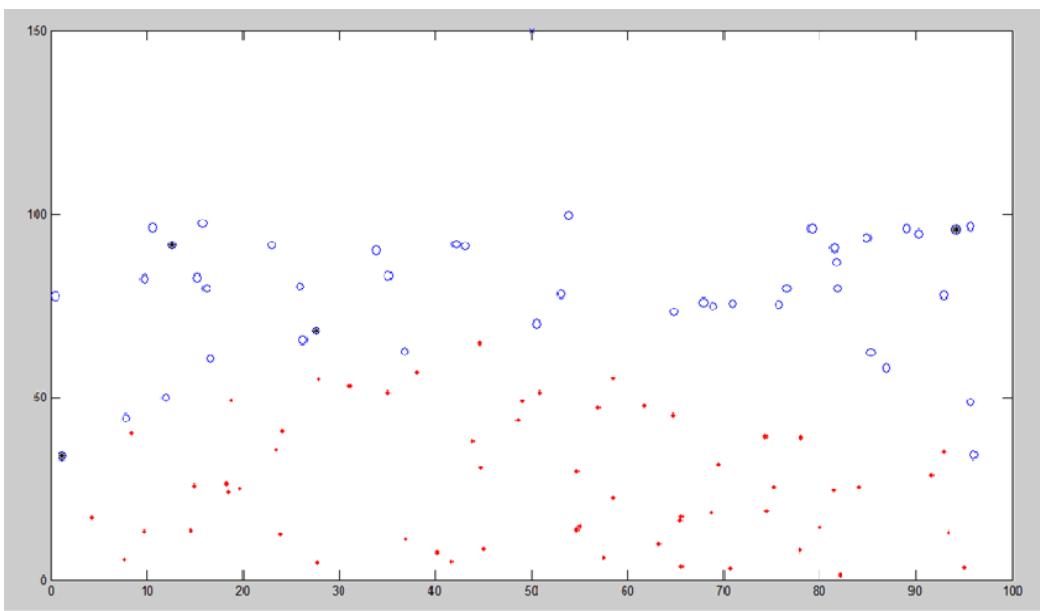
#### B. The Network Model Of Leach Initialization

Analysis of these two protocols is done using a 100-node network with randomly distributed nodes in a (100x100) meter area. There is some nodes used equal energy. Normal nodes are represented by 'o' and dead nodes are represented by '•'. And after Cluster formation cluster heads are represented by '\*'.



**Fig – 2 (Simulation result after some rounds in Leach Duty - Cycle)**

Here the length of each signal is 5000 bits and the energy required for data aggregation is 5nJ/bit/signal. Moreover, the initial energy of each node is 0.5 joule. This figure-2 shows the normal nodes, Cluster head and dead nodes after some rounds. The parameters are used in the implementation of these protocols are total dead node, dead first node, dead last node for 1500 rounds.

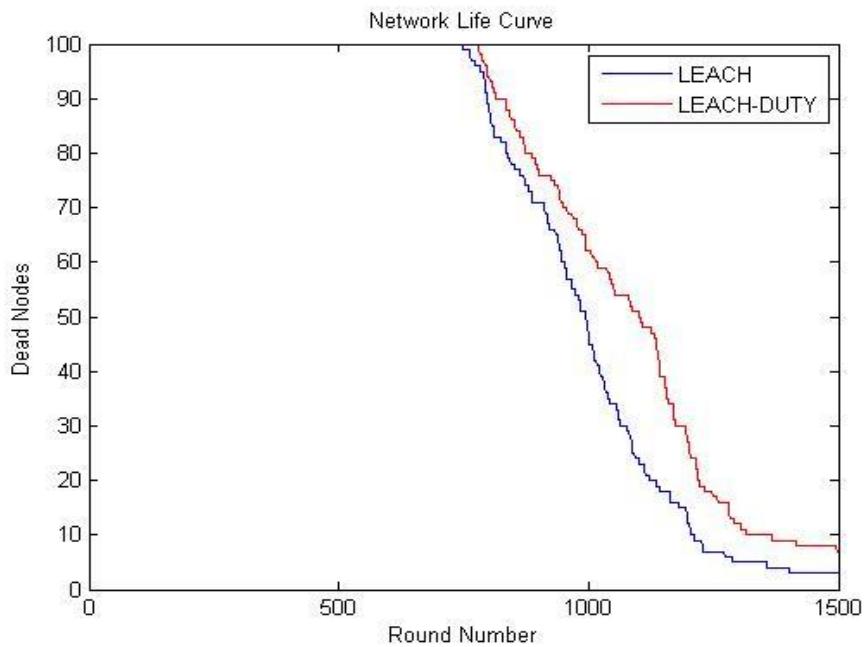


**Fig – 3 (Simulation result after some rounds in Leach Duty - Cycle)**

After 1300 rounds only the proposed protocol System sensor nodes are under dead position in figure - 3.

## V. RESULTS

I execute 1500 rounds of the simulator for this protocol & for each phase. The readings from these rounds were averaged & plotted. Performance of a clustering protocol is also determined by duration of stable period [13] in figure - 4. Analysis of first dead node is represented in Table 1.1



**Fig – 4 (Number of Dead node per round)**

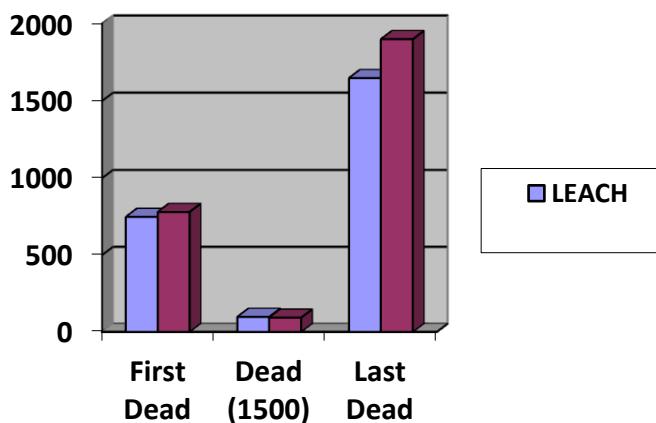
**Table 1.1:** First Dead node in LEACH and LEACH - DUTY

Routing Protocols	First dead node
LEACH	748
LEACH – DUTY	780

In homogeneous LEACH protocol first dead node is at 748 round but in proposed LEACH - DUTY routing protocol first dead node is at round 780 which improve the network Lifetime of LEACH - DUTY. It shows the behaviour of the network in Duty leach originally done in LEACH. Here taking into account the modified radio energy model. It shows the performance comparison for various parameters & transferring the PKTS to CH & Base Station.

#### A. Performance comparison

Now this Figure - 5 and table 1.2 shows the comparison of LEACH and Efficient Leach Protocol based on Duty – Cycle in Wireless Sensor Network considering parameter like total dead node, dead first node, and dead last node for 1500 rounds. From this Figure I can conclude that my proposed Efficient Leach Protocol based on Duty – Cycle in Wireless Sensor Network scheme improves the lifetime of the network about 10% than the existing LEACH protocol.



**Fig - 5** (Performance Comparison of considering various parameters (1500 Rounds)

**Table 1.2** Comparison of number of First Dead, Last Dead, no. of packets to CH and BS.

	First Dead	Last Dead	Pkt. to CH	Pkt. to BS
Leach	748	1648	89149	10080
Leach-Duty	780	1800	96466	10877

## VI. CONCLUSION

Leach is one of the routing protocols based on clustering algorithm to calculate the energy efficiency of the network. In this paper, the performance analysing using MATLAB shows by comparing the proposed Leach Duty - Cycle to Prolonging the Network Lifetime in Wireless Sensor Network with the normal homogeneous LEACH protocol. Where normal LEACH finds the first dead node is at 748 round but in proposed Leach Duty - Cycle to Prolonging the Network Lifetime in Wireless Sensor Network routing protocol first dead node is at round 780 which improve the network Lifetime of Efficient Leach Protocol based on Duty – Cycle in Wireless Sensor Network. Thus the proposed protocol is suitable to save the energy of the network, and provides better performance in energy efficiency has increased about 10% over homogenous Leach protocol. Finally, hence introduced Leach Duty - Cycle to Prolonging the Network Lifetime in Wireless Sensor Network as a new energy conservation paradigm with the prolonging the network lifetime.

## REFERENCES

- [1] I.F.Akyildiz, W.Su, Y.Sankarasubramaniam, and E.Cayirci, "A survey on sensor networks", Communication Magazine, IEEE, vol.40,pp.102-114,2002.
- [2] J.I.Hill,"System Architecture for wireless sensor networks", university of California, Berkeley, Ph.D.dissertation 2003.
- [3] A. Seetharam, A. Acharya, A. Bhattacharyya & M. K. Naskar, (2008) "An Energy Efficient Data Gathering Protocol for Wireless Sensor Networks", *Journal of Applied Computer Science*, Vol.2, No1
- [4] Ahmad Khadem Zadeh, Ali Hosseinalipour & Shahram Babaie, (2010) "New Clustering protocol for increasing Wireless Sensor Networks Lifetime" *International Journal of Computer and Network Security (IJCNS)*, Vol. 2, No. 1.
- [5] Laiaali Almazaydeh, Eman Abdelfattah, Manal Al – Bzoor, & Amer Al – Rahayfeh, (2010) "Performance evaluation of routing protocols in wireless sensor networks" *International Journal of Computer Science and Information Technology*, Vol. 2, No. 2.
- [6] Sun Limin, Li Jianzhong, chen Yu, "Wireless Sensor Networks", Tsinghua publishing company Beijing, 2005.
- [7] Kumar, D., Aseri, T. C. and Patel, R. B., "EEHC: Energy Efficient Heterogeneous Clustered Scheme for Wireless Sensor Networks", Computer Communications, Vol. 32, No.4, pp. 662-667, 2009.
- [8] R.Saravanakumar, S.G. Susila, J. Raja "Energy Efficient Homogenous and Heterogeneous System for Wireless Sensor Networks" *International Journal of Computer Applications (0975 – 8887)* Volume 17-No.4, March 2011.
- [9] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, "An Application Specific Protocol Architecture for Wireless Micro Sensor Networks", *IEEE Transaction on Wireless Networking*, vol. 1, Issue 4, pp 660-670, October 2002.
- [10] K. Padmanabhan, Dr. P. Kamalakkannan "Energy Efficient Adaptive Protocol for Clustering Wireless Sensor Networks" *IJCSI International Journal of Computer Science Issues*, Vol. 8, Issue 5, No 1, September 2011 ISSN (Online): 1694-0814.
- [11] S.Cho, A.Chandrakasan, "Energy-efficient protocols for low duty cycle wireless microsensor", In proceedings of the 33<sup>rd</sup> Annual Hawaii International Conference on System Sciences, Maui, HI Vol.2(2000),p.10.
- [12] Padmavathy. T.V, Chitra M "EERR: Performance Evaluation of Energy Efficient and Reliable routing protocol for wireless sensor networks" Data management and network control in wireless sensor networks (SICN), Volume (0), Issue (0): 2011
- [13] L. Alazzawi, A. Elkateeb, "Performance Evaluation of the WSN Routing Protocols Scalability," *Journal of Computer Systems, Networks, and Communications*, 2008, Vol. 14, Issue 2, pp. 1-9

# Flow Prioritization & Isolation Algorithm (FPIA) with Unified Multi link Algorithm (UMLA) for Optimized Vehicular Traffic

<sup>1</sup>Kulwinder Singh, <sup>2</sup>Supreet Kaur Gill

<sup>1</sup>Research Scholar

(Department of Computer Engineering)

UCoE, Punjabi University, Patiala

Email: kulvinder786x@gmail.com

<sup>2</sup>Assistant Professor

(Department of Computer Engineering)

UCoE, Punjabi University, Patiala

Email: supreetgill13@gmail.com

**Abstract**—The vehicular networks are being used in the variety of vehicular application to establish and control the communications in the vehicle to vehicle domain, vehicle to road side unit (RSU) or RSU to RSU, the vehicular networks are incorporated to share the various types of the traffic updates between the vehicles and vehicular networks to the `road side units (RSU). The major problem lies with the proper handling of the congestion of the traffic originated from the vehicular networks, which lies with the traffic congestion and the traffic stream isolation. In this paper, the major focus is established over the development of the adaptive vehicular network flow prioritization according to the urgency in the delivery of the traffic flows. Also the incorporation of the traffic inflow aggregation has been performed to combine the inflow streams for the guidance towards the target nodes or services. The proposed model combines the Unified Multi Link Aggregation (UMLA) along with traffic inflow Prioritization and Isolation Algorithm (FPIA) for the betterment of the vehicular networks. The experiments have been performed over the multifaceted simulation environments and the results have been obtained from each of the proposed model simulation. The experimental results have proved the efficiency of the proposed model in controlling the overall load and end to end delay by utilizing the traffic inflow organization and handling during the vehicular network transmissions.

**Keywords**— Vehicular traffic isolation, Traffic prioritization, Traffic management.

## I. INTRODUCTION

VEHICULAR AD HOC NETWORK or commonly called as VANET is an advancement that takes moving cars as nodes to create a mobile network .A network of wide range is possible with the help of VANET. This network allows communication among cars having maximum distance of 300 km. Every car is assumed to be a node in a VANET. New cars get connected to VANET when the earlier ones fall out of signal range. Police and Fire vehicles

are considered to be implementing it firstly so that safety can be ensured. Implementation of VANET will also aid in traffic management and provide services to public on the roads. Vehicular Networks have edge over traditional traffic management systems. It is due to the fact that VANET works intelligently as compared to traditional systems. Due to the intelligence possessed by VANETs they are being considered for implementation. Enhanced real time traffic signaling, safety of traffic and reduced emissions by vehicle are the advantages VANETs hold. A lot of research is being done on improving and creating a suitable enough VANET for safety purposes. A self sufficient system distributing emergency information between vehicles can be called as VANET. Existing technologies like UMTS, LTE, WiMax do not have any advantage as compared to VANET. Low cost implementation, self-organization and lower information dissemination time are the main advantages of VANET. It would not be wrong to say that VANET will be evolving application of MANET in coming times. Wireless interfaces are used by VANETs for communication. The power required for wireless interfaces is provided by the moving vehicles. Passenger safety and comfort are the major goals of VANET.

## II. LITERATURE REVIEW

**Miguel Sepulcre.** *et. al.* has worked on the Integration of congestion and awareness control in vehicular networks. The authors have proposed the congestion control method named as INTEgRatioN of congestion and awareness control (INTERN). INTERN aims to configure the transmission parameters of each or every vehicle so that its applications requirements can be satisfied and the channel load can be maintained below the target channel bit rate. In this context, all vehicles implementing INTERN will tend to use the minimum transmission setting that satisfy their individual applications requirements under high. The authors have incorporated the proposed model to acquire dynamically adaptable to changing vehicular formation at every second, channel Load Awareness and the Ability to maintain stable levels of channel load, which increases the application efficiency. The INTERN model hasn't been tested with the highly dense traffic congestion and does not utilize the load balancing approach to minimize the traffic load over single link. Also the existing model has been found not capable of the compression or optimization as it does not incorporate any message compression or optimization method.

**Ghaleb F.** *et.al.* has proposed the security and privacy enhancement in VANETs using mobility pattern. Through this paper the authors have assessed about mobility pattern based misbehavior detection approach in VANETs. The author in this paper starts by classifying the two attackers as outsider and insider. An intruder trying to intercept comes under the former type of attack while undesirable or unauthorized actions performed by a trusted node comes under the latter. a) Physical movement and b) information security perspectives are the metrics used by the author to detect misbehavior in VANETs. Anonymous Location-Aided Routing for MANET (ALARM) is implemented in this paper for vehicular network. This paper includes algorithm by which the misbehavior can be find and detected.

**Sharma G.** *et.al* has proposed the mechanism for security analysis of vehicular ad hoc network. The crux of this paper is about problems and challenges faced in VANETs and devising solutions to overcome these. In accordance to this paper each vehicle comprises of an OBU(On Board Unit).which connects vehicles with RSU via DSRC. and the other device is TPD(Tamper Proof Device),which store the vehicle secrets like keys, drivers identity, trip detail,

route, speed etc. DOS, Fabrication Attack, Alteration Attack are some of the attacks mentioned with Selfish Driver behind wheels, Pranksters etc. being the possible attackers.

**Seuwou P. et.al.** has proposed the effective security as an ill-defined problem in vehicular ad hoc networks. The author defined VANET as a mobile network created with the help of using moving cars as nodes. He emphasized the use of V2V and V2I communications for communications. He classified attacks into a) Physical b) Logical. Tamper proof device being the main cause of the former one and virus being that of the latter one.

### III. EXPERIMENTAL DESIGN

The proposed model is based upon the decongestion of the vehicular network by controlling the ingress and egress traffic flows. The traffic inflow (ingress flows) are the traffic inflows produced by the vehicular nodes and sent towards the base transceiver station (BTS or base station) node. The traffic outflow (egress flow), which is originated from the BTS node towards the end destinations (includes the vehicular nodes or other BTS nodes) is handled individually by the base stations and the intermediate stations for the handling of the data. The central concept behind the traffic decongestion solution is to divide and conquer the traffic and its amalgamation with the unified multi link aggregation (UMLA) algorithms. The vehicular nodes are programmed to itself handle the transfers and information exchange with its neighboring nodes. The distances between the vehicular nodes are expressed by the product of average per-hop distances and the shortest route between two nodes. The trilateral measurement technique is applied to obtain the location information of node. The vehicular traffic decongestion is direct numeric calculation of the minimal number of hops between road side units and vehicular nodes. The primary algorithm includes the calculation of minimum hops between each vehicular node and RSU or base station and other nodes by applying the deterministic distances among them. In the last step, the coordinates of each vehicular node is calculated with the help of maximum likelihood estimation approach or four edges measurement.

The vehicular traffic decongestion solution is based upon the combination of flow prioritization and traffic isolation as well as the unified multi link aggregation methods. The three-dimensional coordinates are utilized for the positioning and evaluation of all of the nodes within the vehicular cluster.

*Step 1- Calculate the minimum hops between each vehicular nodes and the road side units.*

The road side units relay its own location packets neighboring nodes. The node that receives record the least number of hops to every road side units and larger no of hops from identical road side unit are avoided. This step allows each and every node in the network to record minimum number of hops with all road side units.

*Step 2- Determine the exact distance between road side units and vehicular nodes.*

Use equation (1) to estimate the average per-hop distance after obtaining hop distances and the location information of other road side units [3].

$$c_i = \frac{\sum_{j \neq i} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}}{\sum_{j \neq i} h_j} \quad (1)$$

Where  $(x_i, y_i, z_i), (x_j, y_j, z_j)$  denotes the coordinates of road side unit  $i, j$  and  $h_{ij}$  depicts the hop-count between road side unit  $i$  and road side unit  $j$ . The vehicular nodes calculate distances to each and every road side units after receiving per-hop distance and transmit information to the adjacent nodes.

*Step 3- The coordinate of vehicular nodes are calculated by using maximum likelihood estimation method or four edges measurement.*

The hop distance of all RSU is used by vehicular nodes in step 2 to determine the coordinates of vehicular nodes by applying maximum likelihood estimation process or four edges measurement. (2) is used to calculate when distance from road side unit to vehicular node is well-known [3].

$$\begin{aligned} (x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2 &= d_1^2 \\ (x_2 - x)^2 + (y_2 - y)^2 + (z_2 - z)^2 &= d_2^2 \\ &\vdots \\ (x_n - x)^2 + (y_n - y)^2 + (z_n - z)^2 &= d_n^2 \end{aligned} \quad (2)$$

The linear equation in (2) can be represented as

$$AX=B \quad (3)$$

$$A = \begin{bmatrix} 2(x_1 - x) & 2(y_1 - y) & 2(z_1 - z) \\ \vdots & \vdots & \vdots \\ 2(x_{n-1} - x) & 2(y_{n-1} - y) & 2(z_{n-1} - z) \end{bmatrix} \quad (4)$$

$$B = \begin{bmatrix} x_1^2 - x^2 + y_1^2 - y^2 + z_1^2 - z^2 + d_1^2 - d^2 \\ \vdots \\ x_{n-1}^2 - x^2 + y_{n-1}^2 - y^2 + z_{n-1}^2 - z^2 + d_{n-1}^2 - d^2 \end{bmatrix} \quad (5)$$

$$X = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad (6)$$

The MMSE technique can be applied to acquire the coordinates of vehicular node P as per given in the following equation [4].

$$X = (A^T A)^{-1} A^T B \quad (7)$$

The original Flow prioritization and isolation with unified traffic aggregation algorithm does not deal with position deviation generated by the road side unit. The position deviation can be defined as the variation between the exact location and estimated position for the vehicular nodes. The enhanced algorithm follow the corrected per-hop distance for distance calculation between RSU and vehicular nodes and applies the total least Square for modification of road side unit position deviation.

In novel 3-D Flow prioritization and isolation with unified traffic aggregation algorithm, the first step involves rectification of distance between vehicular node and the road side unit. In the second step the change n the location of road side unit is corrected which adopts Total Least Square Method which not only examine arbitrary deviation vector consisting of observations and also rand coefficient matrix that includes errors.

---

*Algorithm 1: Flow prioritization and isolation Algorithm (FPIA)*

---

1. The RSU receives the vehicular data from the vehicular node.
2. The RSU perform the evaluation over the traffic inflow and evaluates its criticality.

3. The criticality of the traffic is supervised by the following bit patterns.
  - a. [1 1] for highly critical update.
  - b. [0 1] for moderately critical update.
  - c. [0 0] for normal updates.
4. All of the traffics with bit pattern [1 1] are marked with the highly critical updates.
5. All of the traffics with bit pattern [0 1] are marked with the moderately critical updates.
6. All of the traffics with bit pattern [0 0] are marked with the normal updates.
7. If the critical level depicts the hurdle or collision over the vehicular network.
  - a. Classify the traffic as the highly critical update and mark the node with red color.
8. If the critical level define the current node position parallel to the hurdle or collision over the vehicular network.
  - a. Classify the traffic as the moderately critical update and mark the node with orange color.
9. If the vehicular node have crossed the danger zone defined by critical algorithm.
  - a. Classify the traffic as the normal update and mark the node with green color.
10. And Normal otherwise.

---

*Algorithm 2: Unified Multi Link Aggregation Algorithm (UMLA)*

---

1. The traffic inflow similarity evaluation algorithm evaluates the similarity between the multiple traffic inflows on the basis of their type and criticality description.
2. Analyze the traffic inflows on each ingress flows over the RSU.
3. Perform the inflow segmentation after evaluating the correlation similarity.
4. Estimate the individual and collaborative traffic volumes being received from the multiple nodes.
5. If the overall traffic volume increases the maximum transmission limits.
  - a. Forward the contention limit request to all of the vehicular nodes.
  - b. Initiate the aggregation process over the unified traffics equal or lower than the controlled current contention window.
  - c. Contract the overall size of the aggregated traffic.
  - d. When the traffic according to the new contention window is satisfied.
    - i. Execute the aggregation models.
  - e. Else
    - i. Return to 5(a).
6. Evaluate the overall aggregate contraction level over the controlled inflows controlled with contention inflows limitation.
7. Initiate the UMLA process for the data aggregation.

#### IV. RESULT ANALYSIS

The results analysis is the analysis of the proposed model on the basis of various performance parameters. The performance parameters are the specified entities which defines the result of the implemented model. The performance parameters have to be selected on the basis of the nature of the project, algorithms being used and the testing data. The result analysis of the proposed model has been entirely based upon the various network performance parameters. The network performance parameters of throughput, delay, network load, packet delivery ratio and data drop rate has been obtained from the proposed model simulation.

##### A. End to End Delay

The end-to-end delay is the time from the generation of a packet by the source up to the destination reception, so this is the time that a packet takes to go across the network. This time is expressed in seconds (sec).

$$\text{Delay} = \sum_{t=1}^n \frac{\text{time length}}{\text{total no.of packets}} \quad \text{Eq. (4.1)}$$

Where, t = Time Interval

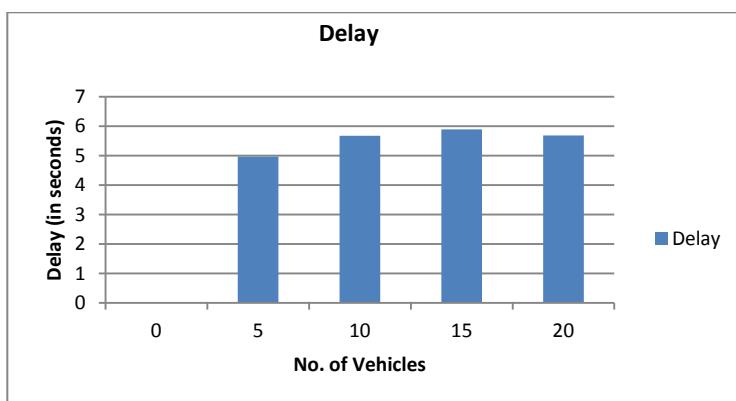


Figure 4.1.1: End-to-End Delay

The end to end delay has been recorded in the form of increasing curve as shown in the Figure 4.1.1. The latency increases with the rise in the traffic volumes.

SIMULATION TIME	DELAY
0	0
5	4.965886791
10	5.680131457
15	5.896184005
20	5.688784159

TABLE 4.1.1: Table of End-to-End delay

The above table 4.1.1 shows the traffic volumes. It has been increased gradually with the increase in the attacker nodes. The traffic volumes range between the 0 and 1.2 seconds, and average nearly at 0.5 seconds.

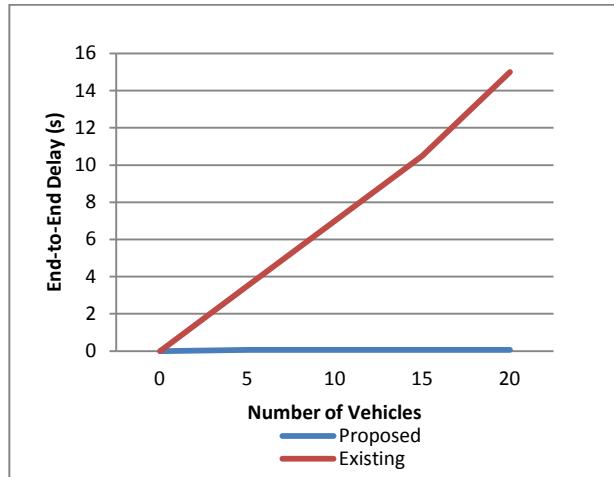


Figure 4.1.2: End-to-End Delay

The end-to-end delay based performance parameter evaluates the overall difference between the delay in the delivery of the data over the given channel for the vehicular network. The maximum end-to-end delay has been recorded nearly at 0.065 microseconds from comparison of the existing and the proposed model simulations.

#### B. Packet Delivery Ratio

Packet delivery ratio is calculated by dividing the number of packets received by the destination through the number of packet generated or sent from the source. It describes the loss rate. The Performance is better when packet delivery ratio is high.

$$PDR = \sum_{t=1}^n \frac{\text{total packets received}}{\text{total packets sent}} \times 100 \quad \text{Eq. (4.2)}$$

Where, t = Time Interval

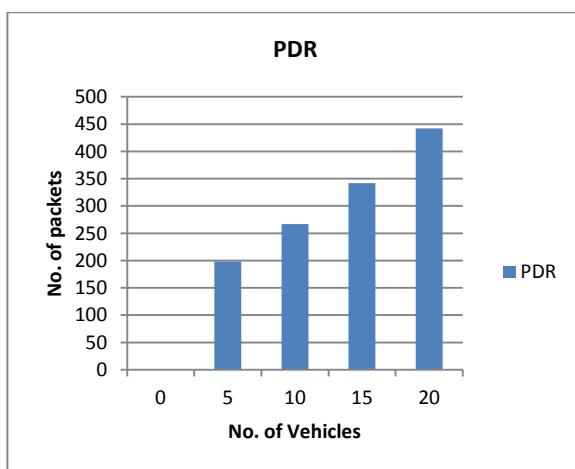


Figure 4.2: Packet Delivery Ratio

The packet delivery ratio is the parameter to indicate the successfully propagated data in comparison with the data volumes from the senders node.

SIMULATION TIME	PDR
0	0
5	198
10	267
15	342
20	442

TABLE 4.2: Table of packet delivery ratio

The above table (table 4.2) indicates the packet delivery ratio obtained from the proposed model simulation. The higher packet delivery ratio indicates the high trust factor of the given network as shown in the Figure 4.2. The packet delivery ratio has been recorded more than 442 packets in the proposed model simulation.

### C. Data Drop Rate

The data drop rate is the parameter which indicates the performance of the proposed model in the form of data loss due to the link bottlenecks or due to traffic overflow or other similar or non-similar reasons.

$$\text{Data Drop Rate} = \sum_{t=1}^{\text{total data sent}} \frac{\text{total data received}}{\text{total data sent}} \times 100 \quad \text{Eq. 4.3}$$

Where, t = Time interval

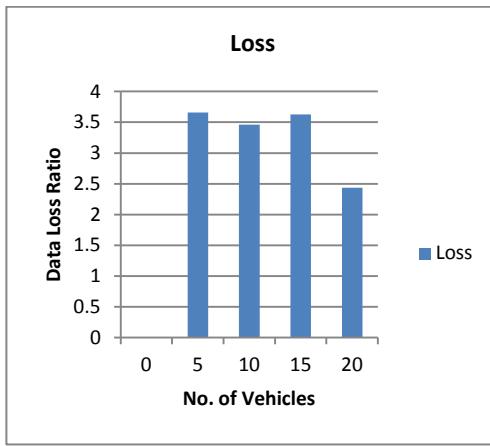


Figure 4.3.1: Data Drop Rate

The data drop rate is the parameter indicates the data drop percentage, and opposes the packet delivery ratio. The data drop rate has been recorded in the total bytes lost till the joining of the new node.

SIMULATION TIME	LOSS
0	0
5	3.656777715
10	3.463058537
15	3.62482501
20	2.433941414

TABLE 4.3: Comparative Analysis of Data Drop Rate

The above table (table 4.3) describes the data drop rate in the unit of bytes per second (bps) obtained from the proposed model simulation. The maximum data loss rate has been recorded nearly at 3 percent as shown in the Figure 4.3.1 in the whole simulation consisted of the 35 vehicular nodes.

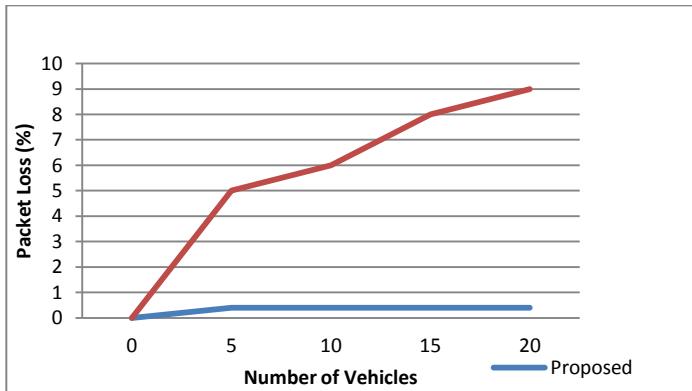


Figure 4.3.2: Packet Loss

The overall data loss has been obtained in the comparison with the existing model for the evaluation of the performance of the existing model and proposed model to handle the heavy traffic inflows over the bottlenecks created by the RSU. The proposed model has been found efficient against the existing model, because it has been recorded with approximately 0.4% lower packet loss than the existing model for the solution of traffic inflow congestion management.

## V. CONCLUSION

The proposed model is based upon the flow prioritization and the isolation algorithm for the isolation of the similar traffic inflows for the efficient delivery across the vehicular networks. The incorporation of the traffic inflow aggregation in the proposed model invokes the collaboration of the similar vehicular traffic streams for the efficient data propagation to improve the overall performance of the proposed vehicular model. The proposed model is evaluating the vehicular traffic inflows and outflows for the load balancing across the given vehicular network by

using the multiple vehicular stream based aggregation. The experimental results have been obtained from the simulation performed over the variable number of the nodes. The proposed model results have shown the performance of the results in comparison with the existing model for the vehicular networks.

## REFERENCES

1. Sepulcre, Miguel, Javier Gozalvez, Onur Altintas, and Haris Kremo. "Integration of congestion and awareness control in vehicular networks." *Ad Hoc Networks* 37 (2016): 29-43.
2. Ghaleb, Fuad A., M. A. Razzaque, and Ismail Fauzi Isnin. "Security and privacy enhancement in vanets using mobility pattern." In *Ubiquitous and Future Networks (ICUFN), 2013 Fifth International Conference on*, pp. 184-189. IEEE, 2013.
3. Samara, Ghassan, Wafaa AH Al-Salihy, and R. Sures. "Security issues and challenges of vehicular ad hoc networks (VANET)." In *New Trends in Information Science and Service Science (NISS), 2010 4th International Conference on*, pp. 393-398. IEEE, 2010.
4. Seuwou, Patrice, Dilip Patel, Dave Protheroe, and George Ubakanma. "Effective security as an ill-defined problem in vehicular ad hoc networks (VANETs)." In *Road Transport Information and Control (RTIC 2012), IET and ITS Conference on*, pp. 1-6. IET, 2012.
5. Javed, Muhammad A., and Jamil Y. Khan. "A geocasting technique in an IEEE802. 11p based vehicular ad hoc network for road traffic management." In *Australasian Telecommunication Networks and Applications Conference (ATNAC), 2011*, pp. 1-6. IEEE, 2011.
6. Hung, Chia-Chen, Hope Chan, and EH-K. Wu. "Mobility pattern aware routing for heterogeneous vehicular networks." In *Wireless Communications and Networking Conference, 2008. WCNC 2008. IEEE*, pp. 2200-2205. IEEE, 2008.
7. Dias, João A., João N. Isento, Vasco NGJ Soares, Farid Farahmand, and Joel JPC Rodrigues. "Testbed-based performance evaluation of routing protocols for vehicular delay-tolerant networks." In *GLOBECOM Workshops (GC Wkshps), 2011 IEEE*, pp. 51-55. IEEE, 2011.
8. Moser, Steffen, Simon Eckert, and Frank Slomka. "An approach for the integration of smart antennas in the design and simulation of vehicular ad-hoc networks." In *Future Generation Communication Technology (FGCT), 2012 International Conference on*, pp. 36-41. IEEE, 2012.
9. Sumra, Irshad Ahmed, Halabi Hasbullah, J. A. Manan, Mohsan Iftikhar, Iftikhar Ahmad, and Mohammed Y. Aalsalem. "Trust levels in peer-to-peer (P2P) vehicular network." In *ITS Telecommunications (ITST), 2011 11th International Conference on*, pp. 708-714. IEEE, 2011.
10. Sumra, Irshad Ahmed, Halabi Hasbullah, and J-L. A. Manan. "VANET security research and development ecosystem." In *National Postgraduate Conference (NPC), 2011*, pp. 1-4. IEEE, 2011.
11. Chen, Lu, Hongbo Tang, and Junfei Wang. "Analysis of VANET security based on routing protocol information." In *Intelligent Control and Information Processing (ICICIP), 2013 Fourth International Conference on*, pp. 134-138. IEEE, 2013.
12. Khabazian, Mehdi, and M. K. Mehmet Ali. "A performance modeling of vehicular ad hoc networks (VANETs)." In *Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE*, pp. 4177-4182. IEEE, 2007.

# Study on Industrial Electricity Forecasting using Artificial Neural Networks in Data Mining

Mallika Wadhwa<sup>#1</sup>, Er. Amrit Kaur<sup>#2</sup>, Sumit Bansal<sup>#3</sup>

<sup>#1</sup> wadhwa12miley@gmail.com

<sup>#2</sup> amrit.tiet@gmail.com

<sup>#3</sup>sb002@ymail.com

Department of Computer Engineering,

Punjabi University, Patiala-147002

**Abstract:** Accurate knowledge of electric power demand is the most basic need for the operation of industries. Artificial Neural Networks (ANNs) are the ones that can provide us the knowledge about the load demand of electric power systems. These are the predominant load forecasting technique used now days.

In this study, various ANNs with inputs, outputs and number of hidden neurons are examined, techniques for their optimizations are proposed and their results are discussed. This model is designed under MATLAB.

**Keywords:** Artificial neural networks, electric power, short term load forecasting, industries.

## I. INTRODUCTION

The industrial sector requires electric power in different ways. Electricity is often needed directly to raise the temperature of components used in the manufacturing process which is called process heating. ANN is used to solve the problems of electricity consumption. ANN is a kind of nonlinear system that simulates the structure, character and function of human brain. Neural network is able to own organization and map the information it receives during learning phase. It has capability of extracting important details from very large and accurate data. The multilayer perceptron model has been used for this purpose. Actual recorded input and output data that influence short-term energy consumption are used in the training, validation and testing process.

## II. NEURAL NETWORK

### A. Architecture of the ANN models

An artificial neural network contains collection of connected nodes with input, output and processing at each node. Nodes are weighted on the connection between input and hidden layers and also between output and hidden layers.

The multilayer perceptron model (MLP) is used for electricity forecasting. The MLP is a feed forward neural network with an input layer of source neurons, one or more hidden layer and one output layer.

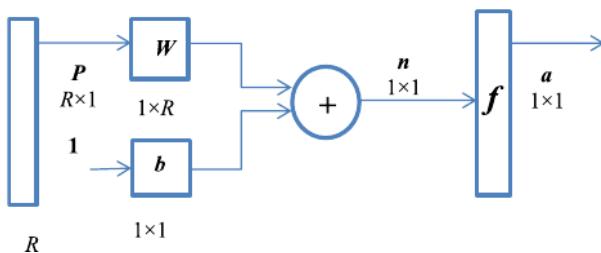


Figure 1. Multilayer Perceptron Model

Each processing element determines a net input value based on all its input connections. In the absence of special connections, the net input has been calculated by summing the input values multiplied by their corresponding weights. The net input can be written as

$$\text{Net}_{ij} = X_j \times W_i$$

#### *Node Properties*

The activation levels of nodes can be discrete (e.g. 0 & 1) or continuous across the range or unrestricted. This depends on the activation function chosen. If it is a hard limiting function then the activation levels are 0 (or -1) and 1. For a sigmoid function the activation levels are limited to a continuous range of real [0, 1]. The Sigmoid function at can be mathematically given as -

$$a_{t=1} / (1+e^x)$$

where x is the input variable. In the case of a linear activation function, the activation levels are open. The activation function is mentioned again in the case of the system dynamics.

### III. LITERATURE REVIEW

This section discusses previous studies that have been reviewed during the course of the research. Various research paper has been studied during this period of time.

Different Methods for electricity demand prediction:

1) *Regression Based Models* - The regression techniques involve formulating a mathematical model to examine the relationship between dependent variable and one or more independent variables. Difficulty of regression techniques is that correlation between component variables and observation variables is not stationary infect it depends somewhat on spatial temporal components.

2) *Time series*- Time series may be univariate or multivariate. It presents the load forecasts as a function of its past observed values while ignoring the exogenous factors. Disadvantage of time series technique is that we

know the actual shape of the distribution. Other disadvantages are they are time consuming and requires a lot of human intervention,

ANN is used to solve the problem of Load forecasting and give accurate result. This model indicates better results compared to that of a time series. For short-term load forecasting, we use the multi-layered feed forward ANN-based model. The study concludes that the ANN-based model gives better results.

#### IV. ENERGY CONSUMPTION

##### A. Network model

The network model explained in the preceding subsections can be expressed diagrammatically as-

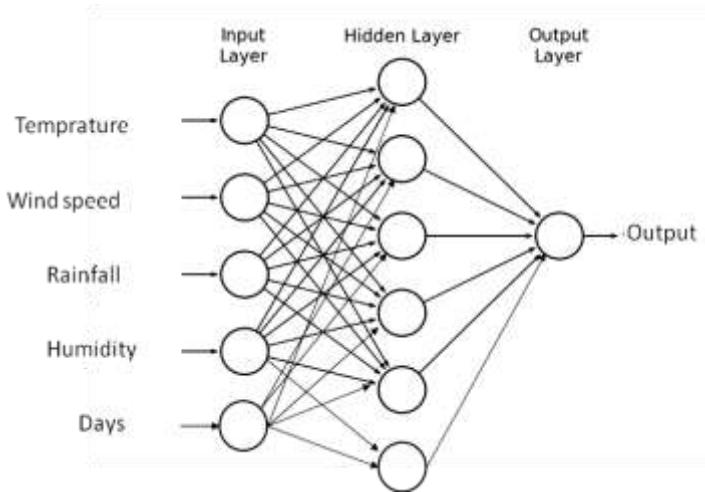


Figure 2. Neural Network Model

The back propagation algorithm is used for the above diagrammatic presentation.

The training data consists of two types of parameters Metrological parameters & electricity consumption data.

*Metrological parameters*- It is the input parameters of neural network.

Input variables	unit
Temperature	centigrade. ( $^{\circ}\text{C}$ )
Wind speed	$\text{m}\cdot\text{s}^{-1}$
Rainfall	Millimeter (mm)
Humidity	Percentage (%)
Days	Value is set to 1 if day under is not a working day otherwise 0.

*Electricity consumption Data-* It is the output of the neural network. This parameter is used to construct the performance function of neural network. The unit of electricity consumption data is KWH (kilowatt Hours).

### B. Back propagation (BP) neural network

The core of BP neural network is the back-propagation that tries to improve the performance of the neural network by reducing the total error.

The various steps involved in the Back Propagation Network algorithm.

Step1- Apply the input vector to the input units.

$$X_p = (X_{p1}, X_{p2}, X_{p3} \dots X_{pn})$$

Step2 - Calculate the net-input values to the hidden layer units.

$$Net_p^h = \sum W_{ji}^h X_{pi}$$

Step3 - Calculate the output from hidden layers.

$$I_{pj} = f_j^h (net_{pj}^h)$$

Step4 – It moves to the output layer and calculate the net-input values to each unit.

$$Net_{pk}^o = \sum W_{ki}^o i_{pj}$$

Step5 - Calculate the outputs from the output layer.

$$O_{pk} = f_k^o (net_{pk}^o)$$

Step6 - Calculate the error terms for the output units.

$$\delta_{pj}^o = (Y_{pk} - O_{pk}) f'_k (net_{pk}^o)$$

Step7 - Calculate the error terms for the hidden units

$$\delta_{pj}^h = f_{jn} (net_{pj}^h) \sum \delta_{pk}^o W_{kj}^o$$

Step8 - Update the weights on the output layer.

$$W_{kj}^o(t+1) = W_{kj}^o(t) + n \delta_{pk}^o i_{pj}$$

Step9 - Update the weights on the hidden layer.

$$W_{ji}^h(t+1) = W_{ji}^h(t) + n \delta_{pj}^h X_i$$

Where.

h - Number of the hidden layer in the network.

j - Number of nodes in the hidden layer

k - Number of nodes in the output layer.

i - Number of nodes in the input layer.

w - Connection strength or Weight.

$\delta$  - Error between actual and predicted value.

n - Learn Rate of the network.

O - Output demand calculated by the network

X<sub>p</sub> - Input variable to the neural network

Y - Actual demand.

### C. Forecasts Accuracy

To evaluate forecasting accuracy of the models obtained for ANN; actual data is compared with the predicted values obtained in the forecasting process using mean absolute percentage error (MAPE) and the coefficient of determination. For MAPE, the lower estimate is more reliable, while a higher is less reliable. The techniques were found to be reliable for determining accuracy for non-linear models. The MAPE function is used to measure the performance the models obtained. The equation is written as:

$$MAPE = \frac{1}{N} \sum_{n=1}^N \left( \frac{|Actual - Forecast|}{|Actual|} \right) \times 100\%$$

## V. CONCLUSION

The paper based on industrial electricity consumption by back propagation algorithm is to give the accurate result as compare to a time series. The BP neural network improves the performance of the neural network by reducing the total error. Five input variables are used that are calculated by BP neural network and give the output.

## REFERENCES

- [1] Filipe Rodrigues, Carlos Cardeira, J.M.F.Calado,2014 The daily and hourly energy consumption and load forecasting using artificial neural network method Energy Procedia 62 ( 2014 ) 220 – 229
- [2] G. Tamizharasi,2014 energy consumption by neural network Vol. 3, Issue 3
- [3] G. Macías-Bobadilla, feb 2013 Estimated electric power consumption by means of artificial neural networks and autoregressive models with exogenous input methods Vol. 8(14)
- [4] Xingping Zhang, Rui Gu,2007 Improved BP neural network for forecasting industrial electricity consumption in China Issue 1, Volume 1, 2007
- [5] Francis K. Kioko,2013 A review paper on Neural Networks for Short-Term Electricity Load Forecasting of Volume 63– No.2, February 2013
- [6] M. Apperley,2014 International Journal of Computer Applications (0975 – 8887) Volume 88 – No.15, February 2014
- [7] Karin Kandananond,2011 This paper is based upon Forecasting Electricity Demand with an Artificial Neural Network Approach *Energies* 2011, 4, 1246-1257

# WEB MINING RESEARCH: A STUDY

Pardeep Kaur<sup>1</sup>, Dr. Rekha Bhatia<sup>2</sup>

<sup>1</sup>(Punjabi University Regional Centre for Information Technology and Management, Mohali)

<sup>2</sup>(Punjabi University Regional Centre for Information Technology and Management, Mohali)

**Abstract:-**The massive information available on the web makes it a very prolific area of research using data mining techniques. The commercial potential of web is developing intense growth in the adoption and usage of web. The increase in usage of web is generating more content, structure and usage data day by day and thus the value of web mining is increasing. In this paper various web mining techniques have been studied. To understand these techniques in a better way this study focused on the algorithms used, machine learning methods and applications. The merits and demerits of these techniques have been situated. In addition to this the research gaps also have been determined.

**Keywords:** Web Mining, Unsupervised Machine Learning techniques, Information recollection, Information Extraction.

## I. INTRODUCTION

The World Wide Web (WWW) is a vast, interactive and widely known source of information. There is abundant of information available on the web today and it is piling up day by day. This huge information on the web has eased the search for data, but when a user wants to retrieve any kind of information from the web, it becomes difficult for him/her to get exactly what he or she is looking for. Sometimes the result of a query is not relevant or it is too late to be useful. The solution for this problem is web mining. Web mining is the concept of using machine learning and data mining techniques to extract information from web data, including hyperlinks between the documents, web documents, usage logs of web sites etc. Web mining is divided into three categories:

*Web Content Mining:* is an extraction method that extracts beneficial information from the contents of web. The content can be some kind of image, video, text, audio or tables. Major portion of the web content data is unstructured text data.

*Web Structure Mining:* is the process of finding the structure information from the web. This information structure is further divided into hyperlinks and document structure. Hyperlinks connect a location to a different location either within the same page or different page. The document structure organises a web page into tree structured format.

*Web Usage Mining:* technique observes the interesting usage patterns in the web usage data on basis of user's browsing behaviour. The web usage data is developed from web server access logs, user profiles, registrations, proxy server access logs etc. These three types of web mining techniques can be used in combination or separately, depending on the requirement of an application. This paper presents a review of various web mining techniques proposed and the tools used.

## II. LITERATURE REVIEW

Zhong Ji et al [1] proposed a feature extraction algorithm named hypersphere-based relevance preserving projection (HRPP) and a ranking function called hyper sphere based rank (H-Rank). Specifically, an HRPP is a spectral embedding algorithm to transform an original high-dimensional feature space into an intrinsically low-dimensional hypersphere space by preserving the manifold structure and a relevance relationship among the images. Furthermore, to catch user's aim without much human interaction, a reversed k-nearest neighbour (KNN) algorithm is proposed, which harvests enough pseudo relevant images by requiring that the user gives only one click on the initially searched images. The HRPP method with reversed KNN is named one-click-based HRPP (OC-HRPP). Finally, an OC-HRPP algorithm and the H-Rank algorithm form a new ISR method, H-reranking. Extensive experimental results on three large real-world data sets show that the proposed algorithms are effective.

Debina Laishram et al [2] represented method to extract news content from various news web sites on the basis of similarity in their representation like date, place and the content of the news that overcomes the cost and space constraint observed in previous studies which work on single web document at a time. This technique is an unsupervised method that builds a pattern representing the structure of the pages using extraction rules learned from the web pages by creating a ternary tree which expands when a series of common tags are found in the web pages. The analysis and the results on real time web sites validate the effectiveness of this approach.

Pimwadee Chaovalit et al [3] studied movie review mining using two approaches: machine learning and semantic orientation. The approaches are conformed to movie review domain for comparison. The results show that these approaches are better than many previous findings. It also finds that movie review mining is more challenging than other types of review mining. Movie review mining classifies movie reviews into two polarities: positive and negative. As a type of sentiment-based classification, movie review mining is different from other topic-based classifications. Few empirical studies have been conducted in this domain.

Yi Yang et al [4] proposed a new inductive algorithm for image annotation. This algorithm incorporates label correlation mining and visual similarity mining into a joint framework. Firstly a graph model is constructed according to image visual features. A multilabel classifier is trained by revealing the shared structure common to different labels and the visual graph embedded label prediction matrix for image annotation. This mechanism is applied to both web image annotation and personal album labelling using the NUS-WIDE, MSRAMM 2.0, and Kotak image data sets, and the AUC evaluation metric. These experiments showed that this algorithm can utilize both labelled and unlabeled data for image annotation and outperforms other algorithms.

Shanchan Wu et al [5] presented a combinational approach of a learning model and a grouping technology is to identify the actual data. The actual content finding problem is dealt as DOM tree node selection problem. They used DOM tree properties to train a machine learning model and developed multiple features. Further, candidate nodes are selected on the basis of machine learning model. Grouping technology is used to filter out noisy data and to pick the missing content. They conducted wide experiments on a real dataset from the product of HP SmartPrint and this approach showed high quality outputs and surpasses many baseline methods.

Gong-Qing Wu et al [6] conducted extensive case studies on web news extraction and designed a series of tag path extraction features to extract web news content. As every feature is different from the other, they combined all these features with DS (Dempster-Shafer) evidence theory, and then built a new content extraction method CEDS. Experimental results showed that compared with existing methods CEDS performed better on average. The results showed that F1-score with CEDS is 8.08% and 3.08% higher than present big content extraction methods CETR and CEPR-TPR respectively. Moreover, the performance of the CEDS is independent of the structure of the web page.

Clemes Koltringer et al [7] presented an automated web content mining approach. This approach shows the use of web content mining technique to extract destination brand identity and image from online sources. It extracts information from three digital image formation agents. These are the websites of destination marketing organisations (DMOs), user generated content from review pages and blogs, and editorial content of Anglo-American news media sites. It also makes it easier to understand the distinct roles and specialities of different types of online media. A large number of documents have informed the online destination representation in various online sources. Internet provides vital information to travellers and influence decision making. The results show that UGC is the richest and diverse source of online information.

Matthew Michelson et al [8] worked on an unsupervised approach that automatically selects the relevant reference sets and uses it for unsupervised extraction. The proposed algorithm involves three steps. First step is to pick out the relevant reference sets from a repository of reference sets. After choosing the reference sets, each post is matched to the members of reference sets. Finally, the reference set attributes are used to perform unsupervised extraction. Experimental results showed that this approach is competitive with supervised machine learning approaches. Further methods can be discovered to improve the accuracy of extraction and matching.

P. Moreno-Clari et al [9] presented a methodology for data analysis and apply it to the special case of university of Valencia (Spain). The entire analysis was focused on three groups of measurements. These were platform usage data, teaching innovation programs and education quality indexes. Useful information was obtained from these measurements. Although they were successful in the statistical analysis of LMS usage data using SPSS but to standardize their methodology the subsequent automation process is yet to be completed and has been left as a future work. Performance in exams, usage statistics, regression, number of visits, top search terms, number of downloads of e-learning resources is presented. Several DM approaches and techniques (clustering, classification and association analysis) have been proposed for joint use in the mining of student's assessment data in LMS.

Yung-Shen Lin et al [10] proposed a new symmetric measure to compute the similarity between two documents. Various features are embedded with this measure. Similarity of two documents with respect to a feature is computed considering three cases: 1) the feature appears in both the documents, 2) the feature appears only in one document and 3) the feature appears in none of the documents. In the first case, the similarity increases with the decrease in the difference between the feature values. For the second case, a fixed value is used. In the last case an absent feature has no contribution to the similarity. To check the effectiveness of the proposed measure, it is applied on some real world data sets for text classification and clustering problems. The results demonstrate that the performance of proposed technique is better than the other techniques.

TABLE 1

COMPARISION TABLE FOR DIFFERENT TECHNIQUES.

Sr. No.	Author	Title	Technique	Merits/Demerits
1.	Zhong Ji et al[1]	Relevance Preserving Projection and Ranking for Web Image Search Reranking	A feature extraction algorithm named hypersphere-based relevance preserving projection (HRPP) and a ranking function called hyper sphere based rank (H-Rank) is proposed.	This technique is robust and Human interaction is reduced. This is unsupervised technique so features increase exponentially.
2.	Debina Liashram et al[2]	Extraction of web news from web pages using a ternary tree approach	A ternary tree approach is proposed to extract news content from various news websites.	Automatic unsupervised extraction of news content. Extraction of multimedia data is not done.
3.	Pimwadee Chaovalit et al[3]	Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches	Two approaches machine learning and semantic orientation for movie review mining are proposed.	This approach is effective but machine learning techniques requires substantial amount of time. Misclassification occurs sometimes.
4.	Yi Yang et al[4]	Web and Personal Image Annotation by Mining Label Correlation With Relaxed Visual Graph Embedding	A new inductive algorithm is proposed that combines label correlation mining and visual similarity mining into a joint framework.	Efficiency and accuracy is increased comparatively.
5.	Yung-Shen Lin et al[10]	A Similarity Measure for Text Classification and Clustering	A symmetric measure is given to calculate the similarity between two documents.	This method mainly focuses on textual data and work on non textual data is not done.
6.	Shanchan Wu et al[5]	Automatic Web Content Extraction by Combination of Learning and Grouping	Combined approach of learning model and grouping technology is proposed. DOM tree properties are used for training.	The entire extracted pattern is compared, thus classification becomes more complex.
7	Matthew Michelson et al[8]	Unsupervised Information Extraction from Unstructured, Ungrammatical Data Sources on the World Wide Web	An unsupervised technique is presented that automatically selects the relevant reference sets and uses it for extraction.	This approach automatically extract data from unstructured and ungrammatical data sources but it is a challenge if one set of data is structured and one is unstructured.

### III. RESEARCH GAPS

Following are the research gaps, which have been found in the study of various techniques:

1. Techniques that use DOM tree approach for content extraction are more resource consuming and are expensive.
2. Most techniques apply unsupervised learning for content extraction and compare the entire extracted content. Thus there is exponential increase in features.
3. Very less work is done on the pattern of the content that provides important information for classification and analysis of data from the web.
4. Many classification techniques are there but preserving the accuracy in data classification needs to be focused.

### IV. CONCLUSION

In this paper, the reported literature has been reviewed on web mining. Web mining is a vivacious area of research. A lot of commercial research is happening in this area from various research communities such as database, artificial intelligence, and data mining. A great knowledge has been acquired by deeply studying all these web mining techniques and further this cognition will be used to advance web mining research. The existing research gaps and drawbacks of these techniques are also discussed. It is predicted from the study that

the interest in this research area will increase in future as web is becoming huge, irreplaceable source of information.

#### REFERENCES

- [1] Zhong Ji, Member, Yanwei Pang, Senior Member, and Xuelong Li, "Relevance Preserving Projection and Ranking for Web Image Search Reranking ", VOL. 24, NO. 11, NOVEMBER 2015.
- [2] Debina Laishram and Merin Sebastian, "Extraction of web news from web pages using a ternary tree approach," IEEE Second International Conference on Advances in Computing and Communication Engineering,, pp. 628-633, 2015.
- [3] Pimwadee Chaovalit and Lina Zhou, "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches," IEEE proceedings of the 38th Hawaii International Conference on System Sciences , 2005.
- [4] Yi Yang, Fei Wu, Feiping Nie, Heng Tao Shen, Yuetong Zhuang, and Alexander G. Hauptmann, "Web and Personal Image Annotation by Mining Label Correlation With Relaxed Visual Graph Embedding," IEEE transactions on image processing, vol. 21, no. 3, pp.1339-1351, March 2012.
- [5] Shanchan Wu, Jerry Liu, Jian Fan, "Automatic Web Content Extraction by Combination of Learning and Grouping," International World Wide Web Conference Committee (IW3C2), pp. 1264-1274, WWW 2015, May 18-22, 2015, Florence, Italy.
- [6] Gong-Qing Wu, Member, CCF, Lei Li, Member, IEEE, Li Li, and Xindong Wu, Fellow, IEEE, "Web News Extraction via Tag Path Feature Fusion Using DS Theory," journal of computer science and technology, pp. 661–672 July 2016. DOI 10.1007/s11390-016-1655-1
- [7] Clemens Költringer, Astrid Dickinger, "Analyzing destination branding and image from online sources: A web content," Journal of business research, Elsevier Inc, 1 Nov, 2015.
- [8] Matthew Michelson and Craig A. Knoblock, "Unsupervised Information Extraction from Unstructured, Ungrammatical Data Sources on the World Wide Web," International Journal of Document Analysis and Recognition (IJDAR), August 2007.
- [9] P. Moreno-Clari, M. Arevalillo-Herraez, and V. Cerveron-Lleo (2009) "Data analysis as a tool for optimizing learning management systems," in Proc. Ninth IEEE Int. Conf. Adv. Learn. Technol., pp. 242-246, Jul.2009.
- [10] Yung-Shen Lin et al, "A Similarity Measure for Text Classification and Clustering," IEEE transactions on knowledge and data engineering, vol. 26, no. 7, pp. 1575-1590, July 2014.

# Speckle Filtering of Ultrasound Images Using Improved Wavelet Shrinkage Guided Filter

Ramandeep Kaur<sup>1</sup>, Madan Lal<sup>2</sup>

<sup>1</sup> Post Graduation Student, Department of Computer Engineering, Punjabi University, Patiala, Punjab, India

<sup>2</sup> Assistant Professor, Department of Computer Engineering, Punjabi University, Patiala, Punjab, India

<sup>1</sup> deep.91325@gmail.com, <sup>2</sup> mlpbiuni@gmail.com

**Abstract**—Speckle noise is an intrinsic yet undesirable residual part of the medical ultrasound images, which considerably degrades the visual quality and limits the accuracy of automatic diagnostic techniques. Hence speckle suppression is an essential task before the examination and processing of the medical ultrasound images. Aiming at the problem of multiplicative speckle noise, an improved despeckled method is proposed in this paper, which is based on WaveletShrinkage\_Guided filter. The proposed method used the Daubechies20 (d20) wavelet transformation to decompose the ultrasound images and an improved wavelet shrinkage algorithm to filter the high-frequency component. The wavelet coefficients of low-frequency sub-bands are filtered by an improved guided filter with varying regularization parameter and the inverse DWT is used to obtain the noise free image. Synthetic image and an ultrasound image experiments with the comparison of the Guided filter and WaveletShrinkage\_Guided filter are carried out. The quantitative results of the proposed strategy show, that it outperforms the other despeckling methods in terms of favorable speckle suppression and edge preserving factors.

**Keywords**—Speckle suppression, Real Time ultrasound images, Wavelet transform, Guided filter, Edge preservation factor (EPF).

## I. INTRODUCTION

Ultrasound has gotten to be a standout amongst the most famous modalities in clinical imaging that is based on echo imaging technology. Of all the imaging modalities right now utilized as a part of the medical field for diagnostic purposes, ultrasound frameworks are considered to represent a minimal danger to the patients. This is on the grounds that non-perceptible sound waves with frequencies above 20 kHz are not known to cause any adverse effects in patients. Accordingly, the clinical application of ultrasonic imaging technology has turned out to be more essential, particularly in observing the growth status of the fetus in pregnant ladies and the diagnosis of injuries to the stomach organs. On the other hand, like to all coherent imaging techniques, the major weakness of an ultrasonic imaging is that it is contaminated by speckle noise. Speckle noise is produced by an interaction of the reflected waves from different autonomous scatters inside a cell determination [1].

## II. RELATED WORK

The requirement for image processing strategies to reduce the speckle noise has been demonstrated to enhance the image quality and increase diagnostics potential for medical ultrasound images. In this manner image denoising problem has been studied broadly. A number of speckle reduction approaches have been proposed based on a multiplicative model of the speckle noise.

Earlier speckle reduction method was proposed by **K.Thangaval, R.Mandavalan, I.Laurancle, Aroquiaraj (2009)**. They have done a comparative study of spatial filters for speckle noise reduction of ultrasound images and

they have proposed M-3 filter which is the hybridization of a mean and median filter. It performed better than many other filters according to statistical measures (RMSE, PSNR, and SNR) [2].

**T.Ratha Jeyalakshmi and K.Ramar (2010)** proposed a method for cleaning speckle noise in the medical ultrasound image. In this study, researchers have used a mathematical morphological operation. This algorithm uses a different strategy for reconstructing the features that are misplaced while removing the noise [3].

**Buemin et al. (2014)** propose the use of training with selected samples for the estimation of the optimal Boolean function. They study the performance of adaptive stack filters when they are applied to speckled SAR images [4].

**Kumar et al. (2014)** propose a method to reduce the speckle noise which uses the concept of fusion. The performance of the proposed algorithm is evaluated by computing measures like MSE, SNR, PSNR and MSSI (Measuring Structure Similarity Index), which gives information about the degree of feature preservation and denoising [5].

**Zhang et al. (2016)** proposed a new despeckling method based on an improved wavelet shrinkage filter and guided filter. An improved thresholding function is developed based on the universal wavelet thresholding function according to the characteristics of US images in the wavelet domain. The new wavelet shrinkage algorithm is designed by applying the Bayesian maximum a posteriori estimation. The speckle noise of low frequency (LL) coefficients is suppressed by using the guided filter [6].

### III. SPECKLE NOISE MODELING

In medical ultrasound imaging, a speckle pattern is produced due to interferences of backscattered echoes from the scatterers that are classically much smaller than the wavelength of an ultrasound wave. It has been known that speckle has a multiplicative nature [7].

Speckle noise obeys the gamma distribution that is given in the following equation:

$$F(g) = \left[ \frac{g^{\alpha-1}}{(\alpha-1)!a^\alpha} e^{-\frac{g}{a}} \right] \quad (1)$$

where  $F(g)$  represents an intensity level,  $a^\alpha$  is variance and  $g$  is gray level.

The presence of speckle noise degrades the quality of US images and restricts the development of automatic diagnostic methods. For the physician, speckle noise influenced their accurate diagnosis. Therefore, from the perspective of a medical application, reduction of speckle noise becomes an important process prior to the analysis and processing of US images. It is very important to understand the speckle noise model before the noise removal in medical ultrasonic images. When the ultrasonic signal is emitted into the human tissue, the reflected signal (envelope signal) is received and processed by the ultrasonic imaging equipment. The final ultrasonic envelope signal obtained consists of two parts: The useful reflected signal of the human tissue, and the noise itself, which includes additive noise and multiplicative noise. Multiplicative noise is related to the principle of ultrasonic signal imaging, which is caused by a random scattering fact in imaging cell resolution during the communication

procedure of the ultrasonic signal. Additive noise can be considered as system noise, such as sensor noise signal [8]. A generalized model of ultrasonic envelope signal can be written as:

$$g^{pre}(i, j) = h^{pre}(i, j)nm^{pre}(i, j) + na^{pre}(i, j) \quad (2)$$

Where  $h^{pre}(i, j)$  is the noise-free image,  $nm^{pre}(i, j)$  is the multiplicative noise (i.e., speckle) and  $na^{pre}(i, j)$  is the additive noise. By assuming that the additive noise (e.g., thermal and electronic noises) is trivial compared to the multiplicative speckle noise, Eq. (2) can be written as:

$$g^{pre}(i, j) = h^{pre}(i, j)nm^{pre}(i, j), (i, j) \in Y^2 \quad (3)$$

The multiplicative speckle noise can be converted to additive noise by applying the logarithm transform and log transformation is used to compress the envelope signal to fit within the display range. Then, Eq. (3) can be rewritten as:

$$\log(g^{pre}(i, j)) = \log(h^{pre}(i, j)) \log(nm^{pre}(i, j)) \quad (4)$$

Since various additive noise reduction techniques have been developed, speckle representation as an additive noise can be more effectively suppressed. In the proposed technique 2-D Daubechies20 (db20) operator is used to perform the 2-D discrete wavelet transform and the obtained wavelet of an Equation (4) is given as:

$$W_{l,k}^j(g) = W_{l,k}^j(h) + W_{l,k}^j(nm) \quad (5)$$

where  $W_{l,k}^j(g)$ ,  $W_{l,k}^j(h)$  and  $W_{l,k}^j(nm)$  are the wavelet coefficients of noisy images, original images, and speckle noise respectively and  $j = 1, 2, \dots, J, l, k \in Y^2$ . The subscripts  $(l, k)$  correspond to the wavelet domain coordinates,  $J$  is the largest decomposition layer, and  $j$  depicts the decomposition layers of discrete wavelet transformation. To make easy the representation, equation (5) can be rewritten as:

$$G_{l,k}^j = H_{l,k}^j + NM_{l,k}^j \quad (6)$$

#### IV. PROPOSED METHODOLOGY

An improved wavelet shrinkage-guided filter is proposed in this section according to the requirements of despeckling for ultrasound images. The block diagram of the proposed methodology is given in Fig.1.

- Ultrasonic envelope signal obtained by the ultrasonic imaging framework is compressed with logarithmic transformation ( $\log$ ) to fit inside the display range. Consequently, the multiplicative noise model of speckle is transformed to an additive noise model of speckle noise and noise-free signal, which is the basic medical ultrasound image. The noise-free signal will comply with the generalized Gaussian distribution, and the speckle noise will conform to the Rayleigh distribution.

- The 2-D Daubechies 20 (d20) discrete wavelet transformation is applied to the log-transformed image, and gets four frequency domains ( $LL^1, LH^1, HL^1$  and  $HH^1$ ). To carry on the procedure of wavelet decomposition for the low-frequency domain  $LL^1$ , four frequency domains ( $LL^2, LH^2, HL^2$  and  $HH^2$ ) are obtained. This step is repeated until the most extreme decomposition layer  $J$  is reached.
- According to the statistical properties of speckle noise and non-noise signal discussed in step 1, an enhanced wavelet threshold shrinkage algorithm [6] is used to process the wavelet coefficients of the high-frequency subbands in every layer ( $LH^j, HL^j$  and  $HH^j$ ,  $j=1, 2, \dots, J$ ). The standard deviation of speckle noise, image, and non-noise signals is evaluated respectively and threshold function of each layer is obtained.

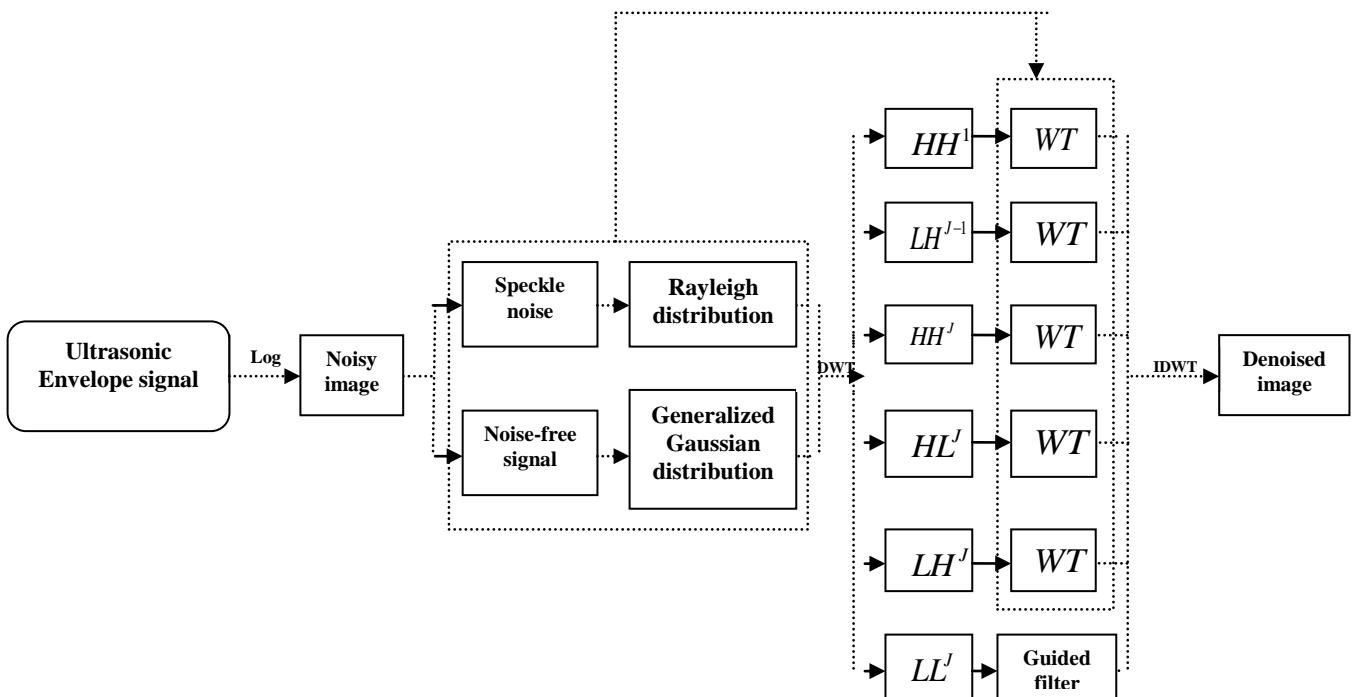


Fig.1. Block diagram of the proposed methodology

- Considering the way that the low-frequency sub-band of the last layer ( $LL'$ ) still exists a large amount of speckle noise and the bilateral filter suffer from two main weaknesses; high complexity and “gradient distortion”. In this way, an improved guided filter with varying regularization parameter ( $\varepsilon$ ) is designed based on a guided filter [6] that is used to filter  $LL'$ .
- Inverse wavelet transformation is used to handle the de-noised wavelet coefficients and obtains the despeckled medical ultrasonic images.

#### A. Threshold Determination

Regarding the wavelet noise reduction methods, the selection of threshold is a vital point of interest. Care should be taken to be able to preserve the edges of the denoised image. A very large threshold  $\lambda$  will shrink the vast

majority of coefficients to zero and may result in over smoothing the image while a small value of  $\lambda$  will result in the sharp edges with details being retained but may fail to suppress the speckle. There exist various shrinkage rules for wavelet thresholding, which rely on choosing a threshold value. Some typically used shrinkage rules for denoising image are Visu Shrinkage, Sure Shrinkage, and Bayes Shrinkage rules. In order to balance between the method used an improved threshold function [6] that is designed based on universal threshold function [9]. It is given by the following equation:

$$\lambda_i = a_j \sigma_n \sqrt{2 \log m} \quad (7)$$

where  $j = 1, 2, \dots, J$  are representing the decomposition layers of a wavelet transform,  $J$  is the largest decomposition layer and  $a_j$  is an adaptive parameter of  $j$  layer that is selected as  $\frac{1}{\ln(j+1)}$ . According to the characteristics of wavelet transform, most noise exists in the high-frequency sub-bands.

#### B. Wavelet Shrinkage Techniques

There are two general categories of thresholding techniques that are soft thresholding and hard thresholding. But these two general techniques do not provide an optimal performance. The proposed system used an improved wavelet shrinkage algorithm [6] that is shown as:

$$\hat{h} = \begin{cases} 0 & g \leq \lambda_i \\ sign(g) \cdot \max \left( |g| - \frac{\sigma_n^2 + \sqrt{\sigma_n^4 + 2\sigma_n^2\sigma_h^2}}{\sqrt{2}\sigma_h}, 0 \right) & g > \lambda_i \end{cases} \quad (8)$$

where  $\hat{h}$  is the estimation of  $h$  and  $g$  is assumed in phase with no-noise signal  $h$  and  $\sigma_n$  is the standard deviation of the noise which can be measured by detailed coefficients as given in the following equation:

$$\hat{\sigma}_n = \frac{median(|G_{i,k}^{HH}|)}{0.6745} \quad (9)$$

#### C. Improved Guided Filter

The guided filter is an edge-preserving filter that generates the filtering output by considering the contents of a guidance image, which can be the input image itself or another different image. The proposed method design an improved guided filter with varying regularization parameter ( $\varepsilon$ ) based on [6], that which is more consistent with the denoising of low-frequency sub bands  $LL'$ . The regularization parameter of guided filter is computed as:

$$\varepsilon = \lambda_i \times a_j \quad (10)$$

## V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the speckle noise simulation experiment is carried out in MATLAB 2016a to check the effectiveness of the proposed methodology. Experiments are carried out with the help of synthetic and ultrasound test images downloaded from [telin.ugent.be/~sanja/](http://telin.ugent.be/~sanja/) after adding speckle noise of different standard deviation

values. The US images are taken as test images and various despeckling methods are applied to those images in order to find out an efficient method. An acceptable filter should achieve an optimal balance between speckle suppression and edge preservation. In our experimentation, various types of performance evaluation parameters are used to measuring the quality of the proposed algorithm, Structural Similarity Index, PSNR, EPI and RMSE etc. We describe the some of the performance metrics in this section as follows:

EPI is used to compare the edge preservation ability of the filters and is computed by using [10]:

$$EPF = \frac{\sum (\Delta x - \bar{\Delta x})(\Delta y - \bar{\Delta y})}{\sqrt{\sum (\Delta x - \bar{\Delta x})^2 \sum (\Delta y - \bar{\Delta y})^2}} \quad (11)$$

where  $\Delta x$  and  $\Delta y$  represents the high pass filtered versions of images  $x$  and  $y$ , obtained with 3\*3 pixel approximation of the laplacian operator. The filter which has high EPF value will preserve the more edges.

The structure similarity index is used to measure the similarity between two images given by [11]:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (12)$$

where  $\mu_x$ ,  $\mu_y$ ,  $\sigma_x^2$  and  $\sigma_y^2$  represents the average and variance of the reference images.  $\sigma_{xy}$  is the covariance of  $x$  and  $y$ ,  $c_1$  and  $c_2$  are two variables to stabilize the division with the weak denominator. The resultant value of SSIM is a decimal value between 0 and 1, and the resultant value is 1 in a case of an identical structure of two images.

The RMSE is given by the following equation [12]:

$$RMSE = \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (x_{ij} - y_{ij})^2} \quad (13)$$

RMSE is used to measure the quality difference between the original image and noise-free image. The low RMSE represents the small difference between the both images.

PSNR used to measure the ratio of a possible power of a signal to the power of corrupting noise that affects the quality of representation an original image. It is characterized in logarithmic scale, in dB (decibel form) and it is easily defined by RMSE as given below:

$$20 \log_{10} \left( \frac{255}{RMSE} \right) \quad (14)$$

where RMSE is a quality difference between an original noise and corrupting noise. The higher ratio of PSNR represents the less obstructive the background noise.

#### A. Experiment on Synthetic Images

To ensure the objectivity of the quantitative assessment of de-noising methods, the speckle noise experiment on simulated synthetic image is carried out in this section. Fig.2(a) is a synthetic test image of size 512\*512 that is corrupted with speckle noise simulated with the MATLAB 2016a. Two different level of wavelet decomposition ( $d = \{1, 3\}$ ) and variance of speckle noise ( $\sigma = \{0.01, 0.03 \text{ and } 0.05\}$ ) are used to evaluate the efficiency of various despeckling algorithms. The despeckled images for the synthetic image experiment of different despeckled methods are shown in Fig. 2.

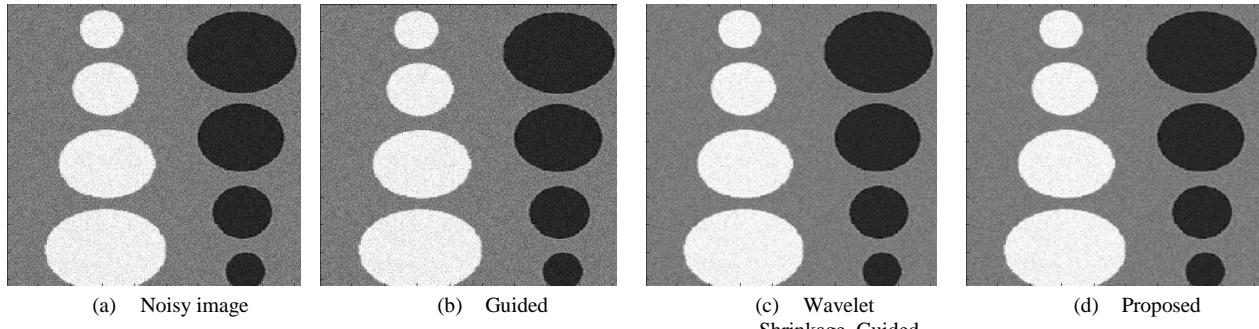


Fig.2. Denoised images of synthetic image experiment

TABLE 1 shows the performance measures obtained by using different speckle filtering methods for the simulated synthetic image experiment and it can be observed that the proposed algorithm obtained the best performance values. The proposed filtering method is an enhanced method based on WaveletShrinkage\_Guided method and quantitative results also demonstrate that the proposed algorithm outperforms the guided filter and WaveletShrinkage\_Guided denoising methods on PSNR value.

TABLE I. COMPARISON OF DIFFERENT DENOISING METHODS FOR THE SIMULATED SYNTHETIC IMAGE IN FIGURE 2(A).

Speckle Reduction Filters	Performance Parameters				Performance Parameters				Performance Parameters			
	EPF	SSIM	PSNR	RMSE	EPF	SSIM	PSNR	RMSE	EPF	SSIM	PSNR	RMSE
	Standard deviation ( $\sigma = 0.01$ )				Standard deviation ( $\sigma = 0.03$ )				Standard deviation ( $\sigma = 0.05$ )			
<b>Guided</b>	0.1294	0.3284	25.0555	7.2588	0.1317	0.1652	20.2971	8.6217	0.1326	0.1156	18.0839	9.0193
<b>Wav_Guided</b>	0.1317	0.6493	29.5984	4.3720	0.1308	0.4181	24.9766	6.3119	0.1321	0.3191	22.7891	7.0948
<b>Proposed</b>	0.1369	0.8130	31.5167	3.0823	0.1337	0.5666	26.4609	5.1968	0.1339	0.4399	24.0697	6.1923

#### B. Experiment on Ultrasound Images

The despeckling performance of various speckle filtering techniques in simulated synthetic image experiment with different level of noise variance can be effortlessly examined. In synthetic images, the comparative contrast between the background and simulated structures could be considered "good" differentiated, where the edges of the structures are effectively recognized. On the other hand, it must be noted that ultrasound images generally show low contrast among various tissues. Therefore, the analysis scheme for simulated synthetic image experiment presented in *Section A* is not the same as that of the experiment for real ultrasound images. In this way, two simulated

experiments must be considered independently and the simulated ultrasound image experiment is introduced in this section in order to analyze the efficiency of the proposed method for real ultrasound images. The medical ultrasound images are obtained from the medical images database [13], [14] and a medical liver ultrasound image of size  $482 \times 584$  pixels is used for the experiment that is shown in Fig.3 (a). The speckle filtering performance of the proposed method is compared with other despeckling methods and the despeckling images are shown in Fig.3.

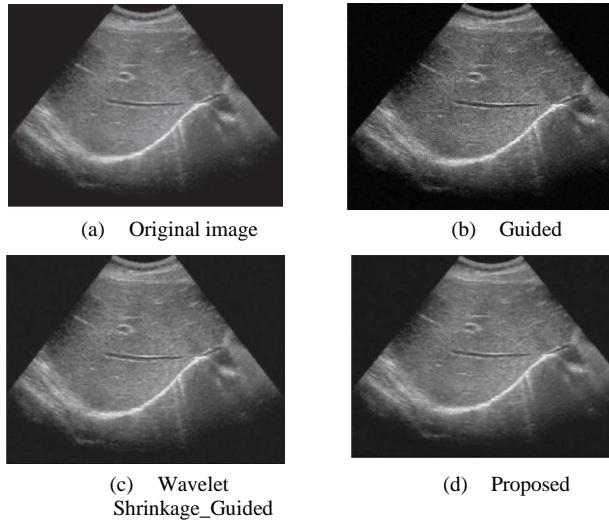


Figure 3. Denoised images of an ultrasound image experiment

TABLE II. QUANTITATIVE RESULTS OF SIMULATED ULTRASOUND IMAGE EXPERIMENT

Speckle Reduction Filters	Performance Parameters				Performance Parameters				Performance Parameters			
	EPF	SSIM	PSNR	RMSE	EPF	SSIM	PSNR	RMSE	EPF	SSIM	PSNR	RMSE
	Standard deviation ( $\sigma = 0.01$ )				Standard deviation ( $\sigma = 0.03$ )				Standard deviation ( $\sigma = 0.05$ )			
<b>Guided</b>	0.9002	0.7057	29.2093	5.6575	0.8507	0.4933	24.4577	7.2885	0.8287	0.3903	22.2591	7.9704
<b>Wav_Guided</b>	0.8590	0.8764	33.6652	3.5499	0.8438	0.7511	29.8100	5.0562	0.8233	0.6095	27.8175	5.7974
<b>Proposed</b>	0.8964	0.9056	34.0254	3.1453	0.8524	0.8150	30.7109	4.4728	0.8385	0.7522	28.7941	5.1770

The comparative analysis of despeckling methods in Fig.4 shows that the filtered ultrasound images of the Guided filter hold a considerable amount of noise and it did not obtain favorable EPF value. As compare to WaveletShrinkage\_Guided method, the proposed strategy has an improvement in every performance metrics. The highest PSNR and EPF values of the proposed strategy at noise level 0.03 and 0.05 shows that it performs better than other despeckling methods in terms of speckle suppression and feature preservation.

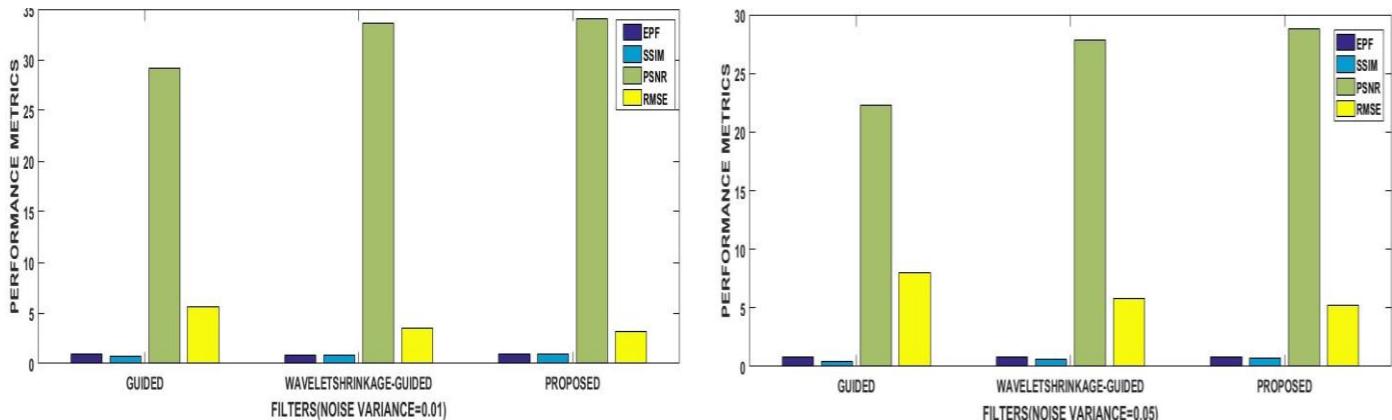


Fig. 4. Comparative analysis of despeckling methods at noise variance=0.01 and noise variance=0.05

## VI. CONCLUSION

In this paper, an improved despeckling method with varying regularization parameter of guided filter is presented for speckle suppression from medical ultrasound images. The proposed technique uses WaveletShrinkage\_Guided filter for its working. The simulated synthetic image experiment and real-time ultrasound image experiment are carried out. The parameters used to compare the performance of different techniques are EPF, SSIM, PSNR, and RMSE. The comparative performance measures of the various filtering methods on these experiments with varying noise levels ( $\sigma=0.01, 0.03$  and  $0.05$ ) and different wavelet decomposition levels ( $d=1, 3$ ) shows that the proposed strategy has a superior performance than other despeckling strategies for speckle noise in low-level frequency components and this method can significantly improve the performance indexes. The proposed strategy obtained the favorable EPF and PSNR values which ensure an improvement in an effective preservation of edges and suppression of speckle noise.

## REFERENCES

- [1] Ju Zhang, Guangkuo Lin, Lili Wu, Yu Cheng, "Speckle filtering of medical ultrasonic images using wavelet and guided filter" Published in Ultrasonic Volume 65, February 2016, Pages 177-193.
- [2] K. Thangavel \*1, R. Manavalan\*\*, I. Laurence Aroquiaraj, "Removal of Speckle Noise from Ultrasound Medical Image based on Special Filters: Comparative Study", ICGST-GVIP Journal, ISSN 1687-398X, Volume (9), Issue (III), June 2009
- [3] T.Ratha Jeyalakshmi and K.Ramar," A Modified Method for Speckle Noise Removal in Ultrasound Medical Images" INTERNATIONAL JOURNAL OF COMPUTER AND ELECTRICAL ENGINEERING, Vol. 2, No. 1, February, 2010 1793-8163.
- [4] Maria Elena Buemin, Alejandro C. Frery b, Heitor S. Ramos," Speckle reduction with adaptive stack filters" Published in Pattern Recognition Letters Volume 36, 15 January 2014, Pages 281–287.
- [5] Indrajeet Kumar, H.S Bhadauria, Jitendra virmani, Jyoti Rawat, "Reduction of Speckle noise from Medical Images using Principal Component Analysis Image Fusion" Published in Industrial and Information Systems (ICIIS), 2014 9th International Conference on Date of Conference: 15-17 Dec. 2014 Page(s): 1 – 6.
- [6] Ju Zhang, Guangkuo Lin, Lili Wu, Yu Cheng, "Speckle filtering of medical ultrasonic images using wavelet and guided filter" Published in Ultrasonic Volume 65, February 2016, Pages 177-193.
- [7] Burkhardt CB. Speckle in ultrasound B-mode scans. IEEE T Son Ultrason.1978; 25(1):1-6.
- [8] E.K. Abd, A. Youssef, Y. Kadah, Real-Time speckle reduction and coherence enhancement in ultrasound imaging via nonlinear anisotropic diffusion, IEEE Trans. Bio. Eng. 49 (9) (2002) 997–1014.
- [9] Donoho, D.L., Johnstone I.M. Ideal Spatial Adaptation By Wavelet Shrinkage, Biometrika, 1994, 81:425-455.
- [10] F.Sattar,L.Floreby ,G.Salomonsson, B.Lovstrom , "Image enhancement based on a nonlinear multiscale method", Image processing IEEE transactions on,vol.6 ,no.6,pp.888- 895,1997.
- [11] Z.Wang, A.Bovik, H.Sheik and E.Simoncelli,"Image Quality assessment: From error measurement to structural similarity, "IEEE Transaction on imageprocessing,Vol.13,No.4,pp.600-612,April,2004.
- [12] R.C. Gonzalez and R.E. Woods: 'Digital Image Processing', Addison- Wesley Publishing Company, 2002.
- [13] Image database. Ultrasound cases. <http://www.ultrasoundcases.info/Category.aspx?cat=73> (accessed 25.08.12).
- [14] Image database. MedPix-diagnosticimageatlas. <http://rad.usuhs.edu/medpix/> parent.php3? mode=image atlas (accessed 22.08.12).

# A Review Paper on Techniques Applied For HealthCare System

Arshbeer Kaur(M.Tech. Research Scholar)  
Computer Engineering Department,  
Punjabi University,  
Patiala, India  
arshbeer09496@gmail.com

Sikander Singh Cheema (Assistant Professor)  
Computer Engineering Department,  
Punjabi University,  
Patiala, India  
cheemasikander8@gmail.com

## *Abstract*

*This paper is a review paper, in which various techniques have discussed referring from many papers. For efficient detection, data mining, analysis of medical records in less time. Main emphasis is given on how we can maintain medical record in order to get fastest access to the practitioners, doctors and also patients and how we can reach remote area people by the help of technology. Evolutionary Algorithm and technology like telemedicine, equipments, interfaces, and tools have made hand in hand with the healthcare system and medical field to get efficient performance. This will help doctors to assess patient's health condition, who can't tell their condition through their mouth like children with mental disorder, deaf or dumb people.*

Keywords: *multi-agent; tools; algorithm.*

## I. INTRODUCTION

In the 21<sup>st</sup> century, health is as important as wealth as we all heard this many times in our life. People are trying to be more and more aware about their health. With the increase in temperatures problems affecting environment are like global warming, heavy rains at the time of harvesting, pollution, ozone layer depletion etc are very common these days. So, along with these problems people have to work for their growth and nation's economic growth as well. But these environment related problems are affecting very badly on their health. They need to keep a check on their health by giving it, the first priority. Researchers have invented a combination of technology with the health care system. Diagnosis, management, e-services, e-learning all are provided through the technology to the doctors, practitioners, patients and very quickly without wasting out their time, so that they don't' come under any loss. Government is also taking many strong steps towards it. By providing hospitals, doctors, medicines to each and every single corner of India. Techniques like multi-objective tools, equipments, multi-agent system are used by hospitals. Evolutionary Algorithms, data mining technique like Fuzzy Logic are applied to maintain records.

## II. LITERATURE SURVEY

*Author Michael A. LEE [1]*, has presented an algorithm for intelligent system design, which is evolutionary by behavior and based on a technique, which is multi-objective optimization technique. The situation like, where conflict occurs, when multiple systems are aggregated, this technique is necessarily used. Algorithm played its role by giving a search mechanism with the help of which, each objective is treated, independently and avoiding aggregation of objectives. The best in technique used here, is that, it does not give a single solution rather gives a set of solution as an output. For a better system design, these techniques are used in task of designing a system, which is basically a ‘fuzzy control’, one. Paper has presented metrics to check performance of multi-objective optimization algorithm and techniques for using these proposed metrics in designing a system which adapt evolutionary algorithm completely dependent on multi-objective optimization. This can be applied in many real world contexts where multiple objectives and relationship among them, which can't be easily handled, are used. Multi-objective tasks having other tasks like financial engineering and component layout, paper has applied techniques, which will easily generalize the tasks on fuzzy system design.

*Author Ping-Tsai Chung [2]*, proposed a software development system mainly to handle PDB i.e. Probabilistic Relational Database Applications and services. The services and applications are like disease analysis, detection, data mining and discovery of all records of medical field, control and prevention of diseases, management of hospital information system. PDB have probability measures to find uncertainty in data and related data items. The author mainly, has applied PDB to medical informatics after getting idea from application like internet comparison shopping, crime fighting information etc. Authors work is mainly focused to reduce diagnostic time, uncertainties, expenses which they did by discovering and developing biomedical informatics application which is PDB, itself. Also they worked, on implementing a strategy for database problems like incompatibility. Improved aspects like reliability, security, computing and trust of software system has made a query processed evaluation system for more advanced application.

*Author E. Sivasankar, [3]*, the paper has applied a technique, which is data mining based, naming Fuzzy Logic which is further based on classifier, in the diagnosis of serious pain of appendicitis in patients. The work is based on already existing statistics taken from patient's medical record. Fuzzy Logic based classifier is used in form of tool in detection of appendicitis pain. This tool easily classify patients to classes of appendicitis i.e. mild, moderate and severe using metrics like pain, nature, pain site, guarding, rigidity, previous surgery and tenderness, temperature.

*Author C.J. Carmona[4]*, aimed to specify rules which describes relationship of different modules and activities available in e-learning and final marks scored by students. Therefore, an application of discovering technique is introduced. Rules are also made in keeping teachers in mind to how to take steps and actions in assessing and further enhancing student's performance. Author has tested its algorithm with classical subgroup discovery algorithm and came to result; this evolutionary algorithm is much more suitable. They have made more interpretable rules for teacher and students and through this they can easily make decision about activities in order to enhance performance of students.

*Author Ellertson[5]*, has used smart mobile tools like tablets, to improve therapies module for autistic children and communicative disorders. For this cloud services and expert engine based on artificial intelligence for tracking

patterns and visually displayed patterns are used by system. After recognizing patterns , therapists and medical professionals working in field are alerted for further proceeding in treatment.

In this thesis, author *Nguyen and Venkatesh [6]*, talked about Autism blogs and posts, which are composed of three basic features: topics, language styles and affective information. These blogs, groups are examined with the help of machine learning and statistical methods. The results have given affective analysis for further monitoring and screening of autism related blogging online.

In this paper, authors *Basilio Noris[7]*, Mandy Barker, have used an eye-tracker WearCam, designed for children, to get the gaze information from viewpoint of child. It also helps in checking vision in early ages of child suffering from autism because autism and ASD also affects vision of child. This gives a chance to see the whole world from the viewpoint of these children.

In paper, the author *Ryoichi Komiya [8]*, has described a telemedicine prototype for future standardization. The problem of different interfaces has been resolved through this paper which results in difficulty of portability of medical equipment. Therefore, there introduced a standardization of interface between telemedicine equipment and telecommunication system. These packages are very efficient in performance for medical treatment in case of disastrous scenarios.

In this thesis, author *Ching-Seh Wu[9]*, proposed a multi-agent framework, which is based on architecture, which is service- oriented, for optimizing of medical data in information system of healthcare. Author has also introduced an Evolutionary Algorithm (EA) for dynamic optimization of quality of medical data. It provide two components, first, this EA is used to optimize medical processes into task according to quality attributes. Second, this multi-agent system helps to discover, monitor and report inconsistent attributes between optimized and actual medical records. EA accuracy in functionality and efficiency on multi-objectives and multi-domains QoS(Quality Of Service) attributes is very good. This algorithm can easily find optimal solution. This framework provides very good platform for medical practitioners and related persons in field of learning.

The author *Philippe De Wilde[10]*, proposed multi-agents ,which are distributed in nature containing nodes which perform functions that can't be easily written down in analytic form. The agents are connected to each other and connections contain information in them. The agents communicate to each other about their status via connections. Due to information stored in connection, they become memory less. This property allows us to design multi-agent system varying according to a factor, named fitness factor of each individual agent. Author built an entropy based definition for degree of instability which checks level of stability. This stability has become a useful tool for multi-agent system analysis and also gives an effective understanding of various systems.

The *author Anna Bonnel[11]*, proposed an analysis in detecting enhanced pitch sensitivity in autistic children specifically. Pitch is further categorized in many types' i.e. high pitch, low pitch, high confidence and sensitivity is also assessed by the help of graphs named Receiver Operating Characteristic (ROC). For detecting pitch, psychoacoustic tasks were taken off to get the better and efficient results.

*Author Bo Dao[12]*, has tried to resolve the problem of uncovering hyper communities. They have used Bayesian Non Parametric method and clustering algorithm which is also non parametric in nature. This all is done for online communities having blogs, posts etc.

The *author Sridari Iyer[13]*, has developed a game for autistic children to make them think out of box to move ahead. As the children, lack in eye contact, social behavior and impairments display a very repetitive behavior. This is very educative game, which will be played on computer. It will also help in improving their communication and recognition skills and also check level of autism and analyze the mental progress level. The system in game is very adaptive as each child is unique, if child is unique, if child is unable to go for next move game automatically and provide audio reminders of images is introduced complexity in game is increased according to repetitive behavior in mind. For social impairments, they will talk to other on screen characters in the form of casual conversations which will help them in talking to other people in real life. Game contains logs which can be analyzed and will provide feedback.

*Author Chunsheng Yang [14]*, has designed a multi-agent system for navigation training specifically and talked about issues like designing multi-agent e-education system, decision support system which is knowledge based, infrastructure of system. These are also good for systems which are multi-robotics based.

### III. PROPOSED METHODOLOGY

The review of all above papers has given a clue to try to make new technology, any algorithm or tool or any technique to make remote area people aware about the diseases, health and technology. For that, I will use any programming language which will be suitable for all platforms and later on also for mobiles. Now a day, mobiles are the most widely used tool, because it can make easy access to internet, which is like solution to all queries for the first step. The main emphasis is to invent user friendly technology especially for people living in remote areas, where internet is limited in reach.

### IV. CONCLUSION AND FUTURE WORK

By developing a technique to make remote area people friendly and aware about the healthcare system and diseases and new technologies. Developers and researchers further can do work in specific areas like for autistic children. If they make such a technique, by that they can make it available on cloud by cloud computing and that will help people to access whatever they want related to health system, very easily and in very less time. This will help them to take first step if any disease occurs and it will reach every part of India.

#### REFERENCES

- [1] LEE Michael A., et al, “Evolutionary Algorithms Based Multi-objective Optimization Techniques for Intelligent Systems Design”, 1996, *IEEE Conference at University of California, Berkeley, USA.*
- [2] Chung Ping-Tsai, et al. “A Software System Development for Probabilistic Relational Database Applications for Biomedical Informatics”, 2009 *International Conference on Advanced Information Networking and Applications Workshops.*
- [3] Sivasankar E., et al., “Knowledge Discovery in Medical Datasets Using a Fuzzy Logic rule based Classifier”, 2010 , *2nd International Conference on Electronic Computer Technology (ICECT 2010).*
- [4] Carmona C.J., et al., “Evolutionary algorithms for subgroup discovery applied to e-learning data”, 2010, *IEEE EDUCON Education Engineering 2010 – The Future of Global Learning Engineering Education.*
- [5] Ellertson Anthony, “Using smart mobile tools to enhance autism therapy for children”, 2012, *Frontiers in Education Conference Proceedings.*
- [6] Nguyen Thin, et al., “Expressed Emotion, Language Styles and Concerns in Personal and Community Settings”, 2015, *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 6.*
- [7] Noris Basilio, et al., “Measuring Gaze of Children with Autism spectrum Disorders in Naturalistic Interactions”, 2011, *33rd Annual International Conference of the IEEE EMBS.*
- [8] Komiya R., “A Proposal for Telemedicine Reference Model for Future Standardization”, 2005, *Proceedings of 7th International Workshop on Enterprise networking and Computing in Healthcare Industry, 2005. HEALTHCOM 2005.*
- [9] Wu Ching-Seh , et al, “Optimizing Medical Data Quality Based on multiagent Web Service Framework”, 2012, *IEEE Transactions on Information Technology in Biomedicine Volume: 16, Pages: 745 – 757.*
- [10] Wilde Philippe De , et al, “Stability of Evolving Multiagent Systems”, 2011, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Volume: 41, Pages: 1149 – 1157.*
- [11] Bonnel Anna, et al, Enhanced Pitch Sensitivity in Individuals with Autism: A Signal Detection Analysis”, 2003, *Journal of Cognitive Neuroscience, Volume: 15, Pages: 226 – 235.*

[12] Dao Bo, et al., "Nonparametric discovery of online mental health-related communities", 2015, *Data Science and Advanced Analytics (DSAA), IEEE International Conference on*, Pages: 1 – 10.

[13] Iyer Sridari, et al, "Research on Educative Games for Autistic Children", 2014, *International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*.

[14] Yang Chunsheng, et al, "Designing Multiagent-Based Education Systems for Navigation Training", 2006 , *5th IEEE International Conference on Cognitive Informatics, Volume: 1, Pages: 495 – 501*.

# An Enhanced watermarking scheme in DWT-SVD-FWHT domain using key based encryption

Dawinder singh

Research scholar,CE section

Yadawindera college of engineering  
Punjabi university, Guru kashi campus  
Talwandi sabo,151302,Punjab,india  
Email:singh2dawinder@gmail.com

Manoj kumar

CE section

Yadawindera college of engineering  
Punjabi university, Guru kashi campus  
Talwandi sabo,151302,Punjab,india  
Email:ermanojchaudhary@gmail.com

**Abstract-** Robustness, imperceptibility and high embedding capacity have been main concerns of researchers in development of new techniques in field of Digital Image Watermarking. Since embedding capacity and imperceptibility of Watermarked Image are contrary to each other, so it becomes tedious to achieve both. Here, in this paper we proposed an enhanced watermarking scheme in discrete wavelet transformation - fast walsh hadamard transform - singular value decomposition (DWT-FWHT-SVD) domain with high embedding capacity and good Imperceptibility shown by experimental results. As far as matter of robustness and security is concerned, watermark is being Encrypted using a key based algorithm to add another security layer. Gray scale cover Images (1024\*1024 size) are used for embedding four gray scale watermarks (256\*256 size) into single cover image. Peak Signal to Noise Ratio (PSNR) and the Normalized Correlation (NC) are parameters used to scrutinize the performance of given technique. To measure the robustness we calculated NC and we also subjected this technique to different attacks to show its robustness against attacks.

**Keywords**—Encryption; DWT; SVD; FWHT;

## I. INTRODUCTION

In this globalised world, exchange of information through internet has been increasing at very sharp pace. A lot of information is being conveyed in form of images in fields like Medical, Advertisement etc. Authenticity and owner identity of an image can be easily compromised over internet, so to evade these problems area of watermarking is one of the most researched area in last decade. Watermarking is a scheme of hiding owner's identity or other valuable information into image for making it more authentic. Watermarking can be done in a spatial domain or in transform domain which highly recommended because spatial domain does not provide robustness and security as transform domain does [1]. A highly used technique in transform domain watermarking is SVD in which data is embedded into cover image by modifying its singular values coefficients. It is used along with another techniques like DWT, DCT (Discrete cosine transform) and fast hadamard transform [5]. We used key based encryption of watermark image before embedding to enhance security of the scheme. Several attacks like speckle attack, salt and pepper can harm the quality of extracted watermarks [4]. We checked robustness against various attacks. In rest of paper, section 2, 3 and 4 contains related work, proposed scheme and experiment results respectively. At the end, conclusion is written in section 5.

## II. RELATED WORK

Many researcher have proposed several watermarking schemes by using different methods for both color as well as gray scale images. In[1], A. Ray et al combined DWT with SVD for embedding watermark in all four sub bands of image using RSA. In[2], S. Lagzian et al. used RDWT with SVD to embed watermark. In[3], G. Çetinel et al. 3<sup>rd</sup> level DWT and SVD is used along with arnold's cat map to provide a robust watermarking scheme. We have used 2<sup>nd</sup> level of DWT to embed watermark in all of sub bands. In[4], Chunlin Song et al. Gave summary of attacks on watermarking. We subjected our scheme against several attacks.

## III. PROPOSED SCHEME

In this scheme we used a DWT and SVD along with FWHT but in a different manner to embed watermark in all of sub bands. Watermark is encrypted before embedding and decryption is performed on extracted watermark.

### A. DWT

Wavelet breaks image into four bands LL (approximate sub-band), HL (horizontal sub-band), LH (vertical sub-band) and HH (diagonal Sub-band) as shown in Fig.1 [1]. DWT can be applied up to n different levels as shown in Fig.1 .

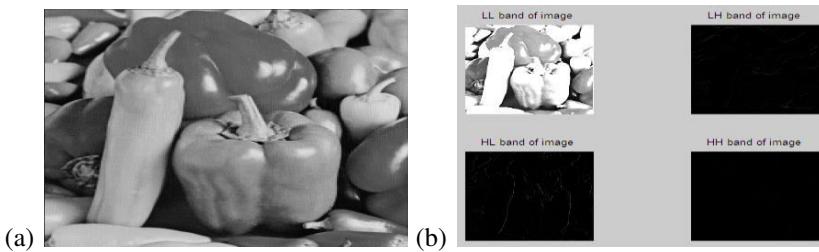


Figure 1 DWT decomposition. (a) Pepper image, (b) 1-level Haar DWT image

We have applied 1-level HAAR wavelet on cover image and then we have applies 2-level HAAR wavelet on each sub-band and selected HH bands from each sub-band for embedding given by 2-level HAAR wavelet.

### B. SVD

SVD is a mathematical tool which decomposes a matrix. Matrix is divided into three matrices, two orthogonal matrices and one diagonal matrix. If A be an MxN matrix. After SVD divides it as:

$$A = U_A S_A V_A^T$$

$$U_A = [u_1, u_2, \dots, u_N]$$

$$V_A = [v_1, v_2, \dots, v_N]$$

$$S_A = \begin{pmatrix} s_1 & \dots & \dots \\ \dots & s_2 & \dots \\ \dots & \dots & s_3 \end{pmatrix}$$

Where,  $U_A$  and  $V_A$  are orthogonal matrices  $S_A$  is a diagonal matrix of  $N \times N$  size singular values. These diagonal values are replaced with  $U_A$  diagonal values of watermark. Singular values correspond to brightness and singular

vectors represent geometry characteristics of the image. Since little variations in singular values of an image don't affect the visual quality, mostly singular values are used for embedding [5].

### C. FWHT

The Walsh-Hadamard transform decomposes a signal into a number of functions that are rectangular or square waves with values of +1 or -1. These basis functions are non-sinusoidal, orthogonal transformation technique. For 2-D images FWHT coefficients are calculated by first row wise then column wise[9]. The 2X2 Hadamard matrix is defined as H1 is given as:

$$H_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad [9]$$

### D. Encryption/Decryption Algorithm

For encrypting watermark image we used an algorithm which first generated key matrix of size depending upon the size of watermark image, then it encrypt each pixel using key matrix. Different keys are used for decryption of pixels of watermark image.

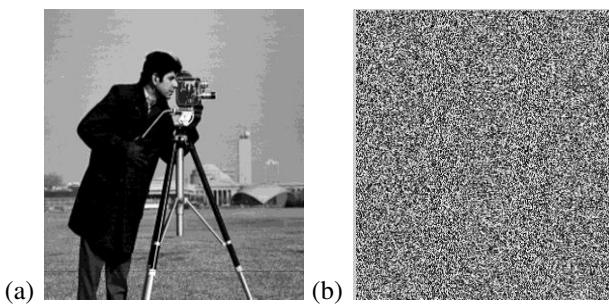


Figure 2 (a) original watermark, (b) encrypted watermark

### E. Watermark Embedding and extraction

The proposed scheme for embedding/extracting watermark into cover image is described in Table1 and Table2 given below.

TABLE 1  
 Steps of watermark embedding scheme

1:	Read 1024x1024 size grey scale 'Cover Image'.
2:	Apply 1-level DWT Haar wavelet on cover image to get sub-bands LL,LH,HL,HH
3:	Apply 2-level DWT Haar wavelet on Each of sub-bands. [LL1,LH1,HL1,HH1]=dwt(LL); [LL2,LH2,HL2,HH2]=dwt(LH); [LL3,LH3,HL3,HH3]=dwt(HL); [LL4,LH4,HL4,HH4]=dwt(HH);
4:	Apply FWHT on sub-bands HH1,HH2,HH3,HH4 obtained from 2-level DWT. HH1=fwht(HH1);
5:	Apply SVD on these sub-bands. [U,S,V]=SVD (HH1);

6:	Load Watermark image of size 256x256.
7:	Encrypt watermark image and apply SVD. $[U_i, S_i, V_i] = SVD;$
8:	Modify singular coefficients of cover_image $S_{ii} = S + (S_i * \alpha)$ ( $\alpha$ has value between 0 and 1) (we take $\alpha=0.004$ )
9:	Apply inverse SVD on each band after modification. $NEW\_HH1 = U * S_{ii} * V'$
10:	Apply Inverse FWHT on all modified sub bands.
11:	Apply inverse DWT for 2-level and 1-level ,respectively.
12:	Show "watermarked_image" and PSNR

TABLE 2

Steps of watermark extracting scheme

1:	Read 1024x1024 size grey scale ' Watermarked Image'.
2:	Apply 1-level DWT Haar wavelet on cover image to get sub-bands LL, LH, HL, HH
3:	Apply 2-level DWT Haar wavelet on Each of sub-bands as in step 3 in embedding process
4:	Apply fwht on sub-bands HH11,HH21,HH31,HH41 obtained from 2-level DWT.
5:	Apply SVD on these sub-bands. $[U, S_{ii}, V] = SVD (HH1);$
6:	Read 1024x1024 size grey scale ' Cover Image'.
7:	Apply 1-level DWT Haar wavelet on cover image to get sub-bands LL,LH,HL,HH and 2-level DWT on each sub-bands.
8:	Apply FWHT and SVD on HH1,HH2,HH3,HH4 sub-bands obtained from 2-level DWT as in step 4 in embedding process. $HH1=fwht(HH1);$ $[U, S, V] = SVD (HH1);$
9:	Get singular values from all sub-bands using equation: $S_i = (S_{ii} - S)/\alpha;$
10:	Get value of $U_i$ , and $V_i$ as in step 7 in embedding process and apply inverse SVD Encrypted watermark $= U_i * S_{ii} * V_i'$
11:	Decrypt the watermark using decryption algorithm.
12:	Show the all of four extracted watermarks and normalized correlation (NC).

#### IV. EXPERIMENTATION AND RESULTS

The proposed scheme is implemented in MATLAB R2014a (8.3.0.532) and experiment is carried out on personal computer Intel(R), Core(TM) i3- 2350M, processors rated at 2.30 GHz, main memory of 4 GB and 32 bit Microsoft

Windows 7 operating system. We used different gray scale images of size 1024x1024 as cover image and cameraman image of size 256x256 as watermark to be embedded. For measuring the performance of the given technique we have used parameters PSNR and NC

#### A. Peak Signal to Noise Ratio

The peak signal to noise ratio is used to measure imperceptibility of watermarked image with respect to original cover image. Higher value of psnr implies the good perceptual quality of image and is measured in dB (decibels) [2]. It is given as:

$$PSNR = 10 \cdot \log_{10} \left( \frac{MAX_I^2}{MSE} \right) \quad [2]$$

Where MSE is Mean square error.

$$MSE = \frac{1}{mn} \sum_0^{m-1} \sum_0^{n-1} ||f(i,j) - g(i,j)||^2 \quad [2]$$

Where  $f(i, j)$  and  $g(I, j)$  are pixel values of original image and watermarked image, respectively [9].

#### B. Normalized correlation

Normalized correlation is a parameter to check the robustness of watermarking scheme. NC is used to measure similarity between extracted watermark and original watermark[9]. It is given as:

$$NC = \frac{\sum_{k=1}^n w_1 w_2}{\sqrt{\sum_{k=1}^n w_1^2} \sqrt{\sum_{k=1}^n w_2^2}} \quad [2]$$

Where  $w_1$  is embedded watermark and  $w_2$  is extracted watermark.

TABLE 3  
 PSNR for different images

Images	Cover Image	Watermarked Image	PSNR
Peppers			34.0723 db
Lena			34.3955 db
Baboon			29.0510 db

Lake			32.8596 db
House			36.6332 db

To check the robustness of proposed scheme watermarked image is subjected to many attacks and extracted watermarks from infected image are compared with original one to calculate NC values as shown in table 4. Value of NC varies between 0 and 1, if we calculate NC for same images, value of NC will be equal to 1. NC values shown in table 4 implies that our proposed scheme makes all of four embedded watermarks almost equally robust against several attacks.

Table 4  
 NC values after different attacks

Attacks	Strength of attack	NC values for Extracted watermarks from all of four bands			
		HH1	HH2	HH3	HH4
No Attack		0.9903	0.9915	0.9911	0.9914
Gaussian noise attack	0.001	0.8643	0.8277	0.8265	0.8150
	0.005	0.7778	0.7871	0.7875	0.7853
Salt and pepper attack	0.001	0.7455	0.7438	0.7434	0.7424
	0.005	0.7443	0.7432	0.7435	0.7426
Speckle attack	0.001	0.8703	0.8324	0.8308	0.8219
	0.005	0.7773	0.7876	0.7880	0.7860
Poison attack		0.7787	0.7836	0.7833	0.7802

#### V. CONCLUSION

We have proposed a watermarking for embedding watermark in all of four bands (LL, LH, HL, HH) and achieved watermarked image with high perceptual quality. Results show robustness of scheme against many noise

adding attacks. For future work this technique can be used with other techniques like CWT (complex wavelet transform) and SWT (stationary wavelet transform). Further, this technique can be extended for audio/video watermarking.

#### ACKNOWLEDGMENT

We thank the whole management of yadawindera college of engineering for their support and to authors of all review papers we have studied related to watermarking techniques.

#### REFERENCES

- [1] A. K. Ray, S. Padhikary, P. K. Patra and M. N. Mohanty, "Development of a new algorithm Based on SVD for image watermarking," Computational Vision and Robotics, Springer India, ,2015, pp. 79-87.
- [2] Samira Lagzian, Mohsen Soryani, Mahmood Fathy, "Robust watermarking scheme based on RDWT-SVD:Embedding Data in All subbands," Artificial Intelligence and Signal Processing (AISP), 2011 International Symposium on,2011, pp. 48-52.
- [3] Gokcen cetinel , Llukman cerkezi, "Chaotic Digital Image Watermarking Scheme Based on DWT and SVD, 9th International Conference on Electrical and Electronics Engineering (ELECO), 2015, pp. 251-252.
- [4] C. Song, S. Sudirman and M. Merabti, "Analysis of digital image watermark attacks," in Proc. IEEE Int. Conf. Consumer Communications and Networking, Las Vegas, United States, 2010, pp. 1-5.
- [5] V. S. Jabade and S. R. Gengaje, "Literature review of wavelet based digital image watermarking techniques," Int. Journal of Computer Applications, Vol. 31, no.1, 2011, pp. 28-35.
- [6] Baisa L Gunjal and Suresh N Mali, " MEO based secured, robust, high capacity and perceptual quality image watermarking in DWT-SVD domain, " SpringerPlus , 2015, pp. 1-16.
- [7] R. Liu, and T. Tan, "An SVD-based watermarking scheme for protecting rightful ownership," IEEE Transactions on Multimedia,vol. 4, 2002, pp. 121–128.
- [8] Navdeep Goel , Gurwinder Singh , " Study of Wavelet Functions of Discrete Wavelet Transformation in Image Watermarking , " An International Journal of Engineering Sciences, 2016, pp.154-160 .
- [9] Elham Moeinaddini, roya ghasemkhani,"A novel image watermarking scheme using blocks coefficient in DHT domain", international symposium on artificial intelligence and signal processing, IEEE, 2015, pp. 159-164.

# Comparative Analysis of Penetration Testing of LAN Using Various Attacks on Windows 7 & Ubuntu

Jiwanjot Kaur Buttar<sup>1</sup>, Himshikha Rahi<sup>2</sup>, Harmandeep Singh<sup>3</sup>

Department of Computer Engineering, Punjabi University, Patiala, 147001

jiwanjot91@gmail.com, hshikha46@gmail.com, harmanjhajj@yahoo.co.in

**Abstract-** Organizations use a large number of systems which holds sensitive data. So, its security has become a major issue for the organizations. One of the solutions for testing the security of an organization is penetration testing. It is a series of activities undertaken to identify and exploit security vulnerabilities. On a daily basis extremely skillful hackers violate the security and acquire the benefit of vulnerabilities to access the classified data. In this paper, for the purpose of comparison between Windows 7 and Ubuntu Operating systems, some attacks are implemented using Kali Linux to check the vulnerabilities of the operating systems connected in a network through Ethernet. Attacks performed are Denial of Service, Code injection attack, Database sniffing, Man in the Middle attack and DNS spoofing. Various graphs are plotted to show the comparison between operating systems for different attacks.

**Keywords-** Vulnerability, network security, Penetration Testing, Attacks, Denial of Service, Code injection attack, Database sniffing, man in the middle attack, DNS spoofing, Kali Linux.

## I. INTRODUCTION – PENETRATION TESTING

Penetration Testing is a process by which an organization comes to know about the loopholes in their security as a pen-tester is able to breach it. For this purpose, the organization appoint experienced pen-tester or a team so that they can have detailed knowledge about the module that are to be patched to make the system secure. Before starting a pen-test, a contract is made between the organization and the pen-tester regarding the scope of the test i.e. whether the test would be a black box, white box or gray box. In a black box test, it is the responsibility of the tester to collect all the information available to public about the organization and then make the information collected useful to perform the test. In white box testing, the tester is provided with the internal security information i.e. architecture of the system, security applied etc. So, the only responsibility of the tester is to perform the test. Gray box testing is the combination of black box and White box testing i.e. partial-disclosure of the information is given to the tester [3].

Penetration testing provides the reality of the situation of an organization so the needed security measures can be implemented and the system can be made almost impenetrable for the hackers who want to use the data of the

organization in a bad way which saves the organization from any damage caused to their reputation and finance [1][3].

## II. ATTACKS

In computer science an attack can be described as an attempt to destroy, steal, alter, and disable the data and functions of an institution. Also to gain unauthorized access to these functions and data is considered an attack [8]. The attacks can be categorized into two types:

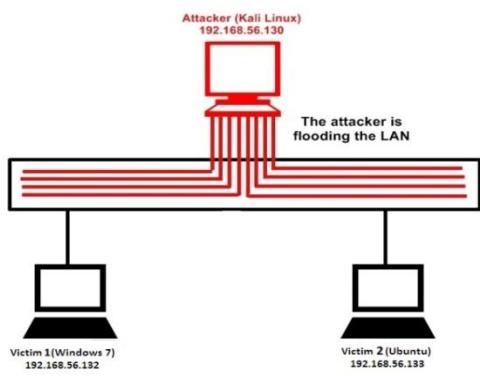
*Active Attacks:* the attacks in which some type of modification is done to the communication between the communication parties by the attacker are active attacks. It is difficult to prevent these attacks; however these can be detected with some effort. Modification, fabrication and interruption can be caused by these attacks [9].

*Passive Attacks:* Passive attacks are those attacks in which the attacker listens to the communication taking place between the two parties. These attacks are very hard to detect as no action is taken by the attacker. No modification occurs during this attack. Mostly prevention methods are adopted for this attack as these are hardly detected [9].

## III. DIFFERENT TYPES OF ATTACKS AND THEIR METHODOLOGIES

### A. Denial Of Service

Denial of Service has become a major threat in networking field. Attack in which the network is flooded with meaningless bogus traffic designed to bring the network down is Denial of Service Attack. With the help of these attacks, hackers deprive the services provided to the intended users which were expected. Everyday new ways of implementing DOS attacks are being designed by the malicious parties or hackers [12].



There are basically two possible forms of DOS Attacks which are as follows:

In Crash Services, the attacker would cause the services to collapse or it will lead the network to broadcast some unnecessary messages in order to offer fake services to a bunch of clients or to all its clients. Flood services are the most common DOS attack used by the attackers in order to halt the services provided by the Organization.

Fig 1: Denial of Service

In this attack, the attackers overflow the network by sending unnecessary traffic to the network so that the authorized users are denied the services provided by the organization or the users connected to that specific network [11] [12].

For implementing Denial of Service Attack firstly, open a New Terminal on Kali Linux. Find IP addresses of all the systems attached on Ethernet using nmap. Choose the IP address you want to flood. Flood the Victim IP address using hping3.

#### B. Code Injection

Code injection is the abuse of a PC bug that is brought on by handling invalid information. Injection is utilized by an attacker to present (or “infuse”) code into a vulnerable PC program and changes the course of execution. It can be implemented in two ways:



Fig 2: Code Injection

For implementing code injection attack by open ports firstly open metasploit console in a new terminal. Then use the exploit found from the list. Find the IP addresses of systems connected to Ethernet. Now set the remote host as victim IP addresses. Find the list of payloads and set the suitable one. And start the exploitation.

For implementing code injection attack by transferring

a file firstly start metasploit console. Find a list of exploits and use the suitable one. Set the local host and local port. In a new terminal, generate a payload in any format and transfer it to the victim machine. Now set the payload from the list found. Open the vnc player and exploitation will be started whenever the payload file is opened on the victim computer.

#### C. Database Sniffing

In this type of attack, the attacker is able to fetch all the information from the database available on the any computer on a LAN network. Firstly the attacker breaks into a system with the help of metasploit tool and then with help of meterpreter locates the file on the victim's machine. Once the required file is located the attacker uses the command to download the required file and view its content to get any benefit from it [16].

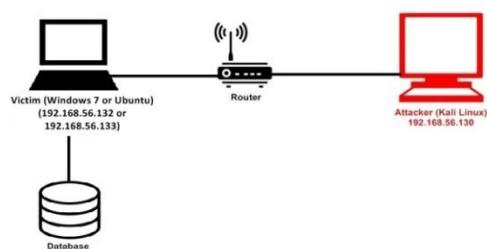


Fig 3: Database Sniffing

For implementing database sniffing attack firstly, start

metasploit console. Find a list of exploits and use the suitable one. Set the local host and local port. In a new terminal, generate a payload in any format and transfer it to the victim machine. Now set the payload from the list found. Check the time from which the victim machine is idle. Use Meterpreter to open the root folder and then list all the contents in it. Then start exploring different files to find the required file. Once the required file is found give the download command.

*D. Man in the Middle Attack*

In Man in the Middle Attack, the attacker gains access to the conversation between the two communicating parties and masquerades itself to both ends and steals the information being exchanged between them for exploitation purposes. It permits the attacker to change the confidential information, halting any conversation taking place or sending false messages to other parties by impersonating itself. It is abbreviated as MiM, MITM, MIM or MitM [18].

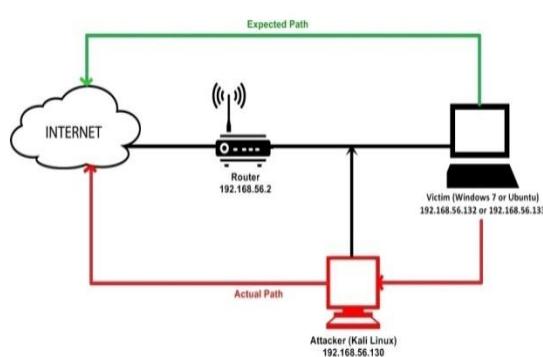


Fig 4: Man in the Middle Attack

For implementing Man in the Middle Attack firstly, open a New Terminal on Kali Linux. Find IP addresses of all the systems attached on Ethernet using nmap. In a new terminal, check whether IP forwarding is enabled or not. If not, then enable it. Redirect the traffic from port 80 to port 8080 in the pre-routing table. Find the interface and Gateway for the Victim IP address. Perform the spoof by portraying to the victim that you are a router and to the router that you are victim. Now listen to the intended port and write all the session details on a log file.

*E. DNS Spoofing Attack*

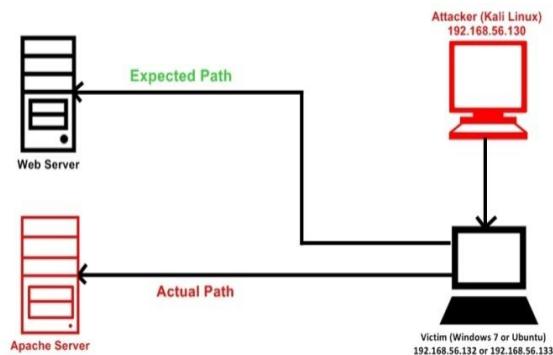


Fig 5: DNS Spoofing

DNS Spoofing also known as DNS Poisoning or pharming attack is a type of attack in which data is altered on the domain name system (DNS) of the victim's computer which leads the name server to redirect to the wrong IP address, thus the traffic is diverted to the attacker's computer. DNS (Domain Name System) is a special server computer which makes it possible to convert the web addresses which can easily be remembered by the users into the IP addresses which are used by the computers to connects to the respective web server [22].

Basically DNS helps the systems to map a route between domain names and their respective IP addresses. For this a procedure named the resolver is called by the application program which maps the web name into an IP address, by passing the web name as a argument. A UDP packet is dispatched to the local DNS Server which in returns finds the IP address corresponding to the web name and returns it to the caller [21][22].

For implementing DNS Spoofing Attack firstly, In a new terminal, initialize postgre. Now start the Metasploit Console. Find the exploits related to the web browser. Then use the exploit earlier found and set Payload for it.

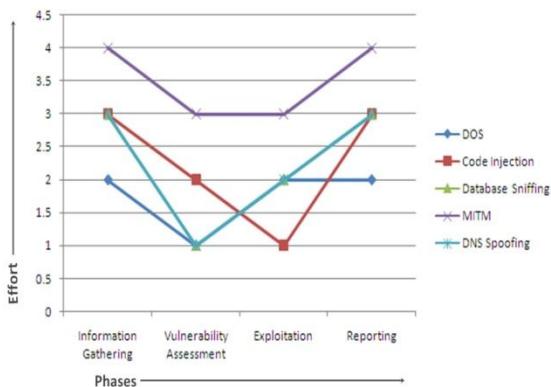


Fig 6: Graph of Effort Vs Phases

The graph shown in Fig. 6 elaborates the effort which is needed to perform the different phases of penetration testing. Different colours shows different attacks which are performed for this paper and the degree of effort being done for each phase.

## V. RESULTS AND DISCUSSIONS

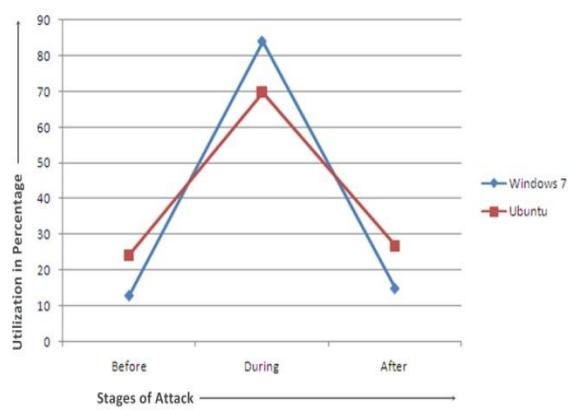


Fig 7: Graph of CPU Utilization vs. Stages

Now set Local Host, Service Host and Service Port. And set the uniform resource identifier path owned. Start exploitation. Now in etter.dns file, add local host under Vim heading and save the changes. In Index html file, change url to local host and port addresses. Now start Apache2 server and ettercap. Start the sniffing and search for hosts. And add the targets and dns\_spoof plugins in the ettercap. You will start seeing that on the victim window whenever user opens any page it will be directed to Apache.

Denial of service attack makes the network busy by sending meaningless bogus traffic so that authorized users are unable to access the facilities provided by the network. This attack can be implemented in two ways: either with crash services or with flood services. In this paper, this attack is implemented using flood services. Hping3 is used to flood the LAN network. In the figure 7 it shows the degree of utilization in terms of CPU when a Denial of service attack occurs. For Windows 7 the utilization increases from 13 to 84% and in Ubuntu the utilization increases from 24.4% to 69.7%. So the total increases in windows 7 the increase is of 71% and in Ubuntu the increase is 43.5%.

In Code injection attack, a bug is infused to the victim's machine so as to get access to that machine for malicious purposes. In this, a malicious file is sent which can be in any format i.e. pdf, doc, exe etc. It is performed in two ways i.e. by open ports and by transferring a bug. Firstly, select the open port through which payload is sent and the victim window is exploited using command shell so as to get control of the command prompt of the victim window. Second way is to transfer a file into the victim's machine. It is performed with the help of metasploit by creating a malicious file that is to be sent to the target machine using a pen drive. When this malicious file is opened on the target machine while the exploit is running, the attacker gets the access to the graphical interface of the target machine. Attack using the open port does not require any action on the victim machine where as in

second method; it is required for the victim to run the file. For code injection attack, none of the method was successful on ubuntu whereas only one was successful on Windows 7.

Database sniffing is used to get all the information residing on a database as the data present on the database is very confidential and if leaked can cause huge loss of money and reputation for the firm. It is implemented with the help of metasploit and meterpreter tool. After gaining access by the meterpreter the attacker tries to find the database which is needed and have some value. Once attacker is able to locate the confidential database he downloads it and can use it to gain benefits in terms of money or reputation. The database sniffing attack was successful on Windows 7 but failed for Ubuntu.

Man in the middle attack helps the attacker to gain the access to the conversation between two parties. The attacker situates himself in between the victim window and the router so as to gain access. First of all we have to check if IP forwarding is enabled or disabled, if disabled enable it. Then in the routing table, the IP's of the victim and router are spoofed from the destination port 80 to 8080 so that the control can be moved to the target. This is implemented with the help of arpspoof command in which we selects the target network i.e. eth0. A log file is created on the attacker's computer named as 'sslstriplog.log' as shown in the above figure in which all the conversation between the communicating parties is recorded. For MITM Attack, Ubuntu is more secure than Windows 7 as the attack on Ubuntu failed because the attack was not able to hinder the security layer in the HTTP protocol whereas as in Windows 7 the attack successfully removed the security layer from the HTTP protocol.

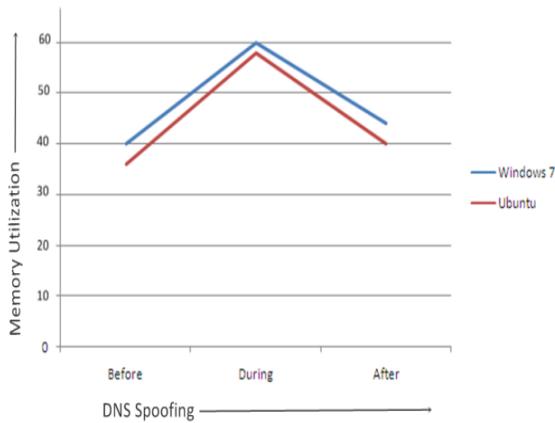


Fig. 8: Graph of Memory Utilization vs Stages

DNS spoofing attack is used to alter the domain name in the domain name system of the victim's machine which makes the name server to redirect the victim machine to wrong IP address. It is performed with the help of metasploit which uses the reverse tcp payload used by the attacker to set up a listener box and then connecting to the victim's machine to listen the active session. After this, the local host i.e. 192.168.56.130, service host i.e. 192.168.56.130 and the service port i.e. 4448 are specified. The uniform resource identifier path is owned so that the victim machine gets redirected to the apache server. Also, ettercap is used to perform the dns spoof.

In DNS Spoofing, we are basically poisoning the Cache memory. So, we compared the results for Windows 7 and Ubuntu on the basis of change in the physical memory or the cache memory before and after the attack. The increase in the usage of physical memory in Ubuntu is from 45.6% to 58.3% i.e. the difference caused by the DNS poisoning in the usage of memory for Ubuntu is 12.7%. Now, for Windows 7 the increase is from 46% to

61% i.e. the difference in the usage of memory for Windows 7 before and after the attack is 15%. So, the increase in the usage of physical memory in Windows 7 is 2.7% than that of Ubuntu.

TABLE I  
SUCCESS OF ATTACKS ON DIFFERENT OS

Attacks	Windows 7	Ubuntu
DOS	Yes	Yes
Code Injection (Ports)	No	No
Code Injection (File Transfer)	Yes	No
Database Sniffing	Yes	No
MITM	Yes	No
DNS Spoofing	Yes	Yes

## VI. CONCLUSION

Security is very important for the organization now days. A lot of money is being invested in making the organization's system impenetrable. We have concluded that Ubuntu is more secure than the windows 7, for Denial of Service Attack, Code Injection Attack and Database Sniffing Attack. Denial of service attack was successful on Ubuntu and the increase in utilization was only 43.5% which was way less than that of Windows 7 (i.e. 71%). For code injection attack, none of the method was successful on ubuntu whereas only one was successful on Windows 7. The database sniffing attack was successful on Windows 7 but failed for Ubuntu. For MITM Attack, Ubuntu is more secure than Windows 7 as the attack on Ubuntu failed because the attack was not able to hinder the security layer in the HTTP protocol whereas as in Windows 7 the attack successfully removed the security layer from the HTTP protocol. During DNS Spoofing Attack, the availability of the physical memory or cache memory was decreased by 15% whereas in Ubuntu the availability of cache memory was decreased by 12.7%. So, the availability in Windows 7 is 2.3% more than in Ubuntu.

## REFERENCES

- [1] Sachin Umrao, Mandeep Kaur and Govind Kumar Gupta, "Vulnerability assessment and penetration testing", International Journal of Computer & Communication Technology ISSN (PRINT): 0975-7449, Volume-3, Issue-6, 7, 8, 2012
- [2] Ankita Gupta, Kavita and Kirandeep Kaur, "Vulnerability assessment and penetration testing", International Journal of Engineering Trends and Technology-Volume4Issue3-2013 ISSN: 2231-5381
- [3] Divya Sharma, Oves Khan, Kanika Aggarwal, Preeti Vaidya, " A new approach to prevent ARP spoofing", International Journal of Innovative Technology and Exploring Engineering (IJITEE)ISSN: 2278-3075,Volume -3, Issue - 1, June 2013
- [4] Manju Khari and Neha Singh, "An Overview of Black Box Web Security Scanners", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014, ISSN 2277 128X
- [5] Victor Velasco, "Introduction to IP spoofing", SANS Institute InfoSec Reading Room
- [6] Roopam and Bandana Sharma, "Review Paper on Prevention of DNS Spoofing", International Journal of Engineering and Management Research, Volume -4, Issue-3, June-2014, ISSN No.: 2250-0758
- [7] Laxman Vishnoi and Monika Agarwal, "Session hijacking and its countermeasures", International Journal of Scientific Research Engineering & Technology (IJSRET), Volume 2, Issue 5, pp 250-252, August 2013, www.ijssret.org, ISSN 2278-0882

- [8] Kefei Cheng, Tingqiang Jia and Meng Gao, "Research and Implementation of Three HTTPS Attacks", JOURNAL OF NETWORKS, VOL. 6, NO. 5, MAY 2011
- [9] Praveen Kumar Mishra, "Analysis of MITM Attack in Secure Simple Pairing", Journal of Global Research in Computer Science, Volume 4, No. 2, February 2013, ISSN: 2229-371X
- [10] Aileen G. Bacudio, Xiaohong Yuan, Bei-Tseng Bill Chu, Monique Jones, "An overview of penetration Testing", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.6, November 2011
- [11] Minakshi Bhardwaj and G.P. Singh, "Types of Hacking Attack and their Counter Measure", International Journal of Educational Planning & Administration. Volume 1, Number 1 (2011), pp. 43-53
- [12] Konstantinos Xynos, Iain Sutherland, Huw Read, Emlyn Everitt and Andrew J. C. Blyth, "Penetration Testing and Vulnerability Assessments: A Professional Approach", 1<sup>st</sup> International Cyber Resilience Conference, Edith Cowan University, Perth Western Australia, 23<sup>rd</sup> August 2010
- [13] Imtiyaz Ahmad Ione and M d. Ataullah, "A survey on various solutions of ARP attacks", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 2, February 2013 ISSN: 2277 128X
- [14] Simar Preet Singh and A Raman Maini, "Spoofing attacks of Domain Name System Internet", National Workshop-Cum-Conference on Recent Trends in Mathematics and Computing (RTMC) 2011 Proceedings published in International Journal of Computer Applications® (IJCA)
- [15] Abhishek Kumar Bharti and Manoj Chaudhary, "Detection of Session hijacking and IP spoofing using sensor nodes and cryptography", OSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727 Volume 13, Issue 2 (Jul. -Aug. 2013), PP 66-73
- [16] Pushpendra Kumar Pateriya and Srijith S. Kumar, "Analysis on Man in the Middle Attack on SSL", International Journal of Computer Applications (0975– 8887) Volume 45–No.23, May 2012
- [17] Suraj S. Mundalik, "Penetration Testing: An Art of Securing the System (Using Kali Linux)", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 10, October-2015 ISSN: 2277 128X
- [18] Ashwani Garg and Shekhar Singh, "A Review on Web Applications Security Vulnerabilities", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 1, January 2013 ISSN: 2277 128X
- [19] Daya Shankar Singh, "Session Hijacking and its Protection", Research Spectra, Volume 1, Issue No. 2-3, August-December 2015, ISSN 2394 9805
- [20] K.Bala Chowdappa , S.Subba Lakshmi , P.N.V.S.Pavan Kumar, " Ethical Hacking Techniques with Penetration testing", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3389-3393, ISSN 0975 9646
- [21] Sheetal Bairwa, Bhawna Mewara and Jyoti Gajrani, "Vulnerability Scanners: A Proactive Approach to Assess Web Application Security", International Journal on Computational Sciences & Applications (IJCSA) Vol.4, No.1, February 2014
- [22] Roi Saltzman, Adi Sharabani, "Active Man in the Middle Attacks", A white paper from IBM Rational Application Security Group
- [23] Franco Callegati, Walter Cerroni and Marco Ramilli, "Man in the Middle Attack to the HTTPS Protocol", Published by IEEE Computer Society, 1540-7993/09
- [24] Thawatchai Chomsiri, "Sniffing Packets on LAN without ARP Spoofing", Third 2008 International Conference on Convergence and Hybrid Information Technology
- [25] Italo Dacosta, Mustaque Ahamed and Patrick Traynor, "Trust No-one Else: Detecting MITM Attacks against SSL/TLS without Third-Parties
- [26] Yu Xi, Chen Xiaochchen and Xu Fangqin, "Recovering and Protecting against DNS Cache poisoning attack", 2011 International Conference of Information Technology, Computer Engineering and Management Sciences, 978-0-7695-4522-6/11
- [27] Gopi Nath Nayak and Shefalika Ghosh Samaddar, "Different Flavours of Man in the Middle attack, Consequences and Feasible Solutions", Published by IEEE Computer Society in 2011, 978-1-4244-5540-9/10
- [28] Nuno Antunes and Marco Vieira, "Penetration Testing for Web Services", Published by IEEE Computer Society in 2014, 0018-9162/14
- [29] Pulei Xiong, Liam Peyton, "A Model-Driven Penetration Test Framework for Web Applications", 2010 Eighth Annual International Conference on Privacy, Security and Trust, 978-1-4244-7550-6/10

# Block Based Image Steganography using LSB and Identical approach

Gursewak singh

Research scholar, CE Section

Yadawindera college of engineering  
Talwandi sabo, 151302, Punjab, india  
Gursewak318@gmail.com

Manoj kumar

CE Section

Yadawindera college of engineering  
Talwandi sabo, 151302, punjab, india  
Ermanojchaudhary@gmail.com

**Abstract-** Steganography plays a very imperative role in secret communication. Many possible techniques are used to embed confidential information in digital images, the least significant bit (LSB) technique has very widely used. In this paper the proposed technique ,a new steganography technique is being developed to hide data in image using pixel based algorithm. Also, to make the algorithm more undetectable data is divided into segments and image into blocks and a data segment is embedded into an image block where it effects , the least image quality. The experimental results prove that the quality of stego image using the proposed algorithm. For the verification of the results, peak signal-to noise (PSNR) and mean square error (MSE) are calculated.

**Keywords**— LSB; stego image; secret image;Arnold's transformation; identical bits.

## I. INTRODUCTION

Steganography is the process of concealing information in a carrier such as text, image, voice, video, or protocol. Digital images are one of the common and most popular ones due to their frequency on the Internet and high capacity of data transmission without degrading effect on images quality [1]. It is a high security technique for long data transmission. For many years it has been considered that security in cryptographic algorithms is directly related to the complexity of the mathematical operations that define the core of the encoding process. However, research using hardware-aided reverse engineering has continuously demonstrated that every cryptographic algorithm has a relatively short lifecycle, defined by the evolvement of computational power. A secure communication system is reliable as long as the cryptographic algorithm on top of which it was built is reliable [3].

In second section of this paper we have written related work to proposed technique and proposed work, experiment and results and conclusion are binded in sections 2,3 and 4 respectively.

## II. LITERATURE REVIEW

Harpal et.al [1] proposed a new steganography techniques for the colored image. Their are many steganography techniques like LSB,DCT, pixel based etc, but these technique have many problems. To improved technique used 2 2 4 LSB insertion three plans images and get best results. R.kumar et al [3] proposed a new method based on parity of the pixel in odd and even case. First of all message is encrypted by vernam cipher algorithm and apply LSB-S method of pixel and XOR operation, after this store the results and verify on different parameters to good results. M.devi et al [3] proposed a new method based on parity of the pixel in odd and even case. First of all

message is encrypted by vernam cipher algorithm and apply LSB-S method of pixel and XOR operation, after this store the results and verify on different parameters to good results.

### III. PROPOSED WORK

#### A.2-bit identical

This method is used to hide the secret message and pixel of the image bits are analysed one by one if identical value finding between the pixel then hide the secret message on those bits.

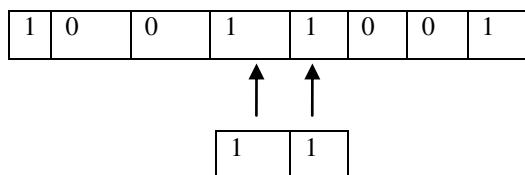


Figure 1. Identical technique

#### B. Least significant bit technique

The most significant technique in steganography. It is used to hide the secret data in grey scale and colored images on its binary coding. fig 1.2 shows the LSB technique and shows the pixel of secret message bits. Algorithm shows the results by shifting right most two bits of LSBs of the pixel [1].

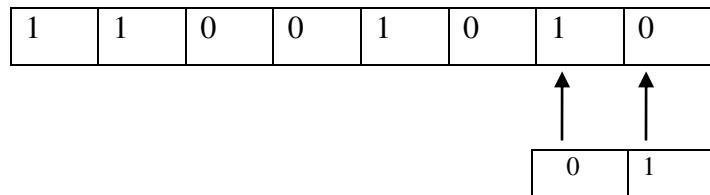


Figure 2. Least significant bit hiding technique

#### C. ARNOLD'S CAT MAP

Arnold's cat map (ACM) or Arnold transform (AT), proposed by Vladimir Arnold in 1960, is a chaotic map which when applied to a digital image randomizes the original organization of its pixels and the image becomes imperceptible or noisy. However, it has a period p and if iterated p number of times, the original image reappears [11].

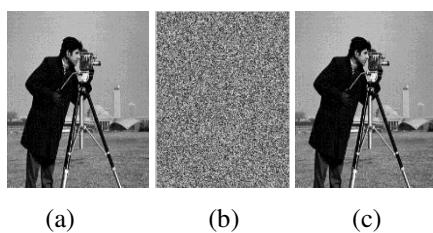


Figure 3. (a) original secret image (b) encrypted secret image (c) decrypted image.

In this figure shows that having original image encrypted with secret image apply ACM random iteration of 20 times which obtain an encrypted image. we can retrieve the original image by applying inverse Arnold's transformation.

*TABLE 1.*

*Steps of hiding Proces:*

1.	Read 256x256 color cover image.
2.	Divide color image into 16 images into 64x64 bits.
3.	Select four blocks of image to hide data.
4.	Find RGB of each of four blocks.
5.	Convert green component of each block into bits, $64 \times 64 \times 8 = 32768$ .
6.	Divide these bits into 4096 groups of 8 bits each.
7.	Read 64x64 grey scale secret image and apply arnold's transformation to get an encrypted image.
8.	Convert encrypted image into bits $64 \times 64 \times 8 = 32768$ .
9.	Divide these bits into 4 blocks having 8192 bits each,then in these groups divide bits into 4096 groups of 2 bit each.
10.	Select group of bits of block 1 we have chosen for hiding data and in first group of bits of secret image,apply proposed algorithm (2-bit identical+LSB),save bit positions in an array(4 arrays for 4 blocks).
11.	Repeat step 10 all of four blocks we have selected for hiding data .
12.	Combine all of 16 blocks of image(12 blocks + 4 updated blocks).
13.	Show stego image and calculate PSNR.

TABLE 2

*Steps of Retreveng process:*

1.	First of all Read the stego image of 256x256 bits.
2.	Divide the stego image into 16 blocks of each bits.
3.	Then select 4 blocks from where we have already fixed
4.	Apply RGB on each of 4 blocks of bits.
5.	Convert RGB's green component of $64 \times 64 \times 8 = 32768$ .
6.	Divide each of the blocks with 8 bits which is =4096.
7.	Repeat 5,6 step until for each 4 blocks.
8.	Select first group and having an array that was defined in first embedding process.
9.	Select position of bits from an array and extract at those positions from corresponding blocks
10.	Repeat the 8,9 steps each of 4 block.
11.	After this combine those extracted bits to get encrypted secret image.
12.	Apply ARNOLD'S inverse transformation, show the extracted image.

#### IV. EXPERIMENTAL RESULTS

The proposed technique has been implemented for cover image. With the help of this stego image has been evaluated. For implementation we have several colored cover images of 256x256 pixel then divide into different blocks of each pixel of 64x64 total 16 blocks. Grey scale image of 64x64x8 bits. Then proposed scheme presented digital image that employ hybrid technique of 2-bit identical. PSNR Peak Signal Noise Ratio is used to measure the quality between cover image and stego image within size.



Figure 4. (a) peeper (b) lena (c) lake (d) cameraman.

#### D. Peak Signal Noise Ratio

PSNR is used to measure the quality between cover image and stego image within size.

$$\begin{aligned}
 PSNR &= 10 \cdot \log_{10} \left( \frac{MAX_I^2}{MSE} \right) \\
 &= 20 \cdot \log_{10} \left( \frac{MAX_I}{\sqrt{MSE}} \right) \\
 &= 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE) \dots\dots (1)
 \end{aligned}$$

In this formula I is the value of each pixel. Greater the value of PSNR better the quality of image. MSE ( Mean Square Error) also a parameter used to test the performance of proposed algorithm

$$MSE = \frac{1}{MN} \sum_{j=1}^{M} \sum_{k=1}^{N} (x_{j,k} - x'_{j,k})^2 \dots\dots (2)$$

where, M and N denotes the total number of pixels in the horizontal and vertical dimensions of an image.  $x_{ij}$  represents the pixels in original image and  $y_{ij}$  represent the pixels of stego image.

#### E. Imperceptibility

To measure the imperceptibility of proposed method. PSNR and MSE are calculated for different data bits. Results are shown in the table.

TABLE 3.

*Shows images and their PSNR, MSE results values.*

Image	Original image	Stego image	PSNR	MSE
Pepper			62.2661	0.0114
Baboon			61.6711	0.01163
Lena			62.6708	0.0031
Lake			61.6607	0.0211
Home			61.9366	0.0147

#### F. Hiding Capacity

The hiding capacity of proposed method is high up to 2bpp. Several images of 256 x 256 pixels are taken and data is hidden in these pixels. The total capacity of the proposed method is 131072 bits. The capacity is calculated as given below:

$$\text{capacity of 1 blocks} = 64 \times 64 \times 2 = 8192$$

$$\text{capacity of 4 blocks} = 64 \times 64 \times 2 \times 4 = 32768$$

$$\text{capacity of 16 blocks} = 64 \times 64 \times 2 \times 16 = 131,072$$

$$\text{Total capacity} = 131,072 .$$

## V. CONCLUSION

In above proposed scheme using 2-bit identical approach with LSB, we have achieved PSNR value above 60 for different images shown in table 3 which implies that our proposed scheme results in good perceptual quality of stego image. Arnold's transformation provides security by encrypting the secret image. For future work, it can be extended to be used with other techniques like DHT,DWT etc. Further it can be extended for video steganography.

## *References*

- [1] A.singh, H.singh, "An Improved LSB based Image Steganography Technique for RGB Images," international conference IEEE, vol.978-1-4799-6085-9/15/\$31.00 ©2015 IEEE.
- [2] M.devi, N.sharma, "Improved detection of LSB Steganography Algorithms in Color and Gray Scale Images," UIET Panjab University Chandigarh, vol. 978-1-4799-2291-8/14/\$31.00 ©2014 IEEE.
- [3] K.joshi, R.kumar, "A New LSB-S Image Steganography Method Blend with Cryptography for Secret Communication", Third International Conference. Vol. 978-1-5090-0148-4/15/\$31.00© 2015 IEEE .
- [4] Shashikala Channalli, Ajay Jadhav , "Steganography An Art of Hiding Data", International Journal of Computer Science and Engineering
- [5] D. Singla and M. Juneja, "Hybrid edge detection-based image steganography technique for color images," in Intelligent Computing, Communication and Devices, ser. Advances in Intelligent Systems and Computing, vol. 309, pp. 277–280,2015
- [6] C.-K. Chan and L. Cheng, "Hiding data in images by simple {LSB}g substitution , " Pattern Recognition, vol. 37, no. 3, pp. 469 – 474, 2004.
- [7] Rajput, A.S., Mishra, N., Sharma, S.: Towards the growth of image encryption and authentication schemes. In: IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI) (2013).
- [8] S.M. M. Karim, M.S. Rahman, and M.I. Hossain, "A New Approach for LSB Based Image Steganography using Secret Key.", Proceedings of 14th International Conference on Computer and Information Technology, IEEE Conference Publications, pp 286 – 291, 2011.
- [9] X. Qing., X. Jianquan and X. Yunhua., "A High Capacity Information Hiding Algorithm in Color Image.", Proceedings of 2nd International Conference on E-Business and Information System Security, IEEE Conference Publications, pp 1-4, 2010.
- [10] N. Ghoshal and J. K. Mandal, "A Steganographic Scheme for Colour Image Authentication (SSCIA)", Proceedings of International Conference on Recent Trends in Information Technology(ICRTIT), India, IEEE Conference Publications, pp. 826 – 83, 2011.

- [11] Minati Mishra, Ashanta Ranjan Routray, Sunit Kumar, "High Security Image Steganography with Modified Arnold's Cat Map", International Journal of Computer Applications (0975 – 8887) Volume 37– No.9,pp.16-20,2012.

# Overview of Data mining Techniques and Tools

Ramanpreet Kaur  
Department of Computer Engineering,  
Punjabi University, Patiala  
Punjab, India  
Ramanpreetkaur234@outlook.com

Gaurav Gupta  
Department of Computer Engineering,  
Punjabi University, Patiala  
Punjab, India  
Gaurav.shakti@gmail.com

**Abstract**—The large amount of data is generated by different Information systems are considered of high business value, and data mining algorithms can be used to extract knowledge from data. This paper gives a way to review data mining in knowledge, technique, and application, including classification, clustering, association analysis, time series analysis and outlier analysis, and their application. Also throw light on tools used for data mining.

**Keywords**—data mining, techniques, clustering, classification, association analysis, applications, tools.

## I. INTRODUCTION

Data mining is used to discovering potentially useful pattern from large data sets. In this we apply algorithms to extract the hidden information from large data sets. The objective of any data mining process is to build an efficient predictive or descriptive model of large amount of data that only best fits or explains it, but is also able to generalize to new data[1].On the basis of the data mining definition, a typical data mining process includes the following steps

1. Data preparation: In this step data is get prepared for mining. This includes 3 sub steps: integrate data sources and clean data; extract some parts of data into data mining system; preprocess the data.
2. Data mining: In this step apply algorithms to the data to find the patterns and evaluate the patterns.
3. Data presentation: In this step visualize the data and present the mined knowledge to the user.

This paper is structured as follows. Section 2 is a survey of data mining techniques including classification, clustering, association analysis, time series analysis. Section 3 is a review of data mining applications and techniques applied to them. In Section 4 we discuss about the open source tools for data mining. Section 5 gives a conclusion.

## II. DATA MINING TECHNIQUES

Data mining techniques include classification, clustering, association analysis, time series analysis, outlier analysis.

### A. Classification

Classification refers to identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing instances whose category membership is known. Classification is an example of pattern recognition. The goal of classification is to accurately predict the target class for each case in data [2]. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risk [3]. There are many methods used to classify data, including

- Frame-based or rule-based expert system,
- Hierarchical classification,
- Neural networks,
- Bayesian network,

- Support vector machines.

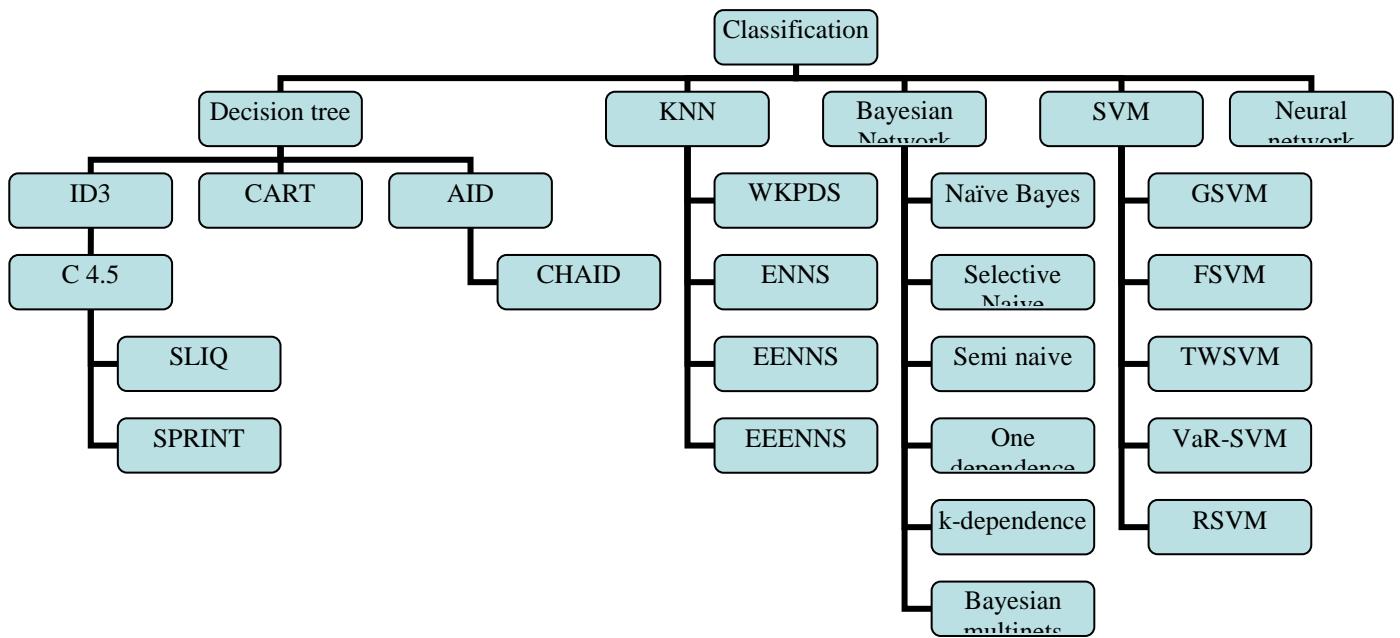


Figure 1: The structure of classification

- I. A decision tree is a tree like structure, where each node is denoted by rectangle and leaf nodes are denoted by ovals. Each leaf node has a class label associated with it. Iterative Dichotomies 3 or ID3 is a decision tree learning algorithm [4]. C4.5 is an improved version of ID3, it uses gain ratio as splitting criteria [5]. Difference between ID3 and C4.5 is that ID3 uses binary splits, C4.5 uses multiway splits. Supervised Learning In Quest (SLIQ) is fast and highly scalable, in SPRINT there is no constraint on large data sets [6]. CART is nonparametric decision tree algorithm, which produce classification or regression tree on the basis of response variable is categorized or continuous.
- II. The K-Nearest Neighbor algorithm is designed to find the nearest neighbor of the observed object. The improvements to k-nearest algorithm are Wavelet Based K-Nearest Neighbor Partial Distance Search (WKPDS) algorithm [7], Equal Average Nearest Neighbor Search (ENNS) algorithm [8], EENNS algorithm [9], EEEENNS algorithm [10] and other.
- III. Bayesian networks are directed acyclic graph; in this nodes represent random variable in Bayesian sense. Edges represent conditional dependencies.
- IV. Support Vector Machines algorithm is supervised learning model with associated learning algorithms that analyze data and discover patterns. It is used for text classification. Further modifications in SVM are GSVM, FSVM, TWSVM, Value-at-risk support vector machine [11], RSVM [12].

#### B. Clustering

Clustering algorithms divide data having similar properties into meaningful groups. There are many clustering methods such as hierarchical clustering method, partitioning algorithms, high dimensional and other (Shown in Figure 2).

- i. Partitioning algorithm divides the given data set into partitions where each partition represents a cluster. The popular partition algorithms are k-mean, k-medoids and their variations.

- ii. Hierarchical clustering forms a hierarchy tree of data sets. This method has two classification approaches, bottom-up approach and top-down approach.

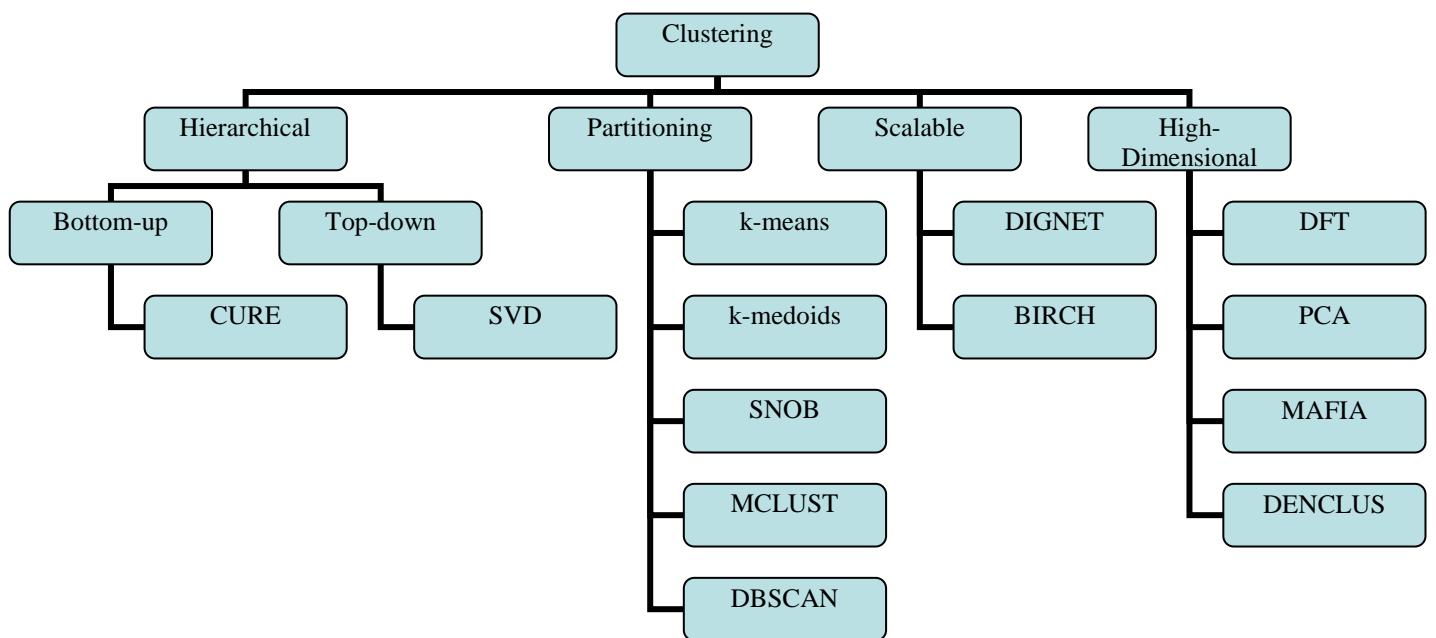


Figure 2 : The categorization of major clustering methods

### C. Association Analysis

Association rule mining deals with transaction data analysis, and discover rules which gives attribute value associated that occur frequently and help in generating qualitative knowledge which in helps in decision making. Association algorithms are

- Sequence( Priori based algorithm, Pattern growth algorithm )
- Temporal sequence(Event-oriented algorithm, Event based algorithm)
- Partition based
- Incremental mining and others.

These association algorithms are shown in Figure 3.

### D. Time Series Analysis

Time series data contains the sequence of values or events which occurred over repeated measurements of time. Time series analysis can be done by using following techniques

- Model based (ARMA, Time series Bitmaps)
- Non data adaptive (DFT, Wavelet functions)
- Data adaptive ( Data adaptive version of DFT, Data adaptive version of PAA)
- Similarity measures (Full Sequence matching, Subsequence matching )
- Indexing ( SAMs, MBR, X-Tree) and others shown in Figure 4.

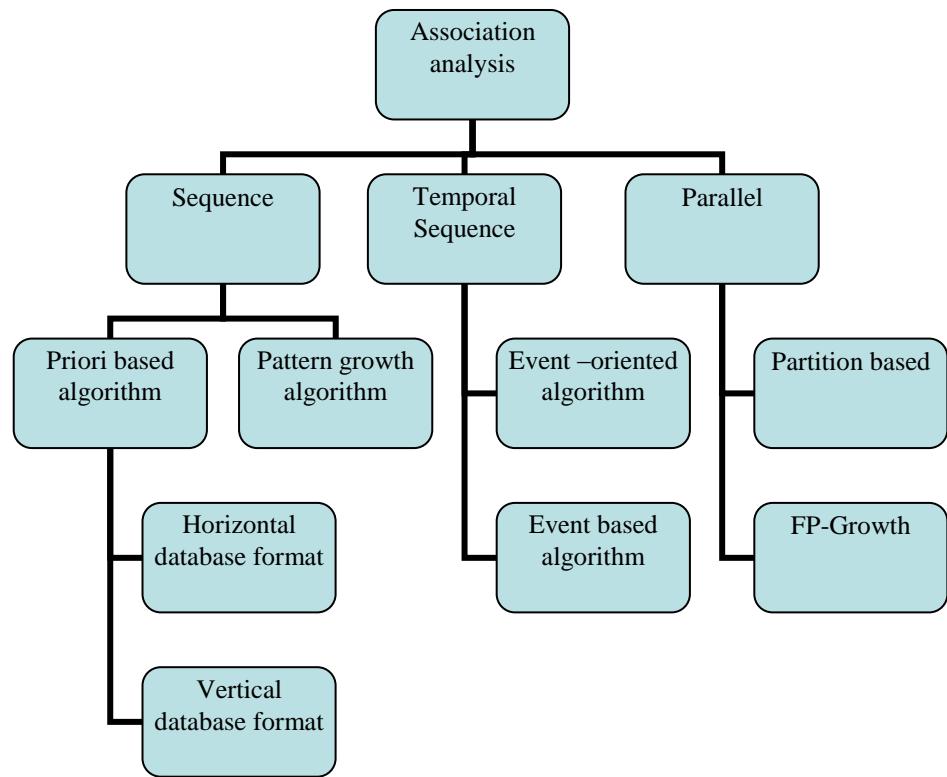


Figure 3: The categorization of Association Analysis algorithm

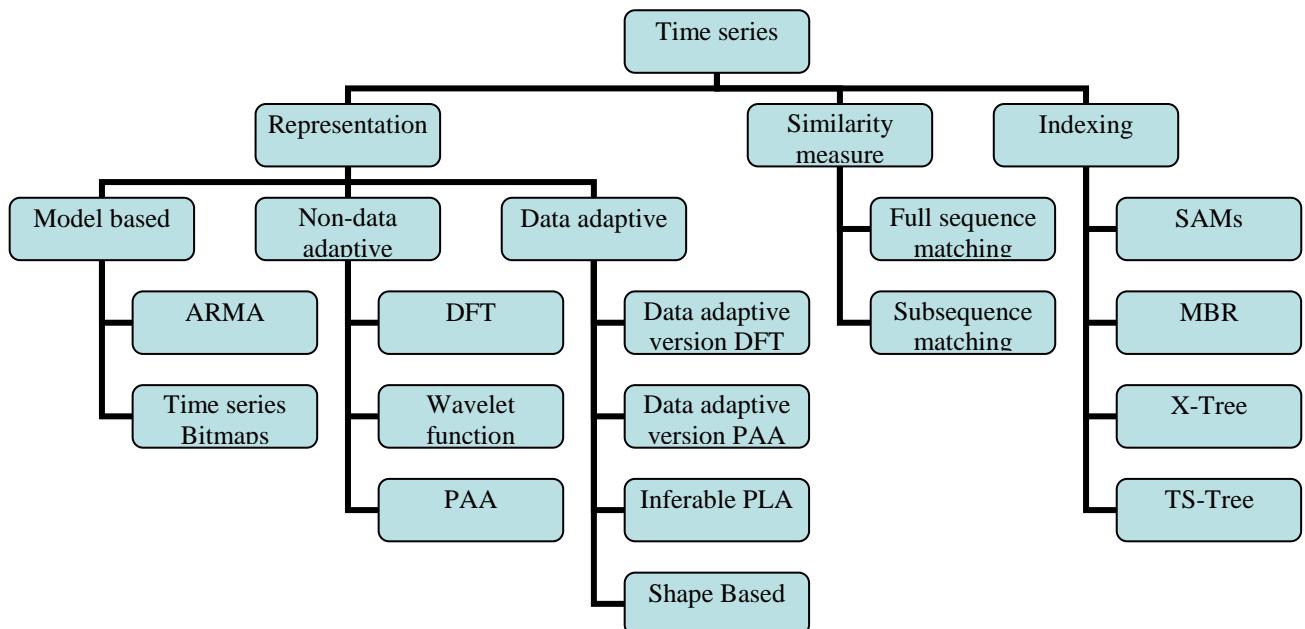


Figure 4 : The categorization of Time series algorithm

### III. DATA MINING APPLICATION

1. Data mining in e-Commerce. Data mining helps the businesses to understand the hidden patterns inside the past purchase transactions. These hidden patterns can help in planning new improved marketing campaigns. Techniques used in e-commerce data mining are clustering, association analysis.
2. Data mining in Industries. Data mining is being used in banking, retail and telecommunication industry. Classification, clustering, association analysis is used in these industries.
3. Fraud Detection use classification technique of data mining.
4. In Health Care data mining is used to provide better medical services. In health care application the data mining techniques are clustering, association analysis, and outlier analysis.
5. E-Governance. In governance applications like e-governance are used to improve the quality of government. In this classification, clustering, association analysis and time series analysis are used. By using data mining government can improve their performance.

**Table 1 Data mining applications**

Application	Technique used
eCommerce	Clustering, Association analysis
Banking	Classification, Clustering, association analysis
Fraud detection	Classification
Health Care	Clustering, Association analysis, Outlier analysis
eGovernance	Classification, Clustering, association analysis, Time series analysis

### IV. TOOLS FOR DATA MINING

There are many open source tools available with good performance and are easy to use. Most popular Data mining tools are

- Rapid Miner: Rapid miner support various type of data mining like text, image, web mining. It is implemented in java and can integrate processes developed by plug-ins.
- R (Rattle): R and Rattle(R Analytical Tool To Learn Easily) is used in research for statistical data analysis. R supports non-linear modeling, classical statistical tests, time-series analysis, classification, clustering, Partial Least Squares Regression, random forest. To use R tool user has to know R programming language.
- WEKA: WEKA (Waikato Environment for Knowledge Analysis) is a java library. It supports many machine learning algorithms such as data preprocessing, classification, regression, clustering, association rules and visualization.
- KNIME: Konstanz Information Miner was developed at the University of Konstanz in 2004 as a machine learning tool.

### V. CONCLUSION

With the rapid growth in information technology the data volume and data complexity is also increased. Data mining is a solution to the problem of large volume and complex data. There are many algorithms by which can be used to mine data from a large amount of data. In this paper there is brief introduction to simple and easy-to-learn open source tools like WEKA, Rapid Miner and KNIME. By using these tools user can easily perform data mining .Different applications of data mining are discussed in the

paper. And techniques which are being used in those applications are also discussed in the paper. There are some applications which can be further studied and work can be done on them.

#### REFERENCES

- [1] A.Mukhopadhyay,U.Maulik, S.Bandyopadhyay, and C. A. C. Coello, “A surevy of multiobjective evolutionary algorithms for data mining: part 1,” IEEE Transaction on Evolutionary Computation, vol. 18, no. 2, pp. 153-160,2011.
- [2] G.Kesavaraj and S.Sukumaran,” A study on classification techniques in data mining ” in Proceedings of the 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT ’13), pp. 1–7, July 2013.
- [3] S. Song, Analysis and acceleration of data mining algorithms on high performance reconfigurable computing platforms [Ph.D.thesis], Iowa State University, 2011.
- [4] J. R. Quinlan, “Induction of decision trees,” Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.
- [5] J. R. Quinlan, C4. 5: Programs for Machine Learning, vol. 1, Morgan Kaufmann, 1993.
- [6] J. Shafer, R.Agrawal, and M.Mehta, “SPRINT: a scalable parallel classifier for data mining,” in Proceedings of 22nd International Conference on Very Large Data Bases, pp. 544–555, 1996.
- [7] W.-J.Hwang and K.-W.Wen, “Fast kNN classification algorithm based on partial distance search,” Electronics Letters, vol. 34, no.21, pp. 2062–2063, 1998.
- [8] P. Jeng-Shyang, Q. Yu-Long, and S. Sheng-He, “Fast k-nearest neighbors classification algorithm,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. 87, no. 4, pp. 961–963, 2004.
- [9] J.-S. Pan, Z.-M. Lu, and S.-H. Sun, “An efficient encoding algorithm for vector quantization based on subvector technique,” IEEE Transactions on Image Processing, vol. 12, no. 3, pp. 265–270, 2003.
- [10] Z.-M. Lu and S.-H. Sun, “Equal-average equal-variance equalnorm nearest neighbor search algorithm for vector quantization,” IEICE Transactions on Information and Systems, vol. 86, no. 3, pp. 660–663, 2003.
- [11] P. Tsyurmasto, M. Zabarankin, and S. Uryasev, “Value-atrisk support vector machine: stability to outliers,” Journal of Combinatorial Optimization, vol. 28, no. 1, pp. 218–232, 2014.
- [12] R. Herbrich, T. Graepel, and K. Obermayer, “Large margin rank boundaries for ordinal regression,” in Advances in Neural Information Processing Systems, pp. 115–132, MIT Press, 1999.

# Hand shape based biometric system: A Survey

Nirpjit kaur<sup>1</sup>, Nirvair Neeru<sup>2</sup>

<sup>1,2</sup> Department of Computer Engineering, Punjabi University, Patiala, Punjab, India

<sup>1</sup> sandhunirpjit@gmail.com, <sup>2</sup> nirvair\_neeru@yahoo.com

**Abstract-** Modern security systems used in various sectors are emerged to use a person's biological characteristics such as fingerprint, eye, hand, palm etc which is referred to as biometrics. Automated biometric systems are most robust so act as a better option to the traditional systems. It widely used as a method for identification and verification based on physical or behavioural characteristics. Among various biometric characteristics, Hand geometry is one of the most well-known biometrics. Because of its ease of collectability, low resolution imaging, Public acceptance, low cost and low template size so have gained popularity in low to medium security systems. A survey has been presented in this paper describing various hand shape-based biometric system Technologies. Component modules used in commercial systems along with recent successful deployments will be reviewed. Finally, analyze the practical issues concerning performance and limitations and put light towards future research related to hand shape biometric systems.

**KEYWORDS:** Biometric, False Acceptance Rate, False Rejection Rate, Hand Geometry, Identification.

## I. INTRODUCTION

Biometrics is derived from Greek language where life described by Bio and metrics is measurement. So digital image processing deals with living beings who make use of application of biometrics. Identification and verification applications decide [9] how to use biometrics under different requirements. Nuclear power plant systems use to restrict access to critical systems. Border is controlled by governments. Immigration service's identify different nationalities for authentication and management. Educational institutions use to maintain controlled access, attendance and library. To keep track of drugs used in Health Care and in private sector & Clubs identification is performed for security and detection of banned visitors.

The paper is organized into following ways. Biometric characteristics are defined briefly in section II. Section III explains hand biometric system which is followed by sub sections describing work done in this particular area by various researchers, basic steps followed in hand geometry recognition and measurement parameters have been described for hand geometry recognition. Finally Conclusion and future scope has been discussed in last section IV.

## II. BIOMETRIC CHARACTERISTICS

Biometrics is a class divided into two characteristics physical and behavioural that means related to physical body features like face, retina, hand, fingerprint etc and latter related to mental state voice, gait,

signature and key stroke etc. These features are used for identification and verification purposes [11]. Visual Biometric includes shape of the ear, Features of eyes from retina and iris are used, Analyzing facial features or patterns, Use of the ridges and valleys (minutiae) found on the surface tips of a human finger and Features like thickness of palm, length of fingers etc from hand geometry are used for individual's identification. On the other hand behavioural Biometric includes voice analyzing information based on pitch as it is different for every person, Key stroke where speed, errors, time are computed. Unique style of every person's to sign so shape, speed, stroke, pen pressure and timing information are analysed and similarly unique style of walking. All these characteristics are depicted through a figure shown below named Figure1.

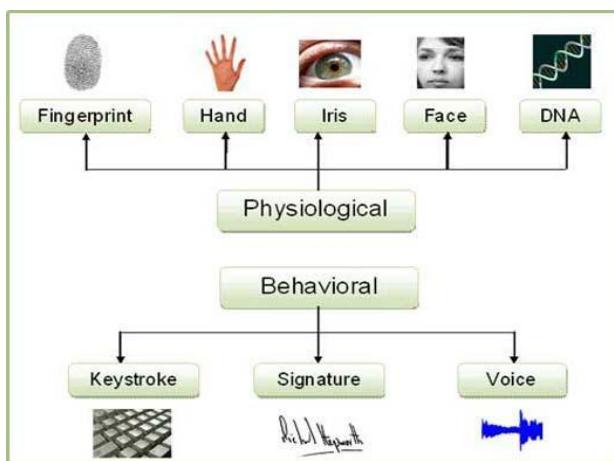


Figure 1. Biometric Characteristics

### III. HAND SHAPE BIOMETRIC SYSTEM

Hand shape geometry is one of the mostly used characteristic [5] as it can easily integrate other characteristics like palm print and fingerprints into an existing system and inexpensive sensors can be used to capture low resolution images in a very user convenient way. Hand geometry take measurements with the use of palm including its shape and size and fingers including their length and width[1]. Environmental factors like weather, skin texture may less likely to affect the performance so easy to employ.

Various researchers have researches and work in field of hand geometry biometric system. Review of researcher's studies has been described in this particular section which is as follows.

Nidhi Saxena et.al [6] proposed a method in their research about personal verification and identification using hand geometry in which features are extracted from fingers through its length and width and also from palm's width. In this technique, testing is performed using different distance functions which are in total six in number. An image is acquired using a camera and converted to a gray scale image which is noise free due to median filter. A binary image is obtained using ostu's method of thresholding. Further extract the features which are the lengths of each finger, the widths of each finger at 3 locations and the width of the palm. This results in 21 features all together. Distance functions are utilized in the matching

process to help differentiate the authorized and unauthorized persons. Matching algorithm compares and test different functions. In their approach the Test data obtained were from different users. Distance function named S1 gives the best results in both verification and identification as compare to other functions. By combining correlation and distance-IV performance is enhanced to good extent. The rates for identification and verification is different although both gives very good results like former gives around 97% and latter around 98%.

Raul Sanchez et.al[9] have implemented a biometric system using hand geometry where hand image is first converted into a binary image and then edge detection is performed based on sobel function to extract contour of hand. The width of palm and distance among the three interfinger points is measured along with the height of fingers. Selected points on the finger and interfinger points are utilized to calculate deviation and angle. For the matching purpose various classifiers are tested and compared among which Gaussian mixture models yields the best results. However other classifiers are Euclidean Distance, Hamming Distance, Radial Basis function and Neural Networks. Experimental results give upto 97% performance which shows that this can be used in medium to high security systems. Among the classifiers for verification GMM's outperforms the other.

Raymond Veldhuis et.al [10] has presented an approach which uses high-level features and low-level features for hand geometry recognition. The high-level features are based on interpreting information like locating a finger and measuring its dimensions such as geometrical features widths, lengths of fingers and angles measured at preselected locations. Landmarks on the contour of the hand are interpreted by features which are low level. However features collected may be very large in number and make process ambiguous so need to reduce the dimension of feature vector to increase the efficiency using combination of two techniques called principal component and discriminant analysis. Data set for testing purposes has been selected from 51 people containing about 850 contour of hand. Experiments evaluation give good results measured in parameter of Equal Error Rate which is about 4%.

Rafael M. Luque- Baena et.al.[7] proposed human identification and verification using hand geometry features. They have used genetic algorithm and mutual information. Their main prospective to design a system for identification rather than classification. The orientation of hand is necessary because images are acquired with no contact to any surface. Segmentation is required for the purpose of extraction so strong hand segmentation is used. Now among extracted features not all are equally important so to select only best suited features an algorithm called genetic algorithm is utilized. GPDS, CASIA and IITD are the databases on which tests are performed and experiments generate almost 100% results in case of GPDS and others produce around 99%.

Nesrine Charfi et.al [4] presented a new concept in biometric systems by combining more than one characteristic in one system. They have fused multiple features to get better performance and features are fused at matching score level. Basically hand geometry biometric system has been employed as hand geometry can easily integrate other features into itself so fingers and palm print are integrated. To present a hand image of a person, shape and texture is extracted from these characteristics. Features are extracted from finger images and contour of hand using a strategy of Scale Invariant Feature Transform(SIFT)

which describes local features. Texture of hand is described from palm of the hand using Gabor filters. Now these two descriptors give advantage of invariance to rotation, translation, scale and lightning conditions. Similarity scores achieved from all the characteristics are finally fused and tested on a database of 230 subjects giving a performance in terms of EER of about 1.95%

Dewi Yanti Liliana and Eries Tri Utaminingshi [2] described how integration may be performed by using multiple features. As palm print has unique characteristics so it has been used in their research which is naturally extracted from hand images. This whole process of extraction of palm print out of hand is done under preprocessing step. Now on processed image one need to extract number of main lines, wrinkle lines, delta lines and minutiae features and this is done using a block based line detection method. Several blocks are constructed by dividing an image and deviation is used as the value of feature and chain code method for hand geometry. Finally Dynamic Time Warping is incorporated to verify the system using both features by measuring distance between them.

Nicolae Dutta[5] reviews general methods and techniques used in biometric systems. Basic steps which are followed and algorithms correspondingly used. She has chosen hand geometry because of its convenience and acceptance by public. She has described the basic first step of any biometric system is pre processing in which image is processed to remove noise, blurring to get smoothed clear image and techniques explained in work are image thresholding, filtering and edge detection. Second basic step is to extract the required information in terms of feature vector using Principal Component analysis, ICA and transformations. Final step is to match the results to decide for this various modules are used few are Euclidean Distance, Hamming Distance, Mahalanobis Distance, Absolute Distance etc.

Salim Chikhi et.al[12] presents a fusion approach. The fused characteristics chosen are ear and iris. Ear is processed by just cropped from face image whereas eye is processed to get a circular area of iris using a transform called Hough transform. Scale Invariant Features which are local Descriptors are extracted from both the characteristics using SIFT and finally for matching purposes these features are stored as template for later use. Similarly images from database are extracted and in matching any kind of matching method generally distance measurement classifier can be applied like here Euclidean Distance has been applied. Results are tremendous as they give approximately 99 % results. Also fused matching Results produce better performance as compared to individual matching.

#### *A .Steps Of Methodology*

From above discussion it becomes clear that almost all hand biometric systems use same kind methodology[3] however differ in terms of technique or algorithm used to be applied according to the requirements of the system. So, common steps are described below which may include a variety of techniques according to what is the requirement and shown in Figure 2.

1.Image Acquisition : This first step includes capturing the image of the hand under different scenarios. Image is acquired using multiple types of cameras.

2. Image Pre-processing : After acquiring image, it may not be in the form like not free of noise or low in resolution due to which cannot be used for further processing. So image is processed so that features can be easily extracted from it. Image smoothing is the general operation performed using morphological algorithms.
3. Feature Extraction : To extract prominent features from hand geometry generally features like finger's and palm's length and width along with hand length are extracted using algorithms based on counting pixels. Basic extraction works on finding useful points to extract to be used further.
4. Matching Module : The features extracted from previous module are now compared with features present in templates which are stored in database. Classifiers are used here which describe the relationship between extracted features and features stored in template by measuring distance or calculating matching score generally used classifiers are SVM, Neural network , GA, fuzzy logic, Euclidean distance etc.
5. Decision Module : The matching score is compared with the threshold. If score comes out to be less than threshold user is accepted otherwise rejected.

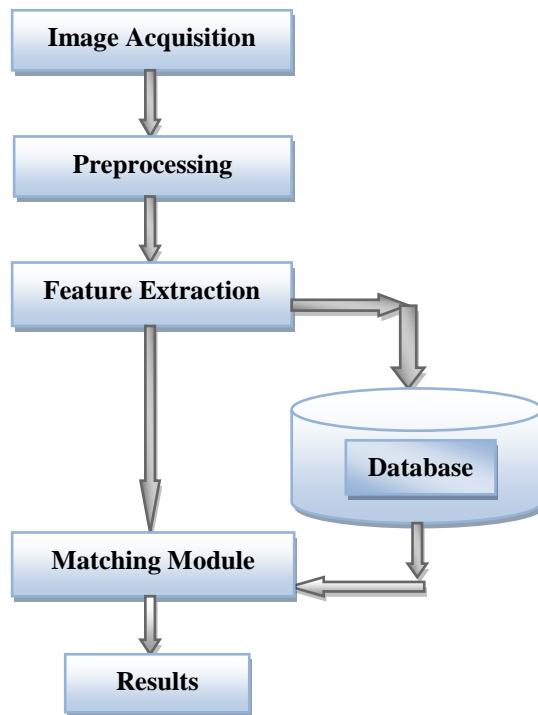


Figure 2. Methodology of Hand Biometric Recognition system

### *B. Performance Parameters*

Performance of any Biometric system verified using different parameters as it perform differently under different techniques so parameters of a biometric system are measured using some standard terms [8] like FAR, FRR, EER and Accuracy

**False Rejection Rate (FRR):** This error happens when a legitimate user is falsely rejected by the system, mathematically measured as the ratio of the number of users rejected by the biometric system which are authorised to the total number of attempts made.

**False Acceptance Rate (FAR) :** This error occur when an imposter is falsely accepted by a biometric system, mathematically measured as the ratio of the number of users accepted by the biometric system which are actually unauthorized to the total number of attempts made for identification.

**Equal Error Rate (EER):** This is the measurement where FRR and FAR are same. The accuracy of a system can be easily detected by the value of EER because the lower the value of EER more accurate will be the system and hence much better. Generally accuracy is calculated by subtracting the sum of FRR and FAR errors from hundred.

## IV. CONCLUSION

Among all biometric systems hand geometry based verification system proved to be accurate and effective so can be used for access control in low to medium security systems. But every system has its own limitations like performance may be effected if hands are prone to variations of illuminations, rotation, and scaling. Also when hands occupy different positions problems may arise like background and translation. So, as suggested by various research studies to overcome such drawbacks one may apply fusion techniques to get better performance as described in studies done in [2],[4], [12] where multiple features like palm, fingerprint, hand, ear and iris are fused. Hand geometry is a debatable biometric feature as some consider it medium in terms of security while other consider it as a strong system which may work in high security systems for verification and identification. To avoid such chances of failure hand geometry may be combined with other features like palm print as studies have shown that thickness of the palm area and width etc. may be considered for identification because of their uniqueness. So, in Future hand geometry may be combined with palm print to build more efficient biometric systems.

## REFERENCES

- [1] Ajay Kumar, David C.M. Wong, Helen C. Shen and Anil K. Jain, "Personal authentication using hand images" , Pattern Recognition Letters, vol. 27 no. 13 pp. 1478-1486, 2006.
- [2] Dewi Yanti Liliana, Eries Tri Utaminingsih, "The combination of palm print and hand geometry for biometrics palm recognition" International Journal of Video & Image Processing and Network Security IJVIPNS-IJENS Vol: 12 No: 01, 2012.
- [3]Mandeep Kaur and Amardeep Singh "A Survey Of Hand Geometry Recognition" International Journal Of Advance Research in Computer Science and Management Studies Vol. 3, Issue 3, March 2015.

- [4]Nesrine Charfi, Hanene Trichili, Adel M. Alimi and Basel Solaiman, "Hand Verification system based on multi-features fusion" IEEE 15<sup>th</sup> International Conference on Intelligent systems Design And Applications(ISDA) Marrakech, pp. 189-194, Dec 2015.
- [5] Nicolae Duta , "A survey of biometric technology based on hand shape" Pattern Recognition Vol .42, no. 11 pp. 2797 – 2806, 2009.
- [6]Nidhi Saxena, Vipul Saxena, Neelesh Dubey and Pragya Mishra, "Hand Geometry: A New Method for Biometric Recognition" International Journal of Soft Computing and Engineering pp: 2231-2307, Volume-2, Issue-6, January 2013.
- [7]Rafael M. Luque-Baena , David Elizondo, Ezequiel López-Rubio, Esteban J. Palomo , Tim Watson, "Assessment of Geometric Features for Individual Identification and Verification in Biometric Hand Systems", Expert Systems with Applications 40 ,pp. 3580–3594,2013.
- [8]Rahul C.Bakshe and Dr. A. M. Patil, " Hand Geometry Techniques : A Review" International Journal of Modern Communication Technologies & Research pp: 2321-0850, Vol. 2, Issue.11, November 2014.
- [9] Raul Sanchez, Carmen sanchez and Ana Gonzalez, "Biometric Identification through Hand Geometry measurements" IEEE Transactions on Pattern Analysis and machine learning vol. 22 no. 10 pp. 1168-1171 October 2000.
- [10]Raymond Veldhuis, Wim Booij, Asker Bazen and Anne Hendrikse, "A Comparison of Hand-Geometry Recognition Methods Based on Low- and High-Level Features", University of Twente, Netherlands, pp 326-330, 2002.
- [11] S. M. Prasad, V. K. Govindan and P. S. Sathidevi, "Bimodal Personal Recognition using Hand Images" Proc. Of the International Conference on Advances in Computing, Communication and Control, pp. 403- 409. ACM, New York, 2009
- [12] Salim Chikhi, Lamis Ghoualmi and Amer Draa, "A SIFT-Based Feature Level Fusion of Iris and Ear Biometrics" F. Schwenker et al. (Eds.): MPRSS 2014, LNAI 8869, pp. 102–112, 2015.

# Review: Simulation Based Crop Modeling

Er. Pankaj Goyal<sup>#1</sup>, Er. Sikander Singh Cheema<sup>\*2</sup>

<sup>#</sup>Computer Engineering Department, Punjabi University  
NH 64 Patiala, India

<sup>1</sup>pankaj90382@gmail.com

<sup>\*</sup>Assistant Professor, Computer Engineering Department  
Punjabi University, Patiala  
<sup>2</sup>Cheemasikander8@gmail.com

**Abstract:** - This paper reviews about two crop models named Dssat, AquaCrop. These models give estimate production of different crops based on various input parameters like soil, water, weather conditions, temperature etc. These models are very useful for estimate yield of different crops. One can judge the future production of different crops of any state or country so based on that information they can build their future planning.

**Key words:** AquaCrop, DSSAT, Crop Models, Decision based Models.

## I. INTRODUCTION

Crop means agglutination of individual plant species developed in a unit region for monetary reason. Model can be simplified evaluation of a system of a process. [1] Simulation is regenerate the logical characteristics of a system without generate the actual system itself. It is done on machines like computers, tablets, servers or any other things which done the computation very fast and handles many types of complex equations. [4]

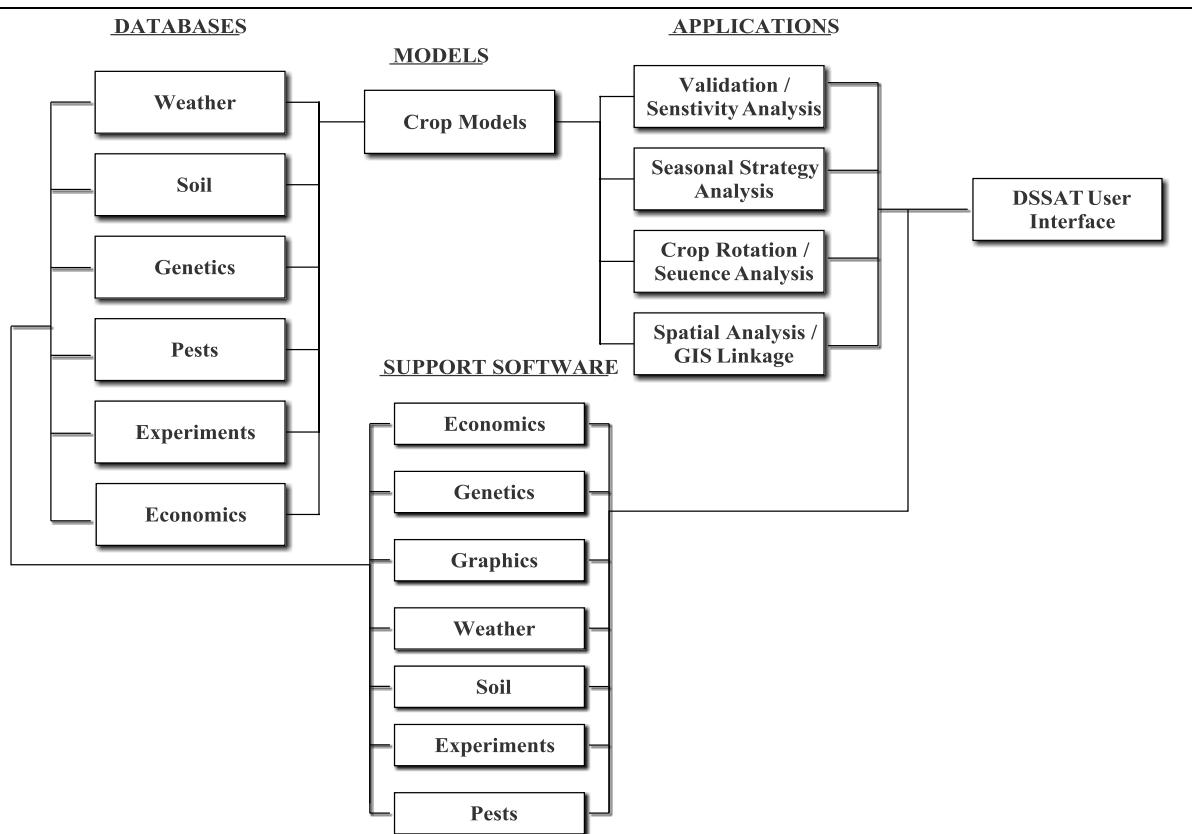
In the simulation of Agriculture field, creating a dynamic environment to predict the growth of a crop over the entire season. How much yield of a crop can be produced from a particular area of land for economic purpose? For planting, growing and harvesting of a crop there is need different types of a parameter for example soil, rain, temperature, humidity, sun radiation, wind direction, wind speed, pests, fertilizers, relative humidity and so on. India having agriculture mainly in the villages so there is a limit to measure these parameters. There is a need to develop that kind of software which needs fewer parameters and produce accurate results.

In a market, there are different types of tools available for simulating the crops to calculate the yield of a crop. These are Infocrop, Dssat, Apsim, Aquacrop, Oryza, Surcos, Stics, Wofost, Epic, Daisy, Corngro, Swheat, Arcwheat, Ceres, Elcros, Bacros, Macros, Papran, Lintul etc. These models developed according to weather of that region, soil conditions of that land. It means using these models in other geographical location may be failed to give the accurate results. Some models which famous worldwide are Dssat and Aquacrop. This paper reviews on working of these models.

## II. DECISION SUPPORT SYSTEM FOR AGRO-TECHNOLOGY (DSSAT)

DSSAT: - Decision support system for agro-technology transfer was developed by the IBSNAT (International Benchmark Sites Network for Agro-technology transfer). [3] The preparatory progress of DSSAT was inspired from a necessity to work together learning about soil, atmosphere, yields, and running for building better choices about exchanging creation innovation from one area to others where soils and atmosphere varied. The frameworks view arranged a structure in which review is coordinated to handle how the framework and its segments work. This seeing then works together into models that allow one to anticipate the conduct of the

framework for given conditions. DSSAT rebound leaders by diminishing the time and HR expected for analyzing elective choices.



**Figure 1: Database, application, module parts and their utilization with harvest models for applications in DSSAT v 4.5. [3]**

The DSSAT is an accumulation of autonomous projects that work together; crop simulation models are at its inside. Databases clarify climate, soil, test conditions and estimations, and genotype data for applying to the models in various cases. The product helps clients to combine these databases and compare reenacted results with perceptions to give them trust in the models or to see whether adjustments are required to enhance accuracy. Also, programs required in DSSAT license clients to recreate choices for harvest management over various years to assess the dangers related to each choice. The Figure 1 shows the main components of DSSAT-CSM.

DSSAT emulate development, augmentation and yield of a crop producing in a uniform region of area underdetermined or simulated management and the transformations in soil water, carbon, and nitrogen that happen under the harvesting framework after some time.

The DSSAT-CSM has a principle driver program, an area unit module, and modules for the essential parts that make up an area unit in a harvesting framework (Figure 2). The Primary modules are for atmosphere, soil, plant, soil-plant-air interface, and management parts. On the whole, these segments depict the time changes in the soil and plants that happen in a solitary area unit because of climate and management. Every module has six operational strides. These are Run initialization, Season initialization, Rate computes, Integration, Daily yield, and Summary yield. The principle program controls when each of these strides is dynamic, and when every module plays out the undertaking that is called for. [7]

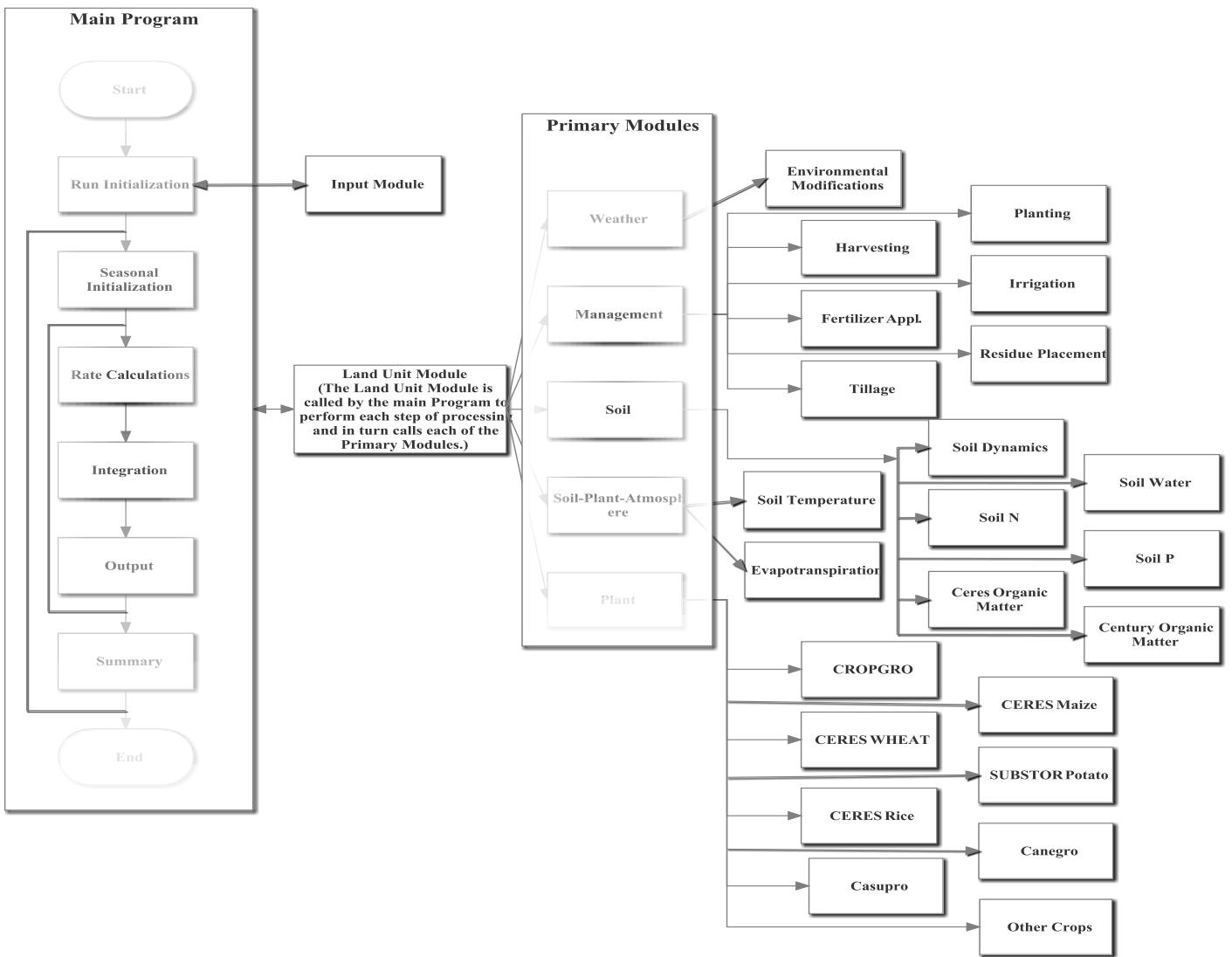


Figure 2: Outline of the parts and measured structure of DSSAT-CSM. [7]

Validation/Sensitivity Analysis accommodates intuitive affectability investigation and correlation of recreated versus watched field information.

Seasonal Strategy Analysis operation recreates crops over various years of climate utilizing the same soil introductory conditions. It allows one to assess the impacts of undetermined future climate conditions on choices made when all soil starting conditions are known.

Crop Rotation/Sequence Analysis works the editing framework modules to reproduce crop revolutions over various years, and soil conditions are instated just at the very begin of the reenactment.

Spatial Analysis/GIS Linkage works the CSM to reenact one or more harvests over space i.e. for exactness farming, land used for growing or other spatial-based applications.

The Land Unit Module confers the interface between the application driver (primary system) and the greater part of the segments that connect in a uniform zone of area. Table 1 delineates the essential and sub-modules used in the CSM and watch their operations. In the first place, sub-modules work precisely like Primary modules. Every sub-module will normally perform six stages, and in this manner it can be changed by various modules that can work with its default information interface

variables and create the characterized module yield interface variables. In this manner, the conviction of "interface" variables is seen to the specific measured methodology utilized as separated of DSSAT-CSM.

**Table 1: Summary representations of modules in the DSSAT-CSM [7]**

Primary Modules	Sub Modules	Behavior
Main Program (DSSAT-CSM)		Manage time loops, finds which modules to be called based on user input information, manage print timing for all modules.
Land Unit		Gives a solitary interface between harvesting system and applications that deal with the utilization of the cropping system. It utilized as accumulation for all segments that communicate on a homogenous zone of area.
Weather		Produces daily weather parameters used by the model. Regulate daily values if needed, and calculates hourly values
Soil	Soil Dynamics	Calculate soil system attribute by layer.
	Soil Water Module	Compute soil water activity including snow ice and defrost, overflow, penetration, immersed stream and water table profundity.
	Soil Nitrogen and Carbon Module	Computes soil nitrogen and carbon forms, including natural and inorganic compost and residue arrangement, deterioration rates, nutrient fluxes between different pools and soil layers.
Soil – Plant – Atmosphere (SPAM)		Redresses dispute for assets in soil-plant-environment framework. Present variant figures subdividing of energy and fix energy equalization forms for soil evaporation, transpiration, and root water extraction.
	Soil Temperature Module	Figures soil temperature by layer.
CROPGRO Crop Template Module		Calculates crop development processes counting phenology, photosynthesis, plant nitrogen and carbon demand, growth partitioning, and pest and disease damage for crops modelled.
Individual Plant Growth Modules	CERES-Maize	Modules that reproduce advancement and yield for individual species. Each is an alternate module that simulates phenology, day by day development and apportioning, plant nitrogen and carbon requirement, senescence of plant material, and so forth.
	CERES-Wheat /Barley	
	CERES-Rice	
	CERES-Sorghum	
	CERES-Millet	
	SUBSTOR-Potato	
	Other (future) plant models	
Management Operations Module	Planting	Discovers planting date shaped on read-in worth or simulated used an input planting window and soil, climate conditions.
	Harvesting	Discovers harvest date, set up on development, read-in value
	Irrigation	Discovers day by day watering system, shaped on read-in qualities or programmed applications set up on soil water depletion.
	Fertilizer	Discovers manure considerations, expand on read-in qualities
	Residue	Use of leftover and other natural organic material

### III. AQUACROP

AquaCrop is a yield water efficiency model created by the Land and Water Division of Food and Agriculture association of the United Nations (FAO). It mimics yield reaction to water of herbaceous harvests and is especially suited to address conditions where water is a key restricting element in productivity creation.

AquaCrop is a Windows-based programming program intended to recreate biomass and yield reactions of field products to different degrees of water accessibility. Its application envelops rainfed and also supplementary shortage and full watering system. It depends on a water-driven development engine that utilizes biomass water productivity (or biomass water use productivity) as key development parameter. [2]

AquaCrop endeavors to adjust exactness, sobriety, and strength. It utilizes a generally little number of unequivocal and for the most part instinctive parameters and information variables requiring straightforward techniques for their determination. The model keeps running on day by day time steps utilizing either date-book time. [8]

An experimental generation capacity in "FAO Irrigation and Drainage Paper n.33" is utilized to judge the yield reaction to water.

$$\left( \frac{Y_x - Y_a}{Y_x} \right) = K_y \left( \frac{ET_x - ET_a}{ET_x} \right)$$

Where  $Y_x$  and  $Y_a$  are the greatest and real yield,  $ET_x$  and  $ET_a$  are the most extreme and real evapotranspiration, and  $K_y$  is the proportionality variable between relative yield deficit and relative lessening in evapotranspiration.

AquaCrop evolves from the past condition by isolating:

- The  $ET_a$  into soil evaporation ( $E_s$ ) and crop transpiration ( $T_a$ ).

$$ET_a = E_s + T_a \quad [8]$$

The detachment of  $ET_a$  into soil evaporation and yield transpiration restricts the unpredictability impact of the non-gainful destructive utilization of water.

- The final yield ( $Y$ ) into biomass ( $B$ ) and harvest index ( $HI$ ).

$$Y = HI(B) \quad [8]$$

The partition of  $Y$  into  $B$  and  $HI$  permits the isolating the primary useful relation and environment condition.

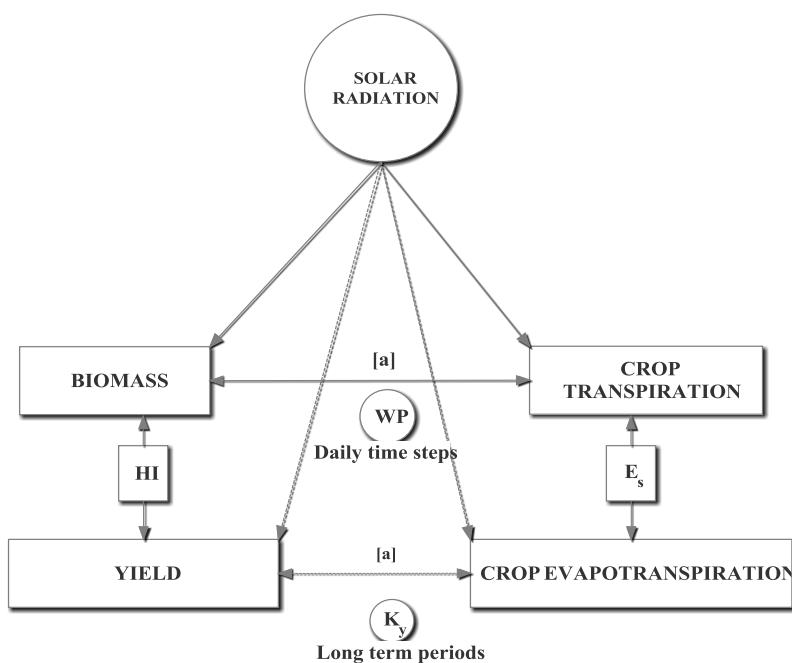


Figure 3: Development of AquaCrop, in light of the presentation of two middle steps: the division of soil evaporation ( $E_s$ ) from harvest transpiration ( $T_a$ ) and the accomplishment of yield ( $Y$ ) from Biomass ( $B$ ) and harvest record ( $HI$ ). [5]

The progressions disclosed prompted the immediate condition at the center of the AquaCrop development engine:

$$B = WP \cdot \sum T_a$$

Where  $T_a$  is the harvest canopy transpiration and WP is the water efficiency parameter. This prompted power of the model because of restoration conduct of WP.

To be operational, this reenactment engine should be embedded in a complete arrangement of extra model segments. AquaCrop join with soil water adjust; the plant, with its advancement, development and yield forms; and the climate, with its polythermal and warm, precipitation, evaporative interest and carbon dioxide fixation. Some running attribute is expressly inspected (e.g., watering system, treatment, and so forth.), as they impact the soil water parity, crop improvement and in that way last yield. The useful connections between the diverse model parts are delineated in the stream graph of Figure 4.

Input prerequisite of AqquaCrop is comprises of climate information, harvest and soil qualities, and management rehearses that characterize the earth in which the yield will create. In the Weather information required is least and most extreme air temperature,  $ET_0$ , precipitation,  $CO_2$  focuses. In the soil parameters are required for information is soil ripeness level, variables influence the soil water equalization, determination of aggregate watering system prerequisite, detail of watering system occasions, watering system plan. The features of AquaCrop that differs from the DSSAT: -

- AquaCrop in view of ground water level.
- The AquaCrop utilizes canopy spread while DSSAT utilizes the leaf zone list.
- The AuaCrop utilizes the water productivity (WP) values standardized for air evaporative interest and  $CO_2$  fixation that give the model a stretched out extrapolation ability to various areas, seasons, and atmosphere, including future atmosphere situations..
- AquaCrop used low number of parameters than a DSSAT.
- The applicability of AquaCrop is to be used in diverse agricultural systems.

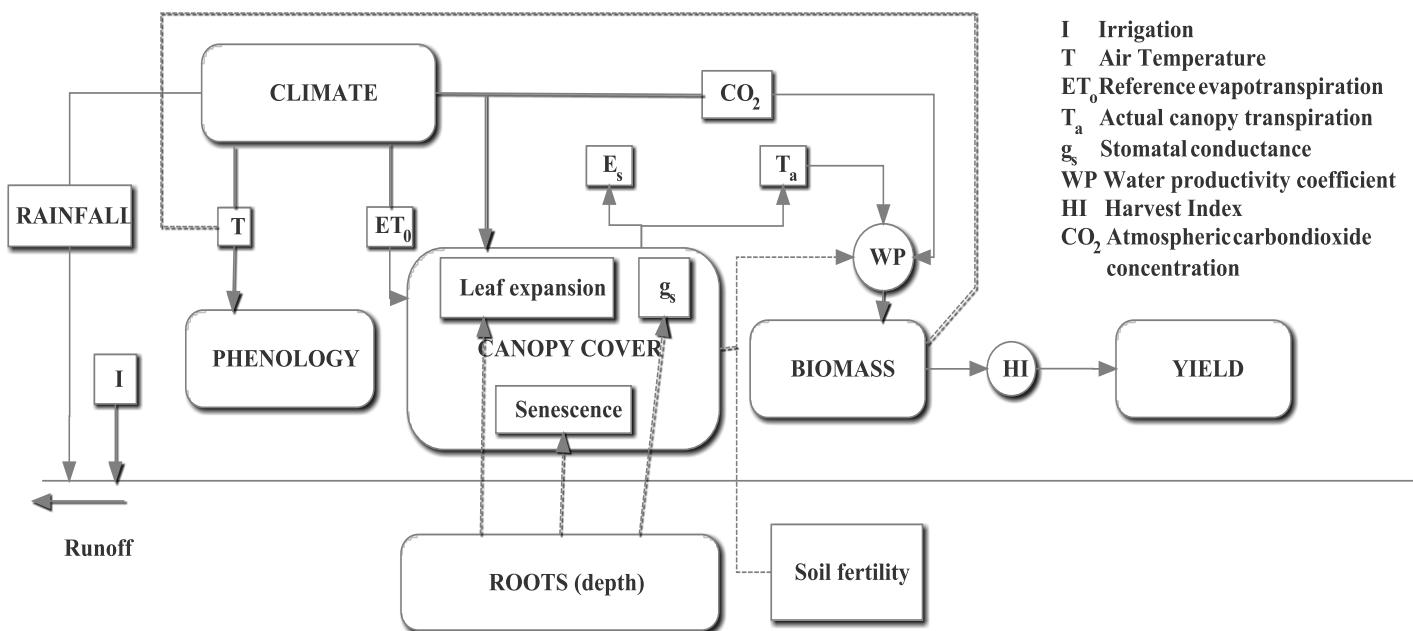


Figure 4: Graph of AquaCrop demonstrating the primary segments of the soil-plant-atmosphere continuum and the parameters driving phenology, canopy spread, transpiration, biomass generation, and absolute yield. [8]

#### IV. FUTURE SCOPE

We are trying to develop a model work for Punjab. In this model, we will try to compare various factors like soil, weather etc. for the optimum output of different crops. Output of different crops can be judge based on the parameters like content of soil, condition of weather. This will help to the farmers of Punjab for the prior production of crops based on conditions of soil and weather.

#### REFERENCES

- [1] Chakrabarti, B. *Crop Simulation Models*. New Delhi: Indian Agricultural Institute.
- [2] FAO crop-model to simulate yield response to water. [http://www.fao.org/nr/water/aquacrop\\_about.html](http://www.fao.org/nr/water/aquacrop_about.html).
- [3] Jones, J. W., et al. "The DSSAT cropping system model." *European Journal of Agronomy* (ELSEVIER Science B.V.) 18 (2003): 235-265.
- [4] Krishna Murthy, V. Radha. *Crop Growth Modeling And Its Application In Agricultural Meteorology*. 235-261: Satellite Sensing And GIS Application in Agricultural Meteorology.
- [5] Steduto, P., et al. "AquaCrop: a new model for crop prediction under water deficit conditions." *Options Mediterraneennes* 80: 285-292.
- [6] What is DSSAT? <http://dssat.net/>.
- [7] G. Hoogenboom, J.W. Jones, C.H. Porter, P.W. Wilkens, K.J. Boote, L.A. Hunt, and G.Y. Tsuji (Editors). 2010. *Decision Support System for Agrotechnology Transfer Version 4.5*. Volume 1: Overview. University of Hawaii, Honolulu, HI.
- [8] Dirk RAES, Pasquale STEDUTO, Theodore C. HSIAO, and Elias FERERES. 2011. *AquaCrop version 4.0 : FAO crop water productivity model to simulate yield response to water chapter 1*
- [9] James Rising, Mark Cane. *Comparison of global agricultural modeling results*.
- [10] Kelly R. Thorp, Kendall C. DeJonge, Amy L. Kaleita, William D. Batchelor, Joel O. Paz. *Methodology for the use of DSSAT models for precision agriculture decision support*.

# Comparison Survey on Different Types of Steganography

Navreet Kaur

M.tech Scholar, Department of Computer Science Engineering, CEC Landran, Mohali, India  
E-mail:tiwananavreet@gmail.com

Isha Vats

Assistant Professor, Department of Computer Science Engineering, CEC Landran, Mohali, India  
E-mail:ishavats90@gmail.com

**Abstract-** Steganography is the form of a stegos Greek words meaning hidden or covered and graphics means writing defining as writing secure covered. Steganography communication support secure data in several mediums such as images, videos and audio. Steganography takes cryptography a step further by hiding a message so that no one suspects there. There are many algorithms proposed in steganography. The main objectives of steganography are robustness against various attacks treatment, the ability of data hidden and undetectable. This article explores the different methods of image and video steganography that are used to hide the message in digital media.

**Keywords:** Steganography, Text Steganography, image steganography, video steganography.

## I. INTRODUCTION

Transmission of data is the biggest challenge for communication from one place to another. There are several ways to provide security and the best is hidden [2] helpful. Steganography is a hidden message in other documents cannot recognize the occurrence of a transmission mode transmission technology in the viewer. It includes a variety of techniques for secure communication. This is different from the encrypted password, because helps to keep secret information is hidden where the information helps to maintain the presence of the message as it is perfect with no secrets. Quality and performance parameters can be used to measure the hidden PSNR video or image and (MSE). Steganographic system is the most important attribute of data, which indicates the existence of which is difficult to determine how hidden messages statistics undetectable (invisible). Another feature is that it refers to how steganography to hide data extraction system resistance and capacity, which can be safely embedded in a robust work [1] the maximum information. Embedded payload and embedding steganographic system efficiency of any data. Amount which may be referred to embed the payload is hidden cover two key parameters in the file. System capacity steganography to hide as much data as it can be used, because it can be called embedding efficiency in the document on the cover [2]. High efficiency embeds any secret system of the main requirements for the least induced distortion. High efficiency means embedding distortion in at least overwrite the file, so it is hard to imagine that there is any secret information in the overlay file. Therefore, different algorithms for higher efficiency are used

embedding. Embedding efficiency and embedding payload usually enjoy inverse relationship. Embed increased efficiency will reduce embedded payload, and vice versa.

## II. STEGANOGRAPHY TYPES

Almost every one digital file may be used for performing steganography, highly redundant format has a variety of known frequency domain, and the image is first transformed helpful. Using bit redundant device / target file, and it provides the necessary accuracy is much better than small objects and display, for example, if we have a clear picture of the sky most of the pixels in the image is that they are blue is considered in this case, the number of redundant pixels below all unnecessary pixels defined in a matrix defined processing, but in fact there is no need, so they called a least significant bit. These redundant bits cannot be changed without producing visible artifacts in the final image. Image and audio files in particular to meet the requirements of this standard [9].The following figure shows the four major categories of file formats in the current steganographic techniques used.

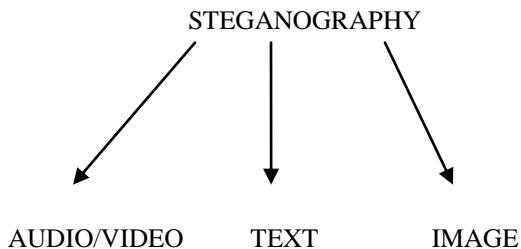


Fig1. Types of steganography

- A. Text information hiding: the number of text messages hidden tabs or spaces or uppercase letters for implementing stealth, like Morse code [2] is used to achieve information hiding
- B. Video Steganography: Video (w.r.t series of pictures to a timeline) is used as the hidden information internet. In the discrete Fourier transform value to be manipulated, it is worth noting that the human eye. May be conducted in secret video formats such as H. 264, MP4, MPEG, AVI, etc.
- C. Audio Steganography: It became very familiar with, due to the popularity of Voice over Internet Protocol (VOIP) growth. Audio steganography can be done in many digital audio formats such as MIDI, AVI and MPEG etc.

## III. STEGANOGRAPHIC PROCESS

The basic process of steganography contains following block Operators, information and secret key. Carrier also referred to override object. Information is embedded in it and used to hide in the presence of an object in covering the news.

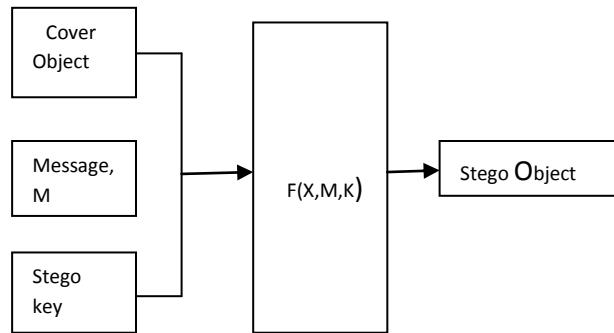


Fig2. Steganographic process

Steganography is a method to hide some particular type of files into a transporting multimedia file. The utilization of the file is dependent Steganography could possibly be more suitable than any former multimedia files, for the reason which is of its size as well as memory necessities

#### IV. COMPARITIVE REVIEW

Author and year	Paper Title	Technique used	Advantage
Abhinav Thakur, 2015	Secure Video Steganography based on Discrete Wavelet Transform and Arnold Transform	Discrete wavelet transformation and inverse DWT	Performance parameters, such as PSNR and MSE can be calculated to determine the quality cover video. Experimental results shows the algorithm has good height for cover concealment perception guarantee
AshishS.Thorat, Prof. Dr. G. U. Kharat ,2015	Steganography based on navigation missile	Least Significant Bit(LSB) technique is used	The idea is that each byte LSB may be little change in the overall use file instead.
Bárbara Emma Sánchez Rinza,2015	Security System for Sending Information Containing Hidden Voice Data by Steganography (SIOVE) Using Matlab	LSB (least significant bit) technique is used	In this a solution proposed to use in MATLAB constructed system, which can be hidden in the image of the speech signal.

M. Kameswara Rao, K. Pradeep Reddy and K. Eepsita Saranya,2015	Security Enhancement in Image Steganography MATLAB Approach	LSB insertion & RSA encryption technique is used	This enhances the security to a higher level because to acquire the steganography image embedded we need to have the key image which will be having only by the receiver
Snehal Satpute ,Sunayana Shahane,Shivani Singh,2015	An Approach towards Video Steganography Using FZDH (Forbidden Zone Data Hiding	Forbidden Zone Data Hiding is the algorithm	The system enables more safety concept using shorthand and password files. This technique, is used for the secure transmission of data.
Ramadhan J. Mstafa and Khaled M. Elleithy,2014	A Highly Secure Video Steganography using Hamming Code (7, 4)	Hamming code (7, 4) before the embedding process	The algorithm is considered to be a highly efficient due to the embedding algorithm so that the data modification covert low quality video has a very good host.

## V. VARIOUS STEGANOGRAPHIC TECHNIQUES

The main function of video Steganography is hide secret message without affecting visual quality, structure and content of the file. Below some of the techniques are given as:-

### 1. HAMMING CODE

Hamming code is the known method of block code, it can prevent the error detection and correction to complete. By increasing the minimum amount of redundancy, which is the so-called n-bit original information [10] Code length codeword Hamming some additional technical data. Add section contains [K] forecast information for the length of the message coding [11] (N-K) parity bits. For example, (7, 4) Hamming code is capable of detecting and correcting data or parity bit single-bit errors. First, add the k message bits (K = 4) (M 1, M 2, M3, M4) Length 3 parity bits (P1, P2, P3) of the length of the codeword is n (n = 7) it is intended for transfer encoding. There are different methods of data (messages and check) for both types, so 2I (P1, P2, M1, P3, M2, M3, M4), where i = 0, 1 parity generally mixed binding site, (NK-1). Hamming code is a linear code, so they have two matrixes: the parity check matrix H and the generator matrix G, encoding and decoding their needs. Encoding side, each message M, comprising up to

four, to produce a product from the matrix, there is applied to the mold 2; As a result, by transmitting codeword  $x \times 7$  Preparation noisy channel.

## 2. LEAST SIGNIFICANT BIT

General information and simple way was at least significant bit (LSB) is inserted into an image embedded in the cover. The least significant (in other words, Section 8) of part or all of the image to change the position of the bit byte secret information. When using the 24-bit images can be used for color components, such as red, green and blue, as they are represented by one byte. In other words, it can be stored in an image of  $800 \times 600$  pixels each pixel, which may be stored or embedded 1,440,000 bits of total of 180,000 bytes of data. For example, for a three pixel image grid 24 may be as follows :( 00101101 0,001,110,011,011,100)

## 3. DISCRETE WAVELET TRANSFORM

DWT wavelet analysis filter bank. Discrete wavelet transform decompose signal into the original signal wavelet coefficients can be a 2-D signal or image is divided into four bands: LL (left), HL (upper right), LH (left), HH.

LL	HL
LH	HH

Fig3. DWT

## 4. SIOVE

SIOVE program is to record an audio file that has been previously recorded, you can play the audio signal. To use this program, or it can be recorded whether the audio signal, and select the button 'select an existing file .WAV files'. After the audio file, 'hidden text' button is selected and activated. User information is written in the text field, then, 'text hidden shown in the program will save the audio file \* .WAV extension. To get the message, select' the recovery text by clicking is selected. Another option that the system will provide is to make sure whether it is included in the hidden text and audio files. To do it, click the user 'and choices. Wav file,' the program will automatically be displayed hidden messages in audio files. in the message 'dialog box, it does not contain hidden text file

## VI. SUMMARY

Steganography is used for the hidden communication. We have reported improved steganographic different systems using different approaches to provide a secure means of communication. Steganography is a method to hide a particular type of file. Using these steganography might possibly be more suitable than any of the above techniques, where each type is different from each other for the reason that is of size and memory requirements. So this paper has presented various techniques of video as DWT steganography, LSB etc.

## REFERENCES

- [1] Abhinav Thakur, "Secure Video Steganography based on Discrete Wavelet Transform and Arnold Transform" International Journal of Computer Applications (0975 – 8887) Volume 123 – No.11, August 2015
- [2] Ashitosh S. Thorat, "Steganography based on navigation missile" International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 4, Issue 6, June 2015.
- [3] Bárbara Emma Sánchez Rinza, "Security System for Sending Information Containing Hidden Voice Data by Steganography (SIOVE) Using Matlab" International Journal of Engineering and Innovative Technology (IJEIT) Volume 5, Issue 1, July 2015
- [4] B. Hao, L.-Y. Zhao, and W.-D. Zhong, "A novel steganography algorithm based on motion vector and matrix encoding," in Communication Software and Networks (ICCSN), 2011 IEEE 3<sup>rd</sup> International Conference on, 2011, pp. 406-409.
- [5] D. Artz, "Digital Steganography: Hiding Data within Data", IEEE Internet Computing, pp. 75-80, May-Jun 2001.
- [6] F.A.P.Petitolas,R.J.Anderson,M.G.Kuhn,"Information Hiding-A Survey", Proceedingof the IEEE, vol. 87, no. 7, June 1999,pp.1062-1078
- [7] J.J. Chae and B.S. Manjunath, "Data hiding in Video" Proceedings of the 6th *IEEE International Conference on Image Processing*, Kobe, Japan (1999).
- [8] Kedar Nath Choudry "A Survey Paper on Video Steganography" Kedar Nath Choudry et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2335-2338.
- [9] M. Kameswara Rao Security Enhancement in Image Steganography a MATLAB Approach Middle-East Journal of Scientific Research 23 (2): 357-361, 2015 ISSN 1990-9233
- [10] Nestler V. J., Conklin W. A., White G. B., Hirsch M. P., Computer Security Lab Manual, McGraw-Hill/Irwin, NY, 2006. ISBN 0-07-225508-0.
- [11] Ramadhan J. Mstafa "A Highly Secure Video Steganography using Hamming Code (7, 4)" in Systems, Applications and Technology Conference (LISAT), 2014 IEEE Long Island , vol., no., pp.1-6, 2-2 May2014doi: 10.1109/LISAT.2014.6845191
- [12] Snehal Satpute An Approach towards Video Steganography Using FZDH(Forbidden Zone Data hiding) IJIACSISSN 2347 – 8616Volume 4, Issue 1January 2015.

# Cloning Attack Detection in Online Social Networks Using Improved Clustering and Similarity Measures

Sanjeev Dhawan<sup>1</sup> Kulvinder Singh<sup>2</sup> and Akshay Sharma<sup>3</sup>

Faculty of Computer Science & Engineering <sup>1, 2</sup>, PG Student of M.Tech. (Software Engineering) <sup>3</sup>

Department of Computer Science & Engineering, University Institute of Engineering and Technology  
(U.I.E.T), Kurukshetra University, Kurukshetra (K.U.K)-136119, Haryana, INDIA.

E-mail(s):<sup>1</sup>rsdhawan@rediffmail.com, <sup>2</sup>kshanda@rediffmail.com, <sup>3</sup>akshaysharmajob@gmail.com

**Abstract:** Today the attractiveness of online social networks is increasing speedily. Users spend their time in popular social networking sites like Facebook, Instagram and Twitter to share the personal data. Clone attack is one of the dangerous attacks in social profile. Attacker stole the personal data and create fake profile in social called cloned profile once cloned profile is created by attacker they sends a friend request using cloned profile. In the proposed system the cloning attack detection using improved clustering (Ikmeans) and similarity measure (Jaro-Winkler) to find out the similarity between cloned profile and real one. After that the clustered similarity data is again tested using similarity function as Euclidean distance to pick up top closest members to be cloned and find out precision, recall and F-measure.

**Keywords:** Cloning attack, Detection, K-Means clustering, Social Networks, Similarity, Online social network (OSN)

## I. INTRODUCTION

Social network popularity is growing massively. Billions of community around the humanity is linked to each other by OSN. OSNs like Facebook, Instagram, and Twitter provide the relation between friends, create new relations with people, and reconstruct relation between old friends and share public interest, hobbies in friendship circles. OSN's stores enormous quantity of sensitive plus private information of users and their conversations. As the content of information is increasing, the public network providers and security companies are bound to provide superior safety features in their public network. Various securities against hacker, spammer, identity cloning, public bots, phishing, and many more threats. But a larger part of online user is not alert with privacy schemes and they frequently disclose an enormous quantity of individual data on their profiles that are able to be seen by anybody in the respite of connections. Clone attack is one of the main attack patterns towards online social networks, where the challenger conceal fake account information by apt to steal and replication real user profile and sends friend requests to the friends of the cloned victim. It is hard for normal users to detect this fake identity because of the same names and alike profile information. In this paper two phases are proposed training phase and testing phase. Training phase collects the user profile information and find the similarity measure using improved kmeans as clustering measures

and after based similarity function. That testing phase is used to predict cluster data and find the performance measure using Euclidean Distance.

## II. RELATED WORK

Social Networks are utilized by many people so duplicacy of profile is occurring day by day. In 2009 Weimin Luo *et al.* [1] proposed the coercion to public networks and examine the targets what the attacker want and the methods how attacker perform the attacks. The authors had proposed some method to divide public networks into two parts name as user networking site and public networking site. Then introduce the related attacks on public networks after that the contented and manner of coercion to public networks. In the last part authors discussed a security framework of public networks and this makes it clear where and of what we should be aware. In 2011, Bhume Bhumiratana [2] proposed a model to find out duplicate (clone) attacks on OSNs (OSN).In this model to develop OSN pathetic conviction model and maintain authenticity of the bogus online identity established by identity clone attack to harvest more private data and argue regarding how the attack can be dissatisfied and avoid by the users and developers of OSN. Their design was used to develop attack methodology to take advantage of cloned bogus profiles and carry authentic conversation between the exploited users. Author presented a system that Works across different public networking sites, and implement a simple experiment to test and fine tune various aspect of the attack. In 2011, Georgios *et al.* [3] proposed a methodology for detecting public network profile clone. The authors had projected the architectural design and execution details of a prototype system that was able to be engaged by users to explore whether they have fall victims to such an attack. In this design three main components was used name as Information Distiller, Profile Hunter, and Profile Verifier. In execution detail authors had consider two approach names as Automated Profile Clone Attacks and Detecting Forged Profiles. Detecting Clone Attack in Public Networks Using Classification and Clustering Techniques was proposed by Kiruthiga *et al.* [4].The first part author had discussed the clone attack detection based on user action time period and users click pattern to find the similarity between the cloned profile and real one in Facebook. The second part author had discussed the users profile information every user's information is stored. Using Naïve Bayes Classifier classify the details for every user information. K-Means clustering is to group the same Network. Clone Spotter is to detect the clone in facebook. In last authors was considering the Cosine similarity and Jaccard similarity to find the similarity for improving the performance. Moreover Morteza and Fatemeh [5] proposed the detection approach was organized by six methods that was Discovering community the public network graph, Extraction user's attribute, Search in community, Selecting profile, Computing strength of relationship, Decision making all these methods to identify and detect profile clone. In addition profile clone detection in social networks was proposed by akshay *et al.* [6] to discuss the profile cloning and detection process of cloned profile. Author had described the profile cloning process. The main targets of profile cloning are user who set their profile to be public. Social profile allows attacker to obtain profile information easily and therefore can duplicate or copy their profile information to create a fake identity. Author also discussed the type of clone name as existing profile clone and cross-site profile clone. In detection of Cloned Profile author was described three method name as Extracting Data from the user's Profile, Searching for the user's profile on other Public networking sites, calculating the similarity index to identify the cloned profile.

From this section, it is clear that although there are various techniques available to improve the recommendations in order to detect profile cloning attacks. However, in most of the recommendation systems (like Naïve bayes and similarity measure), it is very difficult to find the accurate results. In order to avoid this problem, an attempt has been made to define new propose a enhance algorithm using Ikmeans, similarity measures and Euclidean distance.

### III. PROBLEM DESIGN AND IMPLEMENTATION

In the proposed system data set is collected from Facebook. Fig. 1 shows the scenarios for cloning attack in social networks.

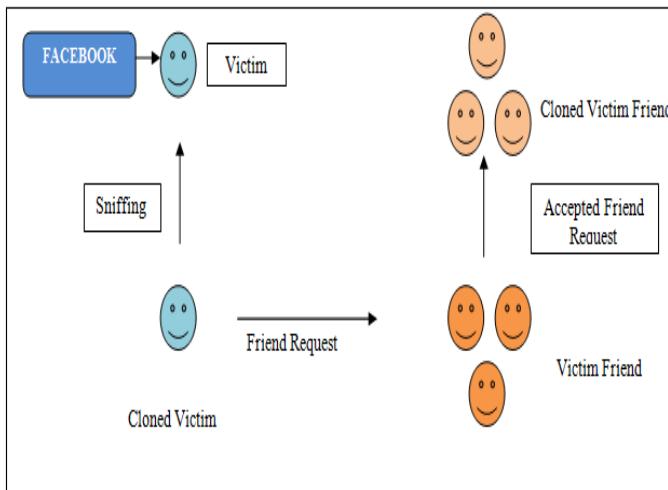


Figure1.Cloning Attack

Most of the users are spending the time in Facebook and other social sites. In that A is one of the users in the Facebook. Attacker clone the A profile and collects the information and sends a friend request to the A's friends. If they accepted the friend request, their information will also be visible to attacker, they will also get cloned. In order to do training phase and testing phase are used as proposed system to achieve more accurate results.

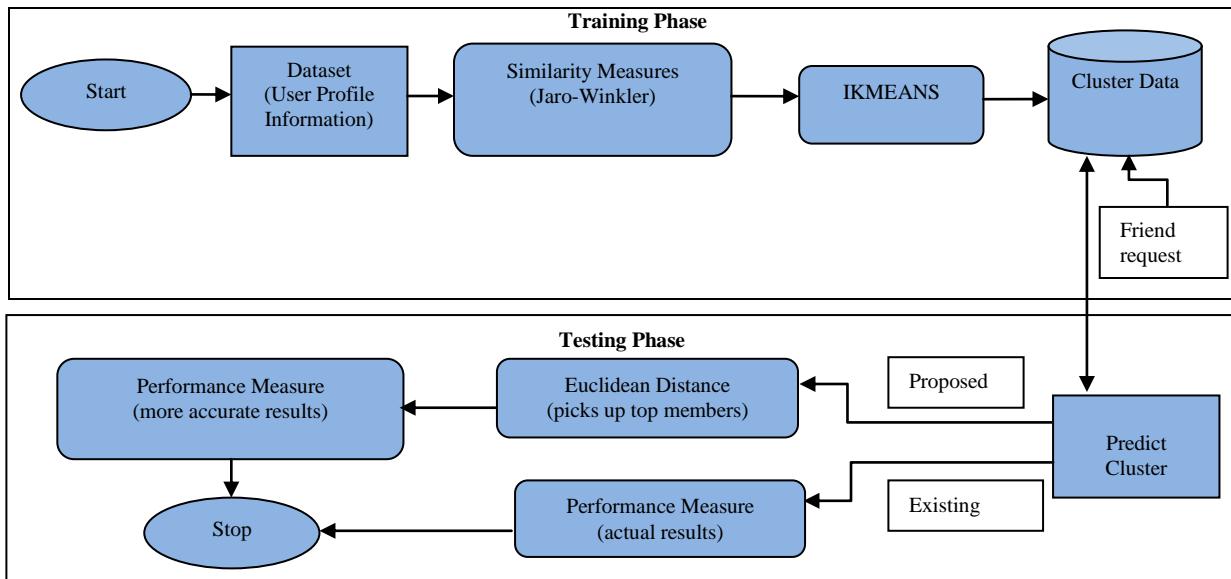


Figure 2. Training phase and Testing phase architectures

In the Proposed Training phase and testing phase architectures in figure 2 are better accurate results as compared to existing system. In training phase used two method names as Jaro-Winkler and Ikmeans for the cluster data.

#### A. Jaro-Winkler

Jaro Winkler is string matching technique which can be used to give a comparing score in-between 0 and 1.Jaro–Winkler distance is a measure of similarity between two strings. Jaro Winkler is adopted for matching the userid, display name, description and finding the similarity score. After finding the similarity score, the similarity matrix was created. The Jaro–Winkler distance metric is designed and best suited for short strings such as person names. The score is normalized such that 0 equates to no similarity and 1 is an exact match.

#### B. Ikmeans

K-Means is a portioning based clustering algorithm having the data objects with similarities in one cluster based on the Euclidian distance. It is the simplest and most widely used algorithm for clustering. It is an unsupervised learning method that follows a simple process for cluster data set. This algorithm always forms  $C$  number of clusters and in every cluster there must be at least one item. In the initial step, K-means chooses fixed centroid points and number of clusters. It then finds the distance of each data point from every centroid. The points are assigned to the centroid at minimum distance from the point. The centroids are recalculated by taking the arithmetic mean of the data points assigned to that cluster. The performance of K- means increases as the number of cluster increases; hence it is advantageous to apply on large datasets.

##### *Algorithm:*

##### *Input:*

C is the random number of clusters

D is the data set having n objects

##### *Output:*

'c' clusters are formed.

##### *Method:*

1. Select c fixed data objects as the cluster centroids.
2. Assign cluster to each data object to its closest centroid.
3. Recalculate the new centroids for each cluster, by calculating the arithmetic mean of data objects in that cluster.
4. If atleast one data object changes its centroid, then move to the step no 2, else move to next step.
5. Output the final clusters.

## IV. EVALUATION

After the training phase is completed then to Testing phase is started. In this phase when a user sends a friend request to cluster data then similarity measures is calculated from predicted cluster data and performance measures is calculated on the basis of precision, recall and Fmeasure. But in proposed system similarity measures is calculated from predicted cluster data set with Euclidean similarity distance .on the basis of result obtained from Euclidean similarity distance the more accurate result are obtained in the form of precision, recall and F-measure.

## V. RESULTS AND ANALYSIS

Both the proposed work and existing work are implemented in Java using Net Beans (Integrated Development Environments) on a computer system with 2.6 GHZ Core i3(third generation) and 4GB RAM . After performing the clustering data process, the output is obtained easily in terms of Precision, Recall and Fmeasure. Table 1 describes the existing and proposed system comparison.

TABLE 1. Comparison between Existing and proposed system

Approach	Precision	Recall	Fmeasure
Existing	0.251751	0.975	0.399476
Proposed	0.604167	0.725	0.659091

### A. Precision

It is defined as the fraction of retrieved results that are relevant to the query. Mathematically it is defined as:

$$\text{Precision} = \frac{\text{Relevant information} \cap \text{Retrieved information}}{\text{Retrieved information}}$$

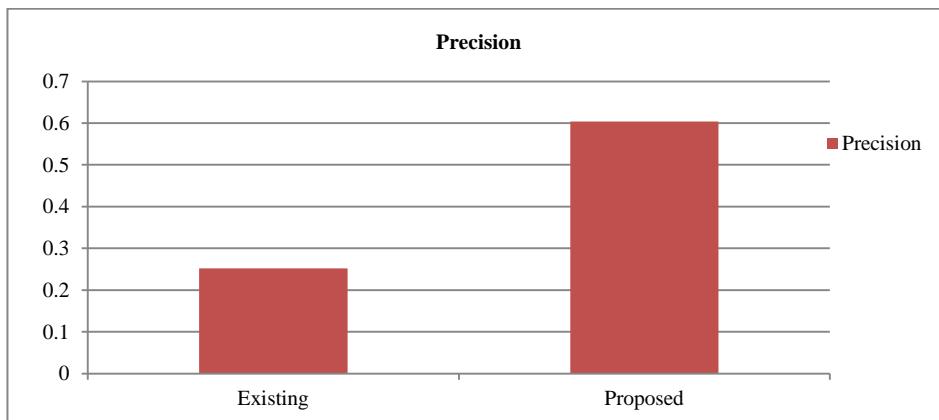


Figure 3. Precision graph of existing and proposed Approach

### B. Recall

It is defined as the fraction of the documents that are relevant to the query that are successfully retrieved. Mathematically, it is defined as:

$$\text{Recall} = \frac{\text{Relevant information} \cap \text{Retrieved information}}{\text{Relevant information}}$$



Figure 4. Recall graph of existing and proposed Approach

### C. F-measure

The measure that combines precision and recall is the mean of them, the traditional F-measure or F score. The F-Measure computes some average of the information retrieval precision and recall metrics. Mathematically, it is defined as:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

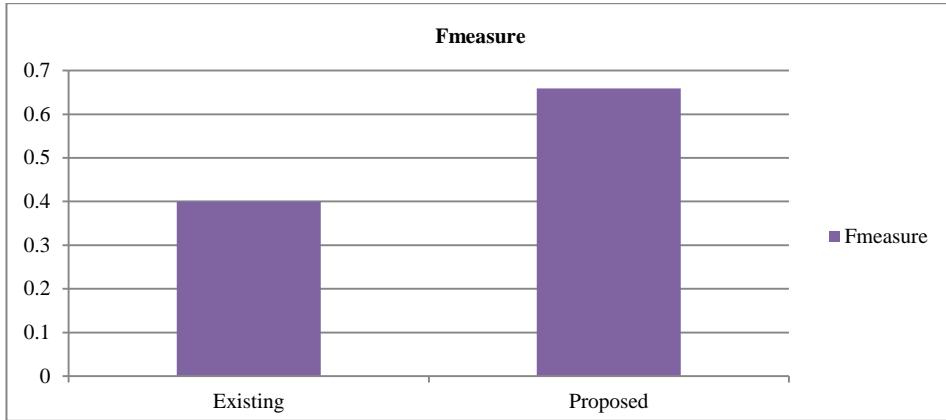


Figure 5. F-measure graph of existing and proposed approach

### D. Performance Comparison

Fig. 6 describes the comparisons of Precision, Recall, F-measure of existing and proposed approaches. It clearly shows that Fmeasure of proposed approach is more accurate result than existing approach.

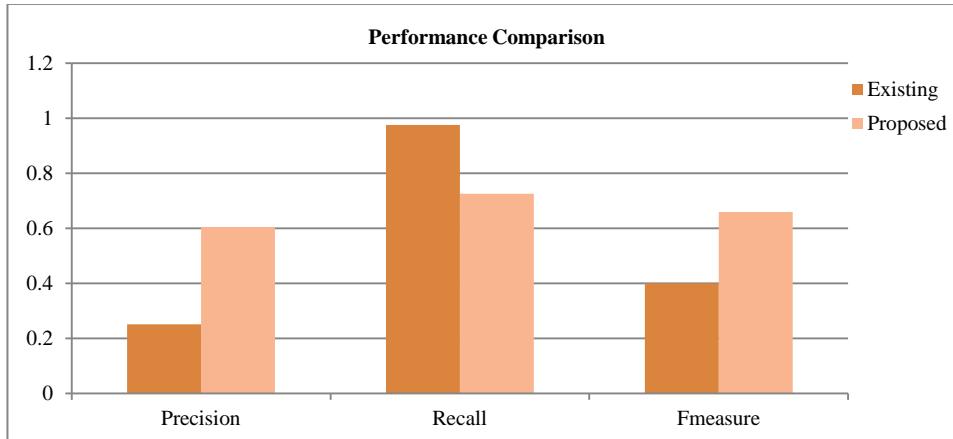


Figure 6. Performance Comparison of existing and proposed approach

## V. CONCLUSION & FUTURE WORK

In this paper, profile cloning attack detection in online social network is proposed using Ikmmeans and similarity measures. In proposed mechanism, Facebook dataset is used for experimental results. To implement proposed mechanism Java Platform with Netbeans is used. Initially similarity measures function is used to find similar data and Ikmmeans clustering is used to group the same school and college of persons and create cluster data. After that the clustered similarity data is again tested using similarity function as Euclidean distance to pick up top closest members to be cloned and find out precision, recall and F-measure. After implementation of proposed work it is to

be concluded that performance of proposed mechanism is more accurate than existing mechanism. In Future, the size of testing dataset can be increased to further analyze the potential of cloned attacks in social networks using other classifiers like k-nearest neighbour (kNN), Random forest (RF) and Naïve Bayes (NB). The authenticity of the data can also be tested using neural network and genetic algorithm for high speed optimization and reliability.

REFERENCES

- [1] Weimin Luo, Jingbo Liu, Jing Liu and Chengyu Fan, “An analysis of security in social networks”, 978-0-7695-3929-4/09, 2009, IEEE, pp.648-651.
- [2] Bhume Bhumiratana, “A Model for Automating Persistent Identity Clone in Online Social Network”, 978-0-7695-4600-1/11, 2011, IEEE, pp.681-686.
- [3] Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis and Evangelos P. Markatos,“Detecting Social Network Profile Cloning”, 978-1-61284-937-9/11,2011,IEEE,pp.295-300.
- [4] Kiruthiga. S, Kola Sujatha. P and Kannan. A, “Detecting Clone Attack in Social Networks Using Classification and Clustering Techniques”, 978-1-4799-4989-2/14, 2014, IEEE, pp.1-6.
- [5] MortezaYousefi Kharaji1 and FatemehSalehi Rizi, “An IAC approach for detecting profile cloning in online social networks”, Vol. 6, No.1, January 2014, IJNSA, pp.1-6.
- [6] Akshay Sharma, Sanjeev Dhawan, and Kulvinder Singh, “A Review paper on profile clone detection in social networks” July 2016, IJESRT, pp.1006-1010.

# Design and Simulation of Folded Arm Miniaturized Microstrip Low Pass Filter

Inder Pal Singh, Praveen bhatt<sup>1</sup>, Nitin Kumar<sup>2</sup>

Shinas College of Technology, Shinas, P.O. Box 77, PC 324, Oman, [ipsinghphys@gmail.com](mailto:ipsinghphys@gmail.com)

<sup>1</sup>Samalkha Group of Institution (SGI), Panipat, Haryana, India, praveen34592@gmail.com

SRM University, Modinagar, Uttar Pradesh, India, [nitin.k@srmuniv.ac.in](mailto:nitin.k@srmuniv.ac.in)

**Abstract**—In this paper we presented two simple designs Chebyshev 3-pole microstrip stepped-impedance low-pass filter in L-band (1 GHz) and folded arm microstrip lowpass filter which is widely being implemented in GPS systems, mobile phones and defence telemetry. Various shapes are designed, simulated and frequency is tuned by altering the stub size and its position. According to the shape of the devices these two filters can be implemented. These two filters are designed for 1 GHz cut-off frequency at -3dB with a passband ripples less than 0.1dB and it shows sharp stopband 1 GHz – 4.9 GHz. Effective permittivity of the substrate is 10.8 and height 1.27 mm. The stepped impedance LPF is a traditional filter and folded arm filter is also a stepped impedance LPF but its inductive arm is bent at 90°. Folded arm filter is the miniaturized form of the traditional LPF. The miniaturized LPF gives the same performance as the traditional stepped impedance LPF. The filter is miniaturized by folding its inductive arm at 90° and its dimensions are optimized. Area of folded arm lowpass filter is reduced by 27.9% with respect to the traditional stepped impedance lowpass filter. These LPF filters are designed and simulated on Ansoft-HFSS platform. Its gives the satisfactory results between theoretical and simulated ones.

**Keywords**—stepped impedance LPF, Chebyshev, L-band, folded inductive arm, HFSS

## I. INTRODUCTION

Microstrip filter is a three layered structure, ground, substrate and patch. Substrate is sandwiched between ground and the patch. Microstrip structure is made on PCB of desired specification. The advantage of microstrip filter is its compact size, not expensive, easy fabrication, light weight, easy troubleshooting. The disadvantage of microstrip filter is its poor power handling capability [1]. Power handling capability of the structure is reduced when high impedance lines are implemented in the design because thin metal lines can't tolerate the high power. At higher microwave frequencies and higher dielectric constant it is difficult to control the heat generation in the conducting structure as well in dielectric. In case of dielectric breakdown, peak power is rapidly reduced. An alternative approach to reduce power loss is, use of low constant or such type of dielectric which has very high thermal resistance e.g. alumina. Power handling capability can be easily controlled at low microwave operating frequency.

Proposed LPF is operating at 1 GHz frequency eventually loss is very less. By using multilayer structure power handling capability can be optimized [2]. If surface of microstrip filter is exposed to air then there is a scope of the interference between surface waves and other unwanted radiation. This can be minimized by shielding the structure by metallic waveguide to prevent entering of any type of electromagnetic radiation inside the structure and it also suppresses a surface modes. Full wave electromagnetic analysis is performed for the modeling of shielding effect. The enclosure cancels the electric field inside the box. The proposed LPF is enclosed by metal housing. It is difficult to fabricate very high impedance line since the line is very thin. Stepped impedance LPF has its impedance high and low in steps from one end to another end [3].

Its design and mathematical formulation is very simple and accurate but its dimensions are larger so it is not always suitable for the compact devices [4]. In some applications, area of the component on the PCB is very precious and can't compromise with the size of the device. In order to design proposed miniaturized LPF, the inductive arms are folded at 90 degree and applied some compensation techniques to retain the total inductance and capacitance of the traditional stepped impedance LPF to get the same frequency response. Since in this paper we compared the properties of two designs at the same center frequency and we achieved to minimize the area of the proposed filter.

## II. DESIGN EQUATION OF STEPPED IMPEDANCE LOWPASS FILTER

### A. Important Parameters In Design Consideration

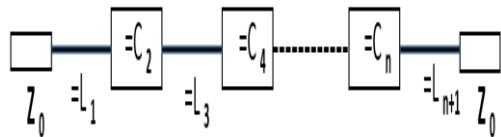


Fig.1 (a)

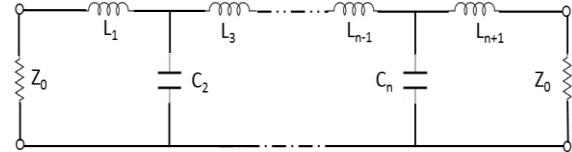


Fig.1 (b)

Consider the Fig.1 in which a traditional layout of the stepped impedance LPF is shown [5]

Fig. 1 (a) General layout of the stepped-impedance LP microstrip filter.

(b) L-C ladder type LPF.

In Fig.1  $Z_0$  is the source and load impedance.  $L_1, L_3 \dots L_{n+1}$  are inductive elements.  $C_2, C_4 \dots C_n$  are capacitive elements.

### B. Design Equation of Microstrip Stepped Impedance Lowpass Filter

Design equations to realize the 3-pole Chebyshev L-C ladder type stepped impedance lowpass filter are given below. Design equations of the filter depends upon the lowpass prototype, type of response, number of elements that filter comprises. Since L-C ladder type lowpass filter consists alternate inductive and capacitive elements.

$$L_1 = L_3 = \frac{z_0}{g_0} \cdot \frac{\Omega_c}{2\pi f_c} \cdot g_1 \quad (1)$$

$$C_2 = \frac{g_0}{z_0} \cdot \frac{\Omega_c}{2\pi f_c} \cdot g_2 \quad (2)$$

Equation (1) & (2) represent the value of inductor (nH) and value of capacitor (pF).

Where,  $\Omega_c$  is normalized cut-off frequency.

$f_c$  is the cut-off frequency based upon the guided wavelength.

$g_0, g_1, g_2$  are the lowpass prototype element.

$$l_L = \frac{\lambda_{gL}}{2\pi} \sin^{-1} \frac{\omega_c L}{Z_{0L}} \quad (3)$$

$$l_C = \frac{\lambda_{gc}}{2\pi} \tan^{-1} \omega_c C Z_{0C} \quad (4)$$

Where

$l_L$  = physical length of inductor ,  $l_C$  = physical length of capacitor,  $\lambda_g$  = guided wavelength,  $\omega_c$  = cut-off frequency

Physical length of inductor and physical length of capacitor should be optimized by satisfying these conditions

$$\omega_c L = Z_{0L} \tan \frac{2\pi l_L}{\lambda_{gL}} + Z_{0C} \tan \frac{\pi l_C}{\lambda_{gc}} \quad (5)$$

$$\omega_c C = \frac{1}{Z_{0C}} \tan \frac{2\pi l_C}{\lambda_{gc}} + 2 \times \frac{1}{Z_{0L}} \tan \frac{\pi l_L}{\lambda_{gL}} \quad (6)$$

### III. DESIGN PARAMETER

#### A. Design specification of Stepped impedance open-stub Lowpass filter

Cut-off frequency, $f_c = 1$ GHz	Normalized frequency = 1 GHz
Relative Dielectric constant, $\epsilon_r = 10.2$	Substrate height, $h = 1.27$ mm
No. of poles =3	Function type = Chebyshev
Characteristic impedance = 50 $\Omega$	Passband ripple = 0.1 dB (return loss $\leq -20$ dB)

#### B. Design specification of Folded arm Stepped impedance open-stub Lowpass filter

Cut-off frequency, $f_c = 1$ GHz	Normalized frequency = 1 GHz
Relative Dielectric constant, $\epsilon_r = 10.2$	Substrate height, $h = 1.27$ mm
No. of poles =3	Function type = Chebyshev
Characteristic impedance = 50 $\Omega$	Passband ripple = 0.1 dB (return loss $\leq -20$ dB)

**TABLE.1 FILTER PARAMETER OF STEPPED IMPEDANCE OPEN-STUB LOWPASS FILTER**

Filter Parameter	Value	Filter Parameter	Value	Filter Parameter	Value
$g_0$	1	$l_0$	8.0 mm	$w_0$	1.1 mm
$g_1$	1.0316	$l_1$	9.84 mm	$w_L$	0.1 mm
$g_2$	1.1474	$l_2$	7.14 mm	$w_c$	8.7 mm
$g_3$	1.0316	$l_3$	9.84 mm	$C_2$	4.6 pF
$g_4$	1	$L_1=L_3$	8.2 nH		

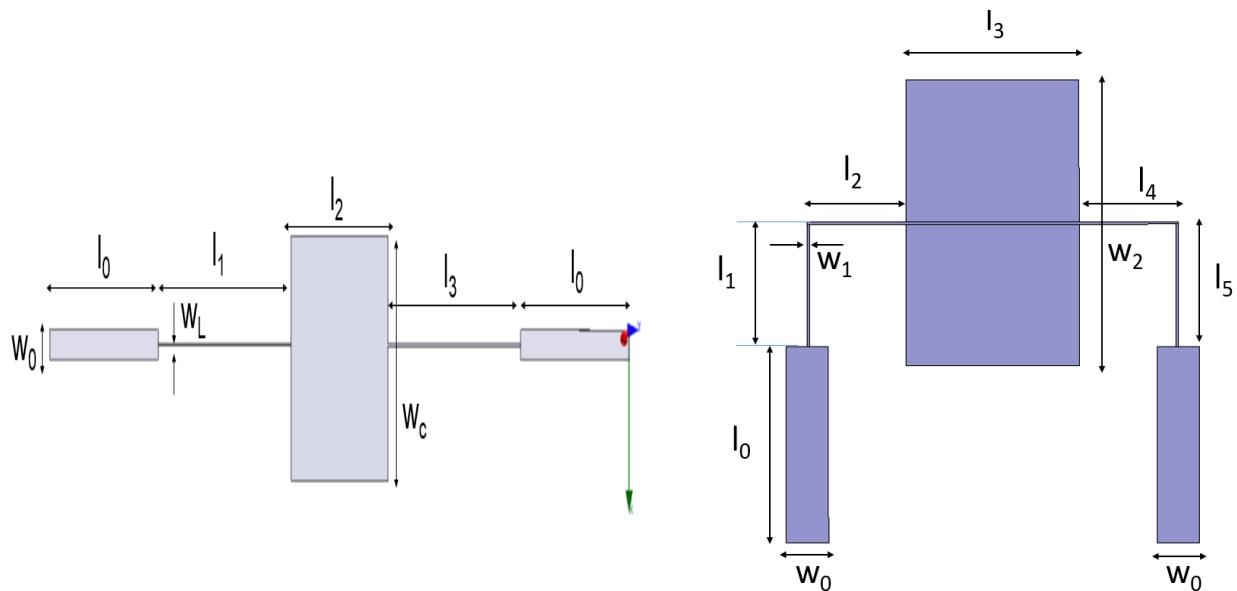


Fig. 2 Layout of stepped impedance open-stub lowpass filter.

Fig. 3 Layout of Folded arm stepped impedance open-stub lowpass filter

**TABLE. 2 FILTER PARAMETER OF FOLDED ARM STEPPED IMPEDANCE OPEN-STUB LOWPASS FILTER**

Filter Parameter	Value	Filter Parameter	Value	Filter Parameter	Value
$g_0$	1	$l_1 = l_5$	5.09 mm	$l_0$	8.0 mm
$g_1$	1.0316	$l_2 = l_4$	4.29 mm	$w_0$	1.1 mm
$g_2$	1.1474	$l_3$	7.5 mm	$w_1$	0.1 mm
$g_3$	1.0316	$C_2$	4.6 pF	$w_2$	11.66 mm
$L_1=L_3$	8.2 nH				

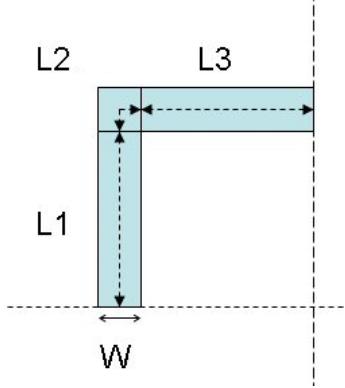


Fig. 4 Perpendicular bent in microstrip line.

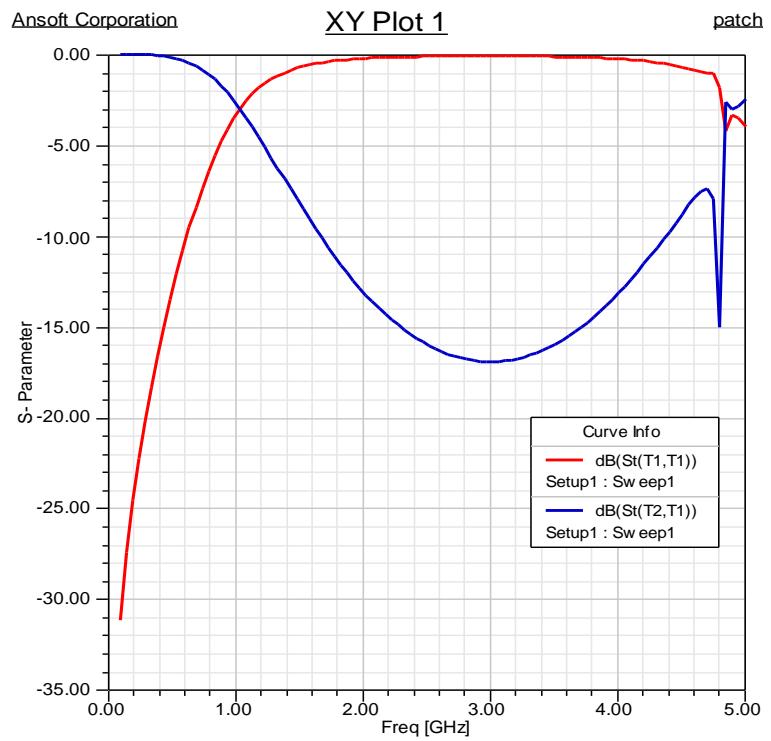


Fig. 5 Plot of S11 and S21 parameter vs frequency.

In the Fig. 3 detailed structure of perpendicularly folded inductive arm stepped impedance lowpass filter is shown. Bent arm acts as a combination of inductor and capacitor like in T-network. The additional inductive and capacitive effects are compensated in the proposed filter by changing the dimension of inductive arm and dimension of central capacitive patch. It could be compensated by designing a mitered bend but in the proposed filter it is not feasible or practical from the fabrication point of view, since we used a very thin inductive line i.e. 0.1 mm. So dimensions of the central capacitive patch are optimized in order to maintain the center frequency 1 GHz.

Microstrip is bent as shown in the fig. 4. When two microstrip are placed over one another at 90 degree. It is very essential to take into account the overlapped part L<sub>2</sub> of the microstrip.

Closed-form expressions for evaluation of capacitance [6]:

$$\frac{C}{W} \frac{pF}{m} = \frac{\frac{14\epsilon_r+12.5}{h} \frac{W}{h} - 1.83\epsilon_r + 2.25}{\frac{W}{h}} + \frac{0.02\epsilon_r}{\frac{W}{h}} \quad (7)$$

The bend in the microstrip increases the return loss since the parasitic discontinuity capacitances increases. Equation no. (7) is used to compensate this loss. This loss is more significant when the operating frequency is  $\geq 3\text{GHz}$ . The proposed filter works on the center frequency 1GHz subsequently loss is less and moreover inductive arm width is only 0.1 mm so the conductive area is less.

#### IV. SIMULATION AND ANALYSIS

##### A. *Stepped impedance open-stub Lowpass filter*

In Fig. 5 simulated S-parameter response of the 3-order stepped impedance open-stub Lowpass filter is shown. S11 and S21 are plotted w.r.t to frequency. The filter works well for the cut-off frequency 1 GHz at -3dB. This filter shows the return loss -31 dB. Its shows a wide and smooth stop band from 1 GHz to 4.9 GHz. Insertion loss is -17 dB. It shows a ripple of – 0.1 dB.

##### B. *Folded Arm Stepped Impedance Open-Stub Lowpass Filter*

In Fig.6 S-parameter (S11 & S21vs. frequency) response of proposed filter, folded arm stepped impedance LPF is shown. In designing of LPF with a wide stop-band techniques are proposed in [7], but the obtained response and transition band equal to 0.4 GHz is not sharp enough. In Fig. 6the cut-off frequency is 1 GHz at -3dB. Smooth and wide stop-band from 1 GHz to 4.95 GHz. The return loss of the filter is -21 dB. The insertion loss of the filter is -24 dB.

##### C. *Combined Analysis of Stepped Impedance and Folded Arm Lowpass Filter*

In the Fig. 7 comparative results of S-parameter are made. In this graph, response of stepped impedance lowpass filter and proposed folded arm stepped impedance lowpass filter are compared. Both the filters show a cut-off frequency 1 GHz at -3 dB. Return loss of folded arm LPF is 10 dB higher than the stepped impedance LPF. Insertion loss of stepped impedance LPF is more than folded arm LPF. Both filter show a good stop-band response. Both filter show same stop-band range 1 GHz to 4.9 GHz.

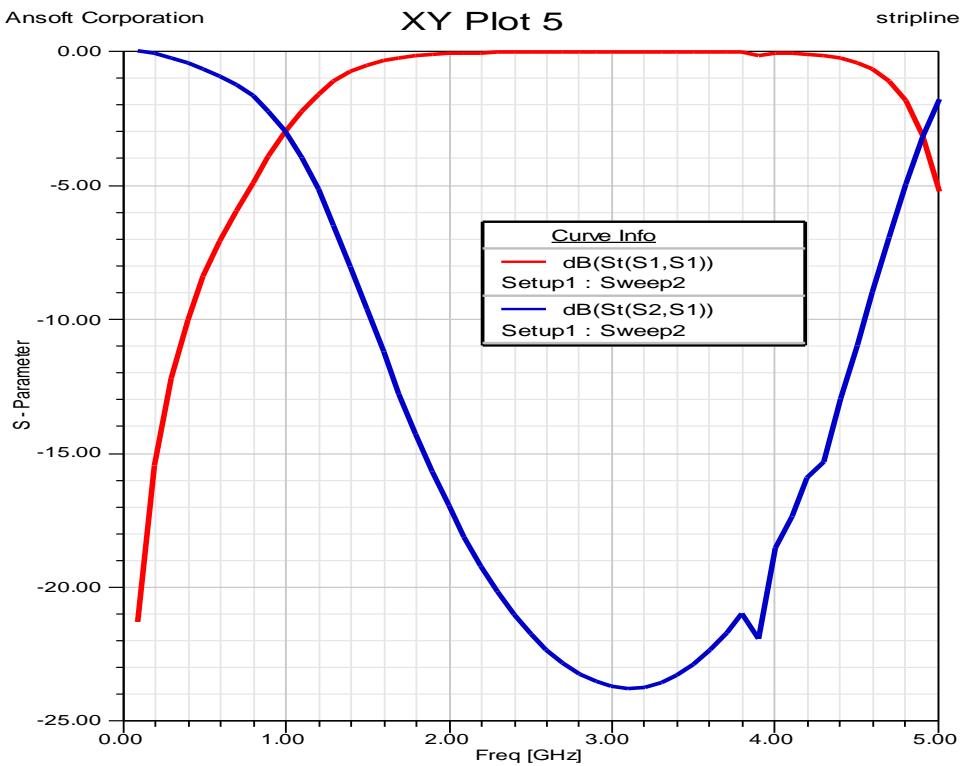


Fig. 6 Plot of S11 and S21 parameter vs frequency of Folded arm LPF

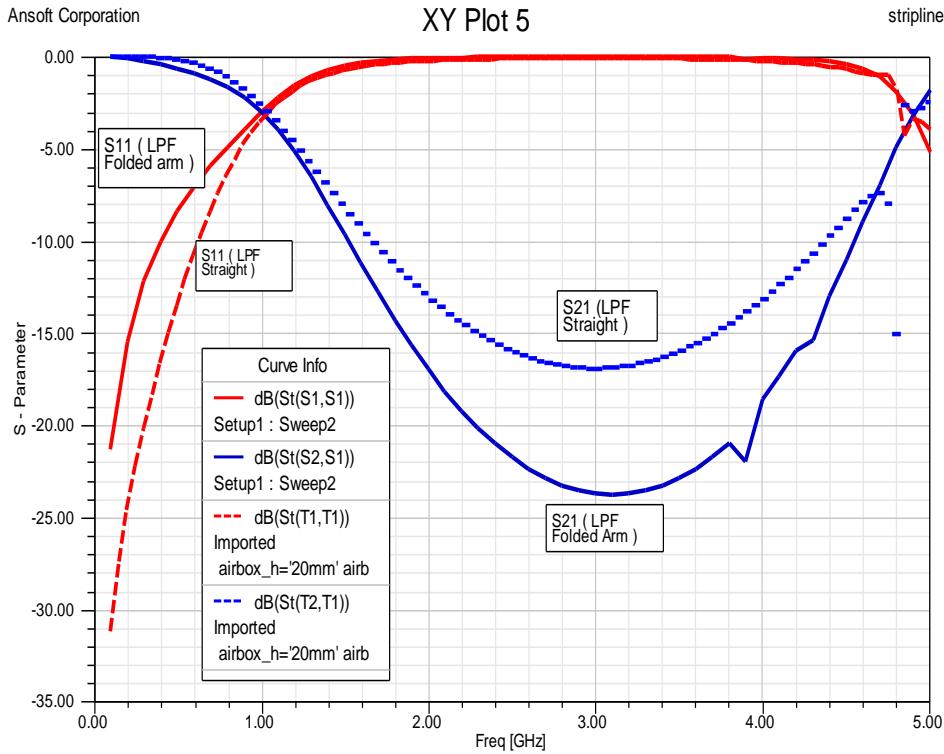


Fig. 7 Comparison of simulated S-parameter response of stepped impedance open-stub lowpass filter and folded arm lowpass filter.

## V. CONCLUSION

Traditional 3-pole stepped impedance microstrip lowpass filter and the proposed 3-pole folded arm microstrip stepped impedance lowpass filter show very good response for the center frequency 1 GHz at -3 dB with minor variations. In this paper we achieved to design for the compactness of the filter without changing the specification of the stepped impedance lowpass filter. Proposed filter shows better and wide stop-band characteristic and sharp rejection at stop-band. Total dimensions occupied on the PCB board by stepped impedance are 42.82 mm x 9.84 mm. Area of stepped impedance lowpass filter is 421.3 mm<sup>2</sup>. Whereas total dimensions occupied on the PCB board by folded arm stepped impedance lowpass filter are 16.08 mm x 18.88 mm. Area of folded arm stepped impedance lowpass filter is 303.5 mm<sup>2</sup>. We achieved to design miniaturized folded arm stepped impedance lowpass filter by 27.9% in the area occupied on the board. Insertion loss of folded arm filter is 7 dB less than the stepped impedance lowpass filter.

## REFERENCES

- [1] K. Rajasekaran, J. Jayalakshmi, T. Jayasanka, "International Journal of Scientific and Research Publications, Volume 3, Issue 8, August 2013.
- [2] J. Inder Bhal, "Average power handling capability of multilayer microstrip lines", International Journal of RF and Computer-Aided Engineering, Volume 11, Issue 6, 2001, pp. 385–395.
- [3] G.L., Young L. Mathaei & E.M.T. Jones, "Microwave Filter impedance matching networks and coupling structures," Artech House, Dedham, Mass, 1980.
- [4] S. Das, S.K. Chowdhury, "Design Simulation and Fabrication of Stepped Impedance Microstrip line Low Pass Filter for S-band Application using IE3D and MATLAB" IJECT Vol. 3, Issue 1, Jan. - March 2012
- [5] J. S. Hong, M. J. Lancaster, "Microstrip Filters for RF/Microwave Applications", John Wiley & Sons, Inc., 2001, pp. 109-110
- [6] K. C. Gupta, R. Garg, I. Bahl, and P. Bhartia, *Microstrip Lines and Slotlines*, Second Edition, Artech House, Boston, 1998.
- [7] J. L. Li, S.W. Qu, and Q. Xue, "Compact Microstriplowpass Filter with Sharp Roll-Off and Wide Stop-Band," Electron. Lett., vol. 45, no. 2, Jan. 2009, pp. 110-11

# A Review on Image Segmentation Techniques

Simerjeet Kaur<sup>1</sup>, Nirvair Neeru<sup>2</sup>

<sup>1</sup>Student, M.Tech (CE) Department UCOE,

Punjabi University Patiala

Punjab, India.

sidhusimerjit90@gmail.com

<sup>2</sup>Assistant Professor, M.Tech (CE) Department UCOE,

Punjabi University Patiala,

Punjab, India.

nirvair\_neeru@yahoo.com

**Abstract**— Bone Age Assessment is an important medical practice in pediatrics radiology to assess any skeletal defect or immaturity based on chronological age. The bone age assessment is performed mostly using a bone atlas which contains the images of carpal bone (left wrist). A trained pediatrician matches the radiograph of carpal bone in atlas with that of the radiograph of carpal bone of subject. Based on his/her analysis then, is he/she predicts the bone age. Automation of such trivial task seems necessary and thus many different techniques for automatic bone age assessment exist today. In this paper, a review of these techniques is presented.

**Keywords**— Bone age, Carpal bone, Region of interest (ROIs)

## I. INTRODUCTION

Bone age assessment could be a procedure oftentimes performed in pediatric radiology. Supported an imaging examination of skeletal development of a left-hand articulation radio carpea, the age is assessed then compared with the age. A discrepancy between these 2 values indicates abnormalities in skeletal development. This examination is universally used because of its simplicity, borderline radiation exposure, and therefore the handiness of multiple ossification centres for analysis of maturity. It is an important procedure in the diagnosis and management of endocrine disorders serving as one index of therapeutic effect. Being a useful procedure in the diagnostic evaluation of metabolic and growth abnormalities [1], it indicates acceleration or decrease of maturation in a variety of syndromes, malformations, and bone dysplasia [2]. Bone age assessment procedure is used to for patients with gonadal dysgenesis [3] or when metacarpal sign occurs [4]. It is also applied in planning for an orthopaedic procedure for correction of angular deformities or abnormalities of length involving the vertebral column or long bones.

In clinical practice, the most commonly used bone age assessment method is atlas matching by a left hand and wrist radiograph against the Greulich & Pyle (G&P) atlas [5] which contains a reference set of normal standard images. However, besides the fact that the data in G&P atlas was collected in 1950s, this method strongly depends on experience of the observer, leading to considerable inter- and intra-observer discrepancy.

## II. RELATED WORK

Most of the work done by others is based on image processing using the bone age atlas [5]. Some of work is done on Active Shape Models [6] [7] [8] and Linear Regression Concepts [9] [10].

**Taani et al.** [6] presented a replacement approach to classifying bones of the hand-wrist pictures into medicine stages of maturity victimization purpose distribution models (PDM). The strategy consists of 2 sections: the coaching section and therefore the classification phase. Throughout coaching, samples of bones from every category area unit collected so the allowable form deformations for every category area unit learnt. A model representing every category is generated. These models area unit afterwards accustomed classify new samples of the bones. Throughout classification all models area unit compared to the input image and therefore the object is assigned to the category whose model is that the nearest match. Experimental results obtained victimization a hundred and twenty pictures of the third distal and middle phalanxes showed the quality of the strategy for classifying these bones into their correct stages of maturity.

**Cootes et al.** [7] described a way for building compact models of the form and look of versatile objects (such as organs) seen in 2nd pictures. The models were derived from the statistics of labeled pictures containing samples of the objects. Every model consists of a versatile form example, describing however the relevant locations of small print on the objects will vary, and an applied math model of the expected gray levels during a region around every model purpose. The paper delineates however the models will be employed in native image search, and provides samples of their application to medical pictures.

**Behiels et al.** [8] evaluated varied image options and completely different search methods for fitting Active form Models (ASM) to bone object boundaries in digitized radiographs. The initial ASM methodology iteratively refines the cause and form parameters of the purpose distribution model driving the ASM by a statistical method match of the form to update the target points at the calculable object boundary position, as determined by an appropriate object boundary criterion. The paper planned Associate in Nursing improved search procedure that's additional strong against outlier configurations within the boundary target points by requiring resultant form changes to be swish, that is obligatory by a smoothness constraint on the displacement of close target points at every iteration and enforced by a lowest value path approach. The paper compared the initial ASM search methodology and our improved search algorithmic program with a 3rd methodology that doesn't have confidence iteratively refined target purpose positions, however instead optimizes a world theorem objective perform derived from applied math a priori contour form and image models.

**Lindner et al.** [9] presented a completely automatic technique to accurately section the proximal femoris in anteroposterior girdle radiographs. Variety of candidate positions is made by a worldwide search with a detector. Every then refined employing an applied mathematics form model beside native detectors for every model purpose. Each international and native models use Random Forest regression to vote for the optimum positions, resulting in strong and correct results. The performance of the system is evaluated employing a set of 839 pictures of mixed quality. The paper showed that the native search considerably outperforms a spread of other matching techniques, which the absolutely machine-controlled system is in a position to attain a mean point-to-curve error of lower than zero.9 millimeter for nine% of all 839 pictures. To the most effective of our information, this is often the foremost correct automatic technique for segmenting the proximal femoris in radiographs nonetheless reportable.

**Adeshina et al.** [10] evaluated the utility of a model of the structure of the carpal space bones within the hand for predicting skeletal maturity in infants (0 - 7 yrs) employing a texture based mostly applied mathematics model of look. Skeletal maturity assessment is vital for diagnosis and observation growth disorders. Applied mathematics models of bone form and look are shown to be helpful for estimating skeletal maturity. The paper situated the carpal bones with associate degree automatic system that uses an unnatural native Model with Random Forest Regression pick. Appearance models were engineered from the metameristic pictures, and texture parameters extracted were wont to estimate skeletal age. By associate degreeanalysis the performance on a data-set of 284 digitized radiographs of traditional infants the paper showed that an mechanically metameristic texture based mostly look model of the carpal region produces terribly satisfactory skeletal age estimation results of a mean absolute error of (0.43; 0.53) years, for male and feminine severally.

### III. RELATED TECHNIQUES

#### A. Active Shape Models

Biomedical pictures typically contain complicated objects, which can vary in look considerably from one image to a different. Trying to live or find the presence of specific structures in such pictures will be an intimidating task. The inherent variability might thwart naive schemes. However, by exploitation models which may traumatize the variability it's doable to with success analyze complicated pictures. [11]

Active form models were developed and introduced by *Cootes et al.* in 1992. Later in 1995, they revealed another article concerning the idea of the idea conferred in [11].

The Point Distribution Model (PDM) is employed to represent a category of shapes supported the distribution of points. These points will represent boundaries and important landmarks (internal locations) of an object. The mean positions of those points and also the main modes of variation describing the movement of points toward mean are given as:

$$x = x + Pb \quad (1)$$

Where,  $x$  represents the  $n$  points of the shape.  $\bar{x}$  is the mean position of points and  $P$  is the matrix of the first  $t$  modes of the variation,  $p_i$  corresponding to the most significant eigenvector in a Principal Component Decomposition of the position variables.  $b$  is a vector of weights for each mode.

By victimization PDM, the corresponding limits area unit obligatory is introduced victimization repetitive approach. These tend to develop a lively form model (ASM).

ASM area unit smart once managing straight forward shapes though if the form is extremely complicated wherever the model points don't seem to be essentially lying on sturdy edges, additional subtle models area unit required. As an alternative, ASM may be combined with Genetic Algorithms (GA) to optimize the load matrix and additional search ways may result into a way promising model.

#### *B. Random Forest Regression*

One of the earliest examples of regression based mostly matching techniques was the work of Covell [13] who used regression toward the mean to predict the positions of points on the face. Random Forest is an ensemble of  $B$  Trees  $\{T_1 | X, \dots, T_B | X\}$ , where  $X = \{x_1, \dots, x_p\}$  is a  $P$ -dimensional vector of descriptor. The ensemble produces  $B$  outputs  $\{Y_1 = T_1 | X, \dots, Y_B = T_B | X\}$  where,  $Y_b, b = 1, \dots, B$  is the prediction for a class by the  $b$ th tree. The average of ensemble's prediction is used for final prediction. [14]

Once a tree has been assembled, the reaction for any perception can be anticipated by taking after the way from the root hub down to the proper terminal hub of the tree, in view of the watched values for the part variables, and the anticipated reaction esteem just is the normal reaction in that terminal hub. [15] An arbitrary timberland is irregular in two ways: every tree depends on an irregular subset of the perceptions, and every split inside every tree is made in light of an arbitrary subset of many competitor variables. Trees are very insecure, so that this arbitrariness makes contrasts in individual trees' forecasts. The general expectation of the woodland is the normal of forecasts from the individual trees - on the grounds that individual trees produce multidimensional stride works, their normal is again a multidimensional stride work that can in any case anticipate smooth capacities since it totals an extensive number of various trees. [15]

#### *C. Tanner and Whitehouse Technique*

The TW2 technique doesn't utilize a scale taking into consideration the age, rather it depends on an appointment of bone's customary development for each age public. [16] In points of interest, within the TW2 strategy twenty square measures of interest (ROIs) set within the principle bones are thought-about for the age assessment. Each ROI is divided in 3 sections: epiphysis, appendage and shaft particularly in children, it's conceivable to tell apart these various natural processes focuses within the phalanx closeness. The advantage of this strategy is that it utilizes the finger and also the carpal bones. Burden is its spread methodology of varied ROIs.

#### IV. CONCLUSION

The analysis of different bone age assessment methods included the techniques like G & P which provided an atlas of chronological bone age. The atlas is good for manual assessment but requires manual assessment and trained person. The automatic method of age assessment is based on the different techniques where two important steps are required. Bone Segmentation from the radiograph and assessment of chronological age based on features. These steps are performed using many different methods. We surveyed a method called Active Shape Model (ASM) for segmentation and Random Forest Regression for Prediction of age. The advantages of each model are discussed. After analyzing we have seen the Active Shape Models have better performance in automatic segmentation scheme. Also, Random Forest Method is faster during training and has performed better.

#### REFERENCES

- [1] A.K. Poznanski, S.M. Garn, J.M. Nagy, and Jr. J.C. Gall, "Metacarpophalangeal Pattern Profiles in the Evaluation of Skeletal Malformations", *Radiology*, vol. 104, no. 1, July 1972 , pp.1-11.
- [2] D.R. Kirks and N.T. Griscom, "Practical pediatric imaging: diagnostic radiology of infants and children", (3<sup>rd</sup> Ed.) [Online]. Available: <http://www.worldcat.org/title/practical-pediatric-imaging-diagnostic-radiology-of-infants-and-children> , 1998.
- [3] J. Kosowicz, "The Roentgen Appearance of the Hand and Wrist in Gonadal Dysgenesis", *The American journal of Roentgenology, radium therapy, and nuclear medicine*, vol. 93, Feb. 1965, pp. 354-361.
- [4] R.M. Archibald, N. Finby and F. De Vito, "Endocrine significance of short metacarpals", *The Journal of Clinical Endocrinology & Metabolism*, vol. 19, no. 10, Oct 19, 1959, pp.1312-1322.
- [5] W.W. Greulich, and S.I. Pyle, "Radiographic atlas of skeletal development of the hand and wrist", *The American Journal of the Medical Sciences*, vol. 238, no. 3, Jun 1, 1959, pp. 393.
- [6] A.T. Al-Taani, I.W. Ricketts, and A.Y. Cairns, "Classification of hand bones for bone age assessment", *In Electronics, Circuits, and Systems, ICECS'96. Proceedings of the Third IEEE International Conference*, vol. 2, Oct. 1996, pp. 1088-1091.
- [7] T.F. Cootes, A. Hill, C.J. Taylor, and J. Haslam, "Use of active shape models for locating structures in medical images", *Image and vision computing*, vol. 12, no. 6, July 1994, pp. 355-365.
- [8] G. Behiels, F. Maes, D. Vandermeulen, and P. Suetens, "Evaluation of image features and search strategies for segmentation of bone structures in radiographs using active shape models", *Medical Image Analysis*, vol. 6, no. 1, March 1, 2002, pp. 47-62.
- [9] C. Lindner, S. Thiagarajah, J.M. Wilkinson, G.A. Wallis, T.F. Cootes, and acogens Consortium, "Fully automatic segmentation of the proximal femur using random forest regression voting", *IEEE transactions on medical imaging*, vol. 32, no. 8, 2013, pp. 1462-1472.
- [10] S.A. Adeshina, C. Lindner, and T.F. Cootes, "Automatic segmentation of carpal area bones with random forest regression voting for estimating skeletal maturity in infants", *In Electronics, Computer and Computation (ICECCO) 11th International Conference IEEE*, Sept. 2014, pp. 1-4.
- [11] T.F. Cootes, C.J. Taylor, D.H. Cooper and J. Graham, "Active shape models-their training and application", *Computer vision and image understanding*, vol. 61, no.1, Jan. 1, 1995, pp. 38-59.
- [12] T.F. Cootes, and C.J. Taylor, "Active shape models—"smart snakes"', *BMVC92 Springer London*, 1992, pp. 266-275.
- [13] M. Covell, "Eigen-points: control-point location using principal component analyses", *Automatic Face and Gesture Recognition, Proceedings of the Second IEEE International Conference*, Oct. 1996, pp. 122-127.
- [14] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling", *Journal of chemical information and computer sciences*, vol. 43, no.6, Nov. 2003, pp. 1947-1958.
- [15] U. Grömping, "Variable importance assessment in regression: linear regression versus random forest", *The American Statistician*, Jan. 1, 2012, pp. 308-319.
- [16] J.M. Tanner, R.H. Whitehouse, W.A. Marshall, and B.S. Carter, "Prediction of adult height from height, bone age, and occurrence of menarche, at ages 4 to 16 with allowance for midparent height", *Archives of disease in childhood*, vol. 50, no. 1, Feb 1975, pp. 14-26.

# Ant MAXMIN approach of Load balancing through VMs in cloud computing

Kanwarpreet Kaur<sup>1</sup>, Amardeep Kaur<sup>2</sup>, Parmeet Kaur<sup>3</sup>

<sup>1</sup>Punjabi University Regional Centre for IT and Management Mohali, India

<sup>2</sup>Punjabi University Regional Centre for IT and Management Mohali, India

<sup>3</sup>Punjab Technical University Jalandhar, India

<sup>1</sup>kanwarpreet74@gmail.com

<sup>2</sup>amardeep\_tiet@yahoo.com

<sup>3</sup>meetnandha36@gmail.com

**Abstract—**Load balancing by virtual machine migration is one of the most significant issues in cloud computing research. A basic approach is to use intelligent algorithms such as Ant Colony Optimization (ACO). However, the main issues with traditional ACO is that it depends on the initial conditions, which affects the convergence speed and final optimal solution. To solve this problem, we propose Max-min Algorithm. Another problem, ACO could arrive at local optimal point, and the convergence speed is typically low. Along this line, we introduce the idea of max-min optimal feature selection to avoid local optimal and accelerate the convergence. Lastly, our experiments show that our improved ACO (Max-min ACO Algorithm) achieves good performance in load balancing. The experimental results show that SLA violations and number of migrations are reduced. The new scheduling strategy was simulated using the Cloud-Sim toolkit package.

**Keywords**—cloud computing, load balancing, ant colony optimization (ACO) and max-min ant system, virtual machine migration

## I. INTRODUCTION

Cloud computing is an emerging computing paradigm in which hosting and delivering of services over the internet is operated by third party [1]. To completely understand the capability of cloud computing, cloud suppliers need to guarantee that they can be adaptable in their virtual machine (VM) conveyance to meet different buyer prerequisites, while keeping the customers detached from the basic datacenter. The shared resources are provided according to the customer request at specific time. The data is stored in the data centers at various hosts, as request of clients can be random to host machines. Request can vary in quantity and thus the load on each host is vary which causes unevenly loaded of tasks over the host machines. Which can reduce the working efficiency of clouds. To solve this problem the virtualization technique is used in which the number of virtual machines are created on the host machines [2]. Live virtual machine migration (VMM) is a technology for achieving load balancing in a cloud environment by migrated an active VM from one host to another as shown in Fig 1. This method has been implemented to reduce the downtime for migrating overloaded and under loaded VMs [3]. To improve the efficiency of Virtual Machine Migration various load balancing algorithms are used to ensure the equal utilization of all the available resources so that no any machine is overloaded or under loaded. The basic working of load balancing algorithms is to provide the path through which various virtual machines are migrated. The basic algorithm used in our approach is Ant Colony Optimization (ACO). ACO is an efficient scheduling algorithm that is used to solve the hard NP-Problem. ACO is a swarm optimization technique based on the foraging of food by ants [4]. The ants while traversing leave the pheromone solution and other ants follow that path. But in this approach the stagnation is occurred. To improve this problem, a hybrid approach of load balancing is used in which the Max-Min Ant System is combined with ACO.

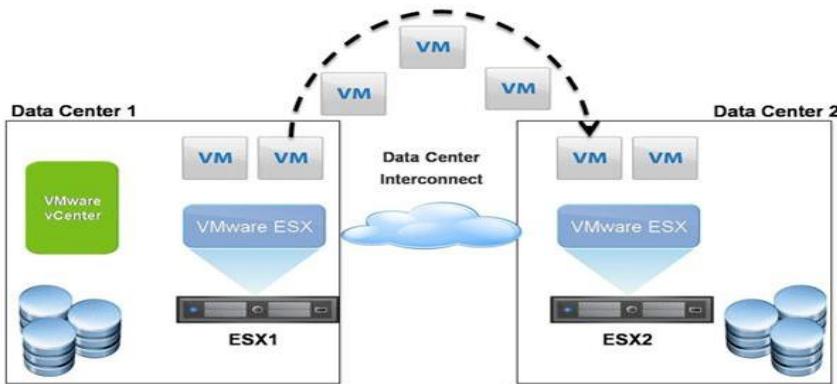


Figure 1. Virtual machine migration in cloud computing

In the MMAS the pheromone range is limited between minimum and maximum value. The low and upper bound avoids ants to converge too soon in some ranges [5].

In our proposed paper, we can find out that hybrid ACO-VMM with Max-Min ant system is capable of migrating the virtual machines more efficiently at lower cost. In order to avoid the unnecessary migration ACO-VMM take the current and previous conditions into account but the stagnation is occurred while migrating the resources from one machine to another. So, to remove this problem, Max-Min ant system is hybrid with this approach. The performance of proposed work is evaluating using Cloud Sim toolkit package [6]. Experiment results show that, the SLA violations and number of migrations are less as compare to previous techniques.

## II. RELATED WORK

There is a growing need for load balancing in cloud data centers. Most of the work in this field leads to data centers or considering the data center hardware aspects. This literature survey will focus on the usage of virtual migration for load balancing.

A PM-LB virtual machine deployment algorithm is proposed in which performance vector method is proposed to deploy virtual machine algorithm [7]. To reduce the time and cost consuming Task based System Load Balancing method using Particle Swarm Optimization is proposed in which system load balancing is achieved by transferring the extra tasks from overloaded VM instead of migrating the whole VM [8]. To speed up the load balancing process in data centers network-topology aware parallel migration is proposed. The VMM based multi resource is migrated to minimum weighted matching problem over the weighed bipartite graph [9]. To minimize the VM downtime and improve the user experience a workload-adaptive live migration algorithm is proposed. This algorithm records the history information of mapping to place the future decisions [10]. For efficient load balancing based on foraging behavior of honey bees, bee colony based algorithm is proposed [11]. In the proposed agent-based problem solving technique, the workloads across commodity, heterogeneous servers is balanced by using the VM live migration [12]. In the ACO-VMM algorithm, the local migration agents independently monitor the resources and their utilization [4]. It takes previous and current condition both into account to avoid the unnecessary migrations [4]. The Ant Colony Optimization algorithm is based on the behavior of ants for finding near optimal solutions like ants foraging

for food. The ants traverse randomly through different paths. When the ant find the food, it comes back with pheromone (markers) to original place. The other ants follow that path with certain probability. As other ants find that path, it gets stronger until the other paths are not common. The age of pheromone is depends upon the two different traversing strategies:

- Positive traversing: the ants traverse through PMs with high pheromone.
- Negative strategy: the ants traverse through PMs with less pheromone.

In our approach, the positive traversing is used based on the cost (pheromone). But in ACO\_VMM the large number of pheromone solution is generated so it is difficult to select better results and another problem is stagnation. Stagnation is the situation when some pheromones are being idle because the ants follow same path for migration. This results in congestion while the other paths are being idle. To remove this problem Max-Min Ant System is used. Max-Min Ant System was proposed by Sutze[13] with some improvements like:

- The initial value of pheromone is set as pheromone solution of ACO-VMM, the algorithm perform better results in finding new explanations.
- When the ants complete one round, only the ants with pheromone solution update the pheromone.
- The value of pheromone is limited between the avg\_bid (max,min) by  $\lambda$  random variable.

When the ants stagnated, MMAS uses the local search and new search technology to improve the algorithm efficiency [14].

### III. PROPOSED WORK

Cloud computing is very often and common in this modern era of technology. In data centers of clouds, the demand of jobs exceed, thus the physical machines doesn't work accurate. The various load balancing algorithms are used to resolve this problem. To reduce the burden on physical machines virtualization technique is used. The virtual machines intakes the memory, space or CPU utilization of physical machines. But the actual problem occurs when the demand of a job for resources exceeds even through the virtual machines. To solve this problem virtual machine migration is used with various load balancing algorithms to migrate the virtual machines from one physical machine to another even in same or different data centers. In this paper, a hybrid algorithm is created which utilizes the features of Ant Colony Optimization and Min-Max Ant System.

The proposed methodology is split into two parts. The first part creates a random environment in which the virtual machines are configured on the physical machines and in second part a hybrid approach of ACO-VMM and Min-Max Ant system is applied for better way to migrate the various VMs on various PMs.

*Algorithm 1: Generation of a random virtual machine migration environment*

```
createEnvironment(h,v)           //h=number of host; // v=number of virtual machines;  
for i=1:h  
    h_ram(i)=xrandom();        //ram for host machines;  
    h_space(i)=xrandom();       //space for host machines;
```

```

h_ramcost(i)=xrandom()           //ram cost for host machines;
h_hddcost(i)=xrandom();          //harddisk cost for host; }

for j=1:vm

v_ram(j)=xrandom();             // ram for virtual machines;
v_space(j) = xrandom();          // space for virtual machines;
v_time(j)=xrandom();             //configuration time;
    
```

The above code generates a dynamic runtime allocation and scheduling environment, which is randomly generated.

The host machines and virtual machines are initialized.

Now, the virtual machines are allocated to the host machines according to their properties.

```

If(h_ram(i)>v_ram(j)&&h_space(i)> v_space(j))
allocation_table[0]=v_id;
allocation_table[1]= h_id;
    
```

The above condition checks virtual machine is applicable for host machine or not. It takes the h\_ram, h\_space, vm\_ram, vm\_spce as the input and the output is allocation table with VM id and Host id. A virtual machine can bid more than single host at same time.

*Algorithm 2: Generate pheromone solution using ACO-VMM*

In the second section ,the Ant Colony Optimization technique is applied and find the pheromone solution on the basis of cost factor. The cost of the Host Machine for bidding is depend upon the hard disk cost, hard disk ram, time and the distance between PM and VM.

```

mytime= time[allocationtable[k][0]];
distance = xrandom;           // distance_cost is from 200 to 500 according to their distance
total_cost=mytime((allramcost[allocationtable[k][0]] + allhddcost[k][0]))+ distancecost;
The pheromone solution is generated according to cost with column: 0=VM,1=Host,3=Cost. The cost factor defines VM configuration on host.
    
```

*Algorithm 3: Generation of Max-Min algorithm*

Implement minmax (pop,d) //d(data)containsthe target elements //pop(population) contains the count of each bid

For 1:eachelement

T=find(d,element); //find element in whole data

For each

$$avg\_bid = \sum_{i=1}^{k.length} d.bid/k.length; \quad (1)$$

$\lambda$  = random(); //  $\lambda$ = random variable. //  $\lambda$  = 0:1

Tp=x\_random(0,1);

If Tp <=0.05

avg\_bid = avg\_bid+  $\lambda$ ;

```

else
avg_bid = avg_bid - λ;
K=find bids minimum than avg_bid

```

The above algorithm generates a  $\lambda$  change in the bid. Only the bids that lie between 0 to avg\_bid are applicable. Then at last, the VMs are migrated according to these bids to host machines.

#### IV. RESULTS AND ANALYSIS

The proposed Ant MAX MIN is compare with five algorithms provided by reference [6] which are ACO\_VMM, Inter Quartile Range Random Selection Algorithm (IQR\_RS), Median Absolute Deviation Random Selection (MAD RS), Local Regression Random Selection algorithm (LR RS) and Static Threshold Random Selection algorithm (THR RS). The performance is evaluated on the basis of SLA Violations and Number of Migrations.

##### A. SLA Violation

To meet the QOS requirements in cloud computing usually SLAs are proposed. The minimum latency or maximum response time is specify by SLAs in enterprise service-level agreement. Beloglazov and Buyya[15] proposed SLA Violation to evaluate the SLA delivered by VM due to over-utilization and performance degradation. They are as:

$$OU\text{ SLA}=\frac{1}{n}\sum_{i=1}^n \frac{T_{t_i}}{T_{x_i}} \quad (2)$$

Where n= no. of PMs,  $T_{t_i}$  is total time for resource utilization,  $T_{x_i}$  is total of i-th PM being in active state.

$$Performance\text{ Degradation} \quad PD\text{ SLA}=\frac{1}{m}\sum_{j=1}^m \frac{P_{d_j}}{P_{r_j}} \quad (3)$$

Where m= No. of VMs,  $P_{d_j}$  estimate performance degradation,  $P_{r_j}$  is total CPU capacity

$$SLAViolation = OU\text{ SLA} * PD\text{ SLA}$$

TABLE I  
 Comparison of Proposed algorithm with different algorithms w.r.t SLA violations

ANT MAXMIN	ACO_VMM	THR_RS_0.8	MAD_RS_2.5	LR_RS_0.8	IQR_RS_1.5
0.301	0.45	3.7	1.57	6.361	1.16

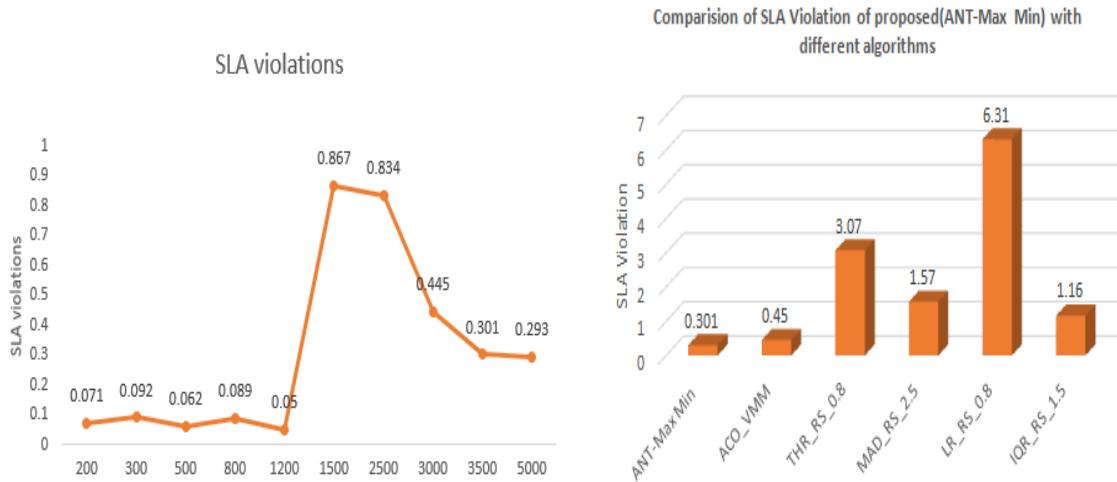


Figure 2.SLA violations of Max-Min ACO.

Comparision of SLA Violation of proposed(ANT-Max Min) with different algorithms

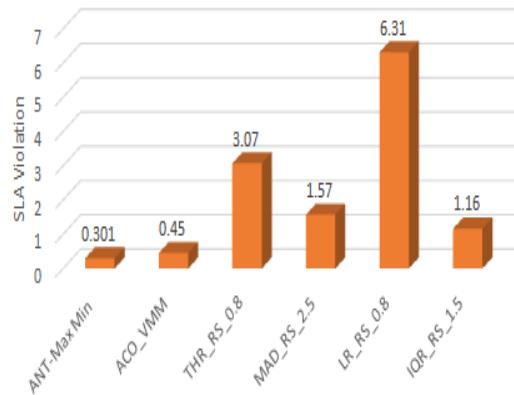


Figure 3 Comparison of Max-Min ACO with different Algorithms.

### B. Number of Migrations

Live VM Migration includes the Cost of Ram and Hard disk, so it is a costly operation. It also include the CPU utilization, link bandwidth, downtime of services and total migration time, so one of our main objective is to minimize the number of Migrations. Proposed algorithm gives better results as compare to other heuristic algorithms.

Comparison of Proposed algorithm with different algorithms w.r.t number of migrations

ANT MAXMIN	ACO_VMM	THR_RS_0.8	MAD_RS_2.5	LR_RS_0.8	IQR_RS_1.5
1050	5530	12530	12869	1306	13124

TABLE II

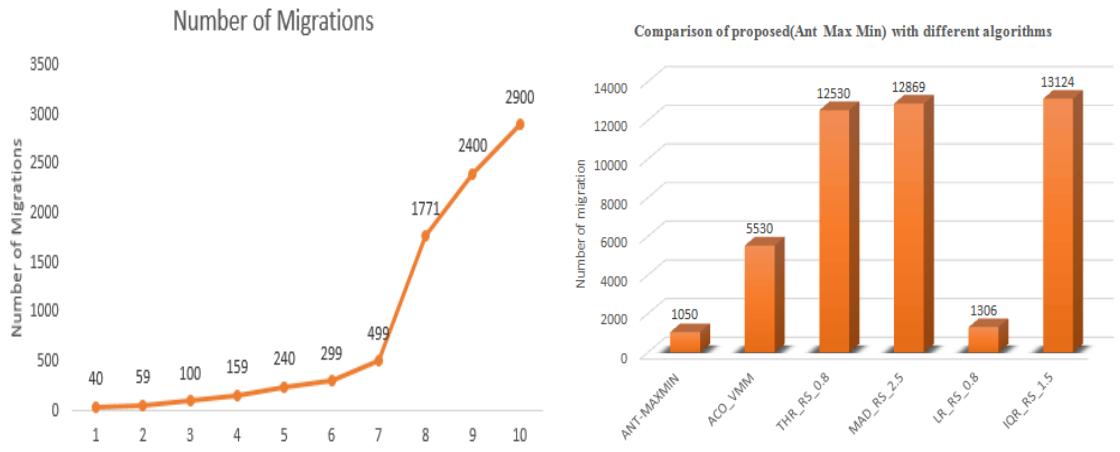


Figure 4 Number of Migrations of Ant Max-Min ACO.

Figure 5 Comparison of Ant Max-Min with different Algorithms

## V. CONCLUSION

The performance of host is degraded due to unbalanced load on the VMs. degradation. Therefore, load balancing algorithms through Virtual Machine Migration are crucial to ensure performance, balance loads evenly on hosts and proper utilization of resources. To achieve the load balancing goal in this paper, we propose a hybrid Ant Colony Optimization (ACO) algorithm with Max-Min Ant system for load balancing of virtual machines in cloud environment. Specifically, we introduce Max-min ant colony optimization ideas for performance improvement with optimal solution. Our research on Load balancing in cloud using Max-min ant colony optimization concentrates on efficient load balancing algorithm.

In future works, we would like to explore more intelligent algorithms and their applications in cloud computing and big data fields like PSO, ABPSO, ABC or hybrid this approach with some heuristic algorithms like Genetic Algorithm.

## REFERENCES

- [1] Tang, Linlin, Jeng-Shyang Pan, Yuanyuan Hu, Pingfei Ren, Yu Tian, and Hongnan Zhao, "A Novel Load Balance Algorithm for Cloud Computing," In International Conference on Genetic and Evolutionary Computing, pp. 21-30. Springer International Publishing, 2015.
- [2] Razali, Rabiatul Addawiyah Mat, Ruhani Ab Rahman, Norliza Zaini, and Mustaffa Samad. "Virtual machine migration implementation in load balancing for Cloud computing." In Intelligent and Advanced Systems (ICIAS), 2014 5th International Conference on, pp. 1-4. IEEE, 2014.
- [3] Liaqat, Misbah, Shalini Ninoriya, Junaid Shuja, Raja Wasim Ahmad, and Abdullah Gani, 'Virtual Machine Migration Enabled Cloud Resource Management: A Challenging Task,' arXiv preprint arXiv:1601.03854 (2016).
- [4] Wen, Wei-Tao, Chang-Dong Wang, De-Shen Wu, and Ying-Yan Xie, "An ACO-Based Scheduling Strategy on Load Balancing in Cloud Computing Environment," In 2015 Ninth International Conference on Frontier of Computer Science and Technology, pp. 364-369. IEEE, 2015.
- [5] Dorigo, Marco, Gianni Di Caro, and Luca M. Gambardella. "Ant algorithms for discrete optimization." Artificial life 5, no. 2 (1999): 137-172.
- [6] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. D. Rose, and R. Buyya, "Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," Software: Practice and Experience, vol. 41, no. 1, pp. 23 – 50, 2011.
- [7] J.-J. LIU, G.-L. CHEN, and C.-X. HU, "Virtual machine migration scheduling strategy based on load characteristic," Computer Engineering, vol. 37, no. 17, pp. 276–278, 2011.
- [8] Ramezani, Fahimeh, Jie Lu, and Farookh Khadeer Hussain, "Task-based system load balancing in cloud computing using particle swarm optimization," International Journal of Parallel Programming 42, no. 5 (2014): 739-754.
- [9] Chen, Kun-Ting, Chien Chen, and Po-Hsiang Wang, "Network aware load-balancing via parallel VM migration for data centers," In 2014 23rd International Conference on Computer Communication and Networks (ICCCN), pp. 1-8. IEEE, 2014.
- [10] Lu, Peng, Antonio Barbalace, Roberto Palmieri, and Binoy Ravindran, "Adaptive live migration to improve load balancing in virtual machine environment," In European Conference on Parallel Processing, pp. 116-125. Springer Berlin Heidelberg, 2013.
- [11] Babu, KR Remesh, Amaya Anna Joy, and Philip Samue, "Load Balancing of Tasks in Cloud Computing Environment Based on Bee Colony Algorithm," In 2015 Fifth International Conference on Advances in Computing and Communications (ICACC), pp. 89-93. IEEE, 2015.
- [12] Gutierrez-Garcia, J. Octavio, and Adrian Ramirez-Nafarrate, "Agent-based load balancing in Cloud data centers," Cluster Computing 18, no. 3 (2015): 1041-1062.
- [13] Li, Liang, Shi-chun Chi, and Gao Lin, "The complex method based on ant colony algorithm and its application to the slope stability analysis," Chinese Journal Of Geotechnical Engineering-Chinese Edition- 26 (2004): 691-696.
- [14] Yanhui, Wang, Xiao Xuemei, Jia Limin, and Qin Yong, "The Research on Searching Path Algorithm of Interoperable Recommendation Trust Based on MMAS," In Computer Modeling and Simulation, 2010. ICCMS'10. Second International Conference on, vol. 1, pp. 423-427. IEEE, 2010.
- [15] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers, <http://onlinelibrary.wiley.com/doi/10.1002/cpe.1867/full>," 2011.

# Region Of Interest based lossless and near lossless Image Compression

Sandeep Kaur<sup>#1</sup>, Gianetan Singh Sekhon<sup>\*2</sup>

<sup>#</sup>*Ycoe Talwandi Sabo, Punjabi University  
Patiala, Punjab, India.*

<sup>1</sup>*bhullarsandeep37@yahoo.com*

<sup>\*</sup>*Assistant Professor, Ycoe Talwandi Sabo  
Punjabi University Patiala, India*

<sup>2</sup>*gianetan@gmail.com*

**Abstract**— Image compression is a method through which we can reduce the storage space of images, videos which will helpful to increase storage and transmission process's performance. In image compression, we do not only concentrate on reducing size but also concentrate on doing it without losing quality and information of image. Image compression may be lossy and lossless. In lossless compression the exact original data to be reconstructed from the compressed data. Lossless is in contrast to lossy data compression, which only allows constructing an approximation of the original data, in exchange for better compression rates. This research is about the image compression based upon the wavelet compressions, and use both lossless and near lossless compression. In this research a Fuzzy Measure based classifiers are specified to classify image dataset for image compression.

**Keywords**— Image Compression, Lossless compression, Near Lossless compression

## I. INTRODUCTION

Digital image process is that the use of pc algorithms to perform image process on digital pictures. As a subcategory or field of digital signal process, digital image process has several blessings over analog image process. It permits a way wider vary of algorithms to be applied to the computer file and may avoid issues corresponding to the build-up of noise and signal distortion throughout process.[1] Since pictures square measure outlined over 2 dimensions (perhaps more) digital image process could also be sculptural within the sort of four-dimensional systems. compression is minimizing the dimensions in bytes of a graphics file while not degrading the standard of the image to associate unacceptable level.[1] The reduction in file size permits additional pictures to be hold on during a given quantity of disk or memory house. It additionally reduces the time needed for pictures to be sent over the web or downloaded from web content.

There square measure many other ways within which image files is compressed. For net use, the 2 commonest compressed graphic image formats square measure the JPEG format and also the GIF format.[2] The JPEG technique is additional usually used for images, whereas the GIF technique is often used for line art and alternative pictures within which geometric shapes square measure comparatively straightforward.[2]

Other techniques for compression embody the employment of fractals and wavelets. These ways haven't gained widespread acceptance to be used on the web as of this writing. However, each ways provide promise as a result of they provide higher compression ratios than the JPEG or GIF ways for a few forms of pictures. Another new technique that will in time replace the GIF format is that the PNG format.[2]

A document or program is compressed while not the introduction of errors, however solely up to an explicit extent. this is often referred to as lossless compression. on the far side this time, errors square measure introduced. In text and program files, it's crucial that compression be lossless as a result of one error will seriously harm the which means of a document, or cause a

program to not run.[3] In compression, atiny low loss in quality is typically not noticeable. there's no "critical point" up to that compression works utterly, however on the far side that it becomes not possible. once there's some tolerance for loss, the compression issue is bigger than it will once there's no loss tolerance. For this reason, graphic pictures is compressed quite text files or programs.

#### A. **IMAGE CLASSIFICATION**

Classification between the objects is straightforward task for humans however it's evidenced to be a fancy downside for machines. The raise of high-capacity computers, the supply of prime quality and cheap video cameras, and also the increasing want for automatic video associate analysis has generated an interest in object classification algorithms. a straightforward organisation consists of a camera mounted high on top of the interested zone, wherever pictures area unit captured and consequently processed.[3] Classification includes image sensors, image pre-processing, object detection, object segmentation, feature extraction and object classification. organisation consists of information that contains predefined patterns that compares with detected object to classify in to correct class. Image classification is a very important and difficult task in varied application domains, together with medicine imaging, biometry, video police investigation, vehicle navigation, industrial visual review, golem navigation, and remote sensing.[4] Classification method consists of following steps:

- A. Pre-processing- atmospherically correction, noise removal, image transformation, main element analysis etc.
- B. Detection and extraction of a object Detection includes detection of position and different characteristics of moving object image obtained from camera. And in extraction, from the detected object estimating the mechanical phenomenon of the item within the image plane.
- C. Training: choice of the actual attribute that best describes the pattern.
- D. Classification of the object-Object classification step categorizes detected objects into predefined categories by mistreatment appropriate methodology that compares the image patterns with the target patterns.[4]

Fuzzy Measure: In Fuzzy classification, varied random associations area unit determined to explain characteristics of a picture. the assorted sorts of random area unit combined (set of properties) during which the members of this set of properties area unit fuzzy in nature. It provides the chance to explain totally different classes of random characteristics within the similar type. It uses random approach. Performance and accuracy depends upon the edge choice and fuzzy integral.[5]

Advantages: expeditiously handles uncertainty. properties area unit describe by characteristic varied random relationships.

#### B. **REGION OF INTEREST**

A region of interest (often abbreviated ROI), could be a elite set of samples inside a dataset known for a selected purpose.[1] The construct of a ROI is often utilized in several application areas. as an example, in medical imaging, the boundaries of a tumour could also be outlined on a picture or in an exceedingly volume, for the aim of measure its size. The endocardial border could also be outlined on a picture, maybe throughout totally different phases of the oscillation, as an example end-systole and end-diastole, for the aim of assessing viscous operate. In laptop vision and optical character recognition, the ROI defines the borders of associate object into consideration. In several applications, symbolic (textual) labels area unit intercalary to a ROI, to explain its content in an exceedingly compact manner. inside a ROI might lie individual points of interest (POIs).[1]

A ROI could be a sort of annotation, typically related to categorical or quantitative data (e.g., measurements like volume or mean intensity), expressed as text or in structured type.

There area unit 3 essentially totally different means that of coding a ROI:

As associate integral a part of the sample knowledge set, with a novel or masking price that will or might not be outside the traditional vary of unremarkably occurring values and that tags individual knowledge cells

As separate, strictly graphic data, corresponding to with vector or electronic image drawing components, maybe with some incidental plain (unstructured) text within the format of the info itself.[6]

As a separate structured linguistics data (such as coded price types) with a collection of abstraction and/or temporal coordinates.[7]

Salience Mapping primarily based Region of Interest: prominence technique uses color, intensity, and orientation parameters in an exceedingly manner compatible with human attention behavior. This technique removes unrelated regions, that have nearly uniform attribute in a picture. Color, orientation, and intensity area unit traditional options for prominence detection. There area unit primarily 3 steps in prominence map technique; initial step is extraction of feature vectors at locations over the image. Next step is activation of maps victimisation those feature vectors; final step is standardization, that normalizes the activation map.

#### C. **IMAGE COMPRESSION**

Image compression is minimizing the dimensions in bytes of a graphics file while not degrading the standard of the image to an unacceptable level. The reduction in file size permits additional pictures to be keep during a given quantity of disk or memory area. It conjointly reduces the time needed for pictures to be sent over the net or downloaded from sites.

There are many alternative ways within which image files is compressed. For net use, the 2 commonest compressed graphic image formats are the JPEG format and also the GIF format. The JPEG methodology is additional usually used for pictures, whereas the GIF methodology is often used for line art and different pictures within which geometric shapes ar comparatively easy.[8]

Other techniques for compression embrace the employment of fractals and wavelets. These strategies haven't gained widespread acceptance to be used on the net as of this writing. However, each strategies provide promise as a result of they provide higher compression ratios than the JPEG or GIF strategies for a few varieties of pictures. Another new methodology which will in time replace the GIF format is that the PNG format.

#### D. **IMAGE COMPRESSION TYPES**

Image compression is can be lossy or lossless. lossless compression is most well-liked for depository functions and sometimes for medical imaging, technical drawings, clip art, or comics. lossy compression strategies, particularly once used at low bit rates, introduce compression artifacts. lossy strategies are particularly appropriate for natural pictures reminiscent of images in applications wherever minor (sometimes imperceptible) loss of fidelity is appropriate to attain a considerable reduction in bit rate. The lossy compression that producible variations is also known as visually lossless.[8]

Methods for lossless compression are:

- Run-length cryptography – employed in default methodology in PCX and jointly of attainable in BMP, TGA, TIFF
- Area compression
- DPCM and prognostic committal to writing
- Entropy cryptography
- Adaptive lexicon algorithms reminiscent of LZW – employed in GIF and run-in
- Deflation – employed in PNG, MNG, and TIFF
- Chain codes

Methods for lossy compression:

- Reducing the colour house to the foremost common colours within the image. the chosen colours ar per the colour palette within the header of the compressed image. Every element simply references the index of a colorize the colour palette, this methodology may be combined with video digitizing to avoid posterization.
- Chroma subsampling. This takes advantage of the very fact that the human eye perceives spatial changes of brightness a lot of sharply than those of color, by averaging or dropping a number of the chrominance info within the image.
- Transform committal to writing. this is often the foremost unremarkably used methodology. specifically, a Fourier-related rework reminiscent of the discrete cosine rework (DCT) is wide used: N. Ahmed, T. Natarajan and K.R.Rao, "Discrete cos rework," IEEE Trans. Computers, 90-93, Jan. 1974. The DCT is usually mentioned as "DCT-II" within the context of a family of discrete cosine transforms; e.g., see discrete cosine network. The a lot of recently developed riffle rework is additionally used extensively, followed by division and entropy committal to writing.
- Fractal compression.

Wavelets ar signals that ar native in time associated scale and customarily have an irregular form. A riffle could be a wave form of effectively restricted length that has a median price of zero. The term ‘wavelet’ comes from the very fact that they integrate to zero; they wave up and down across the axis. several wavelets conjointly show a property ideal for compact signal representation: orthogonality. This property ensures that knowledge isn't over delineated . a symbol may be rotten into several shifted and scaled representations of the initial mother riffle. A riffle rework may be accustomed decompose a symbol into part wavelets. Once done the coefficients of the wavelets can be decimated to get rid of a number of the main points. Wavelets have the nice advantage of having the ability to separate the fine details in an exceedingly signal. terribly little wavelets may be accustomed isolate terribly fine details in an exceedingly signal, whereas terribly massive wavelets will determine coarse details.

**E. HAAR WAVELET COMPRESSION[8]**

Alfred Alfred Haar (1885-1933) was a Hungarian man of science UN agency worked in analysis finding out orthogonal systems of functions, partial differential equations, Chebyshev approximations and linear inequalities. In 1909 Haar introduced the Haar ripple theory. A Haar ripple is that the simplest variety of ripple . In distinct kind, Haar wavelets square measure concerning a computing known as the Haar remodel.. The mathematical prerequisites are unbroken to a minimum; so, the most ideas is understood in terms of addition, subtraction and division by 2. We tend to additionally gift a algebra implementation of the Haar ripple remodel, and mention necessary recent generalizations. Like all ripple transforms, the Haar remodel decomposes a distinct signal into 2 subsignals of 0.5 its length. The Haar ripple remodel encompasses a range of advantages:

- It is conceptually straightforward and quick.
- It is memory economical, since it is calculated in situ while not a brief array.
- It is precisely reversible while not the sting effects that square measure a drag with different ripple transforms.
- It provides high compression magnitude relation and high PSNR (Peak signal to noise ratio).
- It will increase detail in a very algorithmic manner.

The Haar remodel (HT) is one in every of the best and basic transformations from the area domain to an area frequency domain. A HT decomposes every signal into 2 parts, one is named average (approximation) or trend and also the different is thought as distinction (detail) or fluctuation. information compression in transmission applications has become additional important recently wherever compression strategies square measure being chop-chop developed to compress massive

information files akin to pictures. Economical strategies sometimes reach pressure pictures, whereas retentive high image quality and marginal reduction in image size. Recently the utilization of ripple remodels and distinct cosine Transform (DCT) for compression was investigated [1]. The final purpose of compression systems is to compress pictures, however the result's below optimum. though the utilization of ripple Transforms was shown to be additional superior to DCT once applied to compression, a number of the finer details within the image is sacrificed for the sake of saving a bit additional information measure or cupboard space. This additionally implies that lossy compression techniques akin to DCT is utilized in this space. compression mistreatment DCT may be a straightforward compression methodology that was initial applied in 1974 [2]. it's a preferred remodel used for a few of the compression standards in lossy compression strategies. The disadvantage of mistreatment DCT compression is that the high loss of quality in compressed pictures, That is additional notable at higher compression ratios.[1][2]

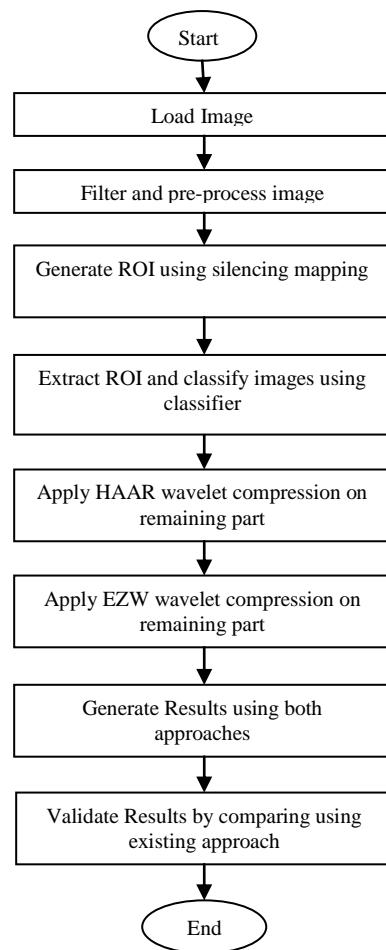
#### F. EZW COMPRESSIONS

Discrete riffle rework (DWT) provides a multiresolution image illustration and has become one in every of the foremost necessary tools in image analysis and writing over the last 20 years. compression algorithms supported dB 9/7 give high writing potency for natural (smooth) pictures. An elementary shift within the compression approach came when the distinct riffle rework (DWT) became in style. Riffle supported compression provides a awfully effective technique for medical pictures. Medical image needs storage of enormous amount of digitized clinical information. because of the high information measure and capability of storage, medical image should be compressed before transmission. a preferred methodology of compression, namely, the embedded zero tree riffle (EZW) that has less degree of loss. In lossy compression, image characteristics square measure sometimes preserved within the coefficients of the domain house in to that the initial image is reworked. the standard of the image when compression is incredibly necessary and quality loss should be inside the tolerable limits that vary from image to image and methodology to methodology, therefore the compression becomes additional attention-grabbing as a locality of analysis of various varieties of medical compression techniques. A preferred methodology of compression is that the embedded zero tree riffle. progressively, medical pictures square measure non-inheritable and hold on digitally. These pictures could also be terribly massive in size, number. Compression offers a method to scale back the value of storage and increase the speed of transmission. Compression minimizes the dimensions in bytes of a graphics file while not degrading the standard of the image. The resolution in file size permits additional pictures to be hold on during a given quantity of disk or memory house. It additionally reduces the time needed for pictures to be sent over the net or transfer from WebPages.

## II. METHODOLOGY

In this section various steps are defined to achieve the image compression based upon the silencing mapping Region of Interest Technique. First of all the image is converted to synthetic image i.e. pre-processed image in which the noise is removed from the image. After pre-processing step the next one is to generate ROI using silencing mapping technique in which the image properties are stored using a test data set and those properties are used to generate the ROI from the other images.

When the Region of Interest is extracted, various compression techniques are applied on that image. Except the region of interest the other portion of image is compressed using the above defined compression techniques i.e. HAAR wavelet for lossless compression and EZW for near lossless compression. After the image compression various metrics like Mean Square Error, Peak Signal to Noise Ration and contrast etc, are calculated to generate and validate the results that is shown in the next section.



**Fig 1: Flow Chart**

### III. RESULTS AND DISCUSSION

In this research work various parameters are considered for results validation part like Mean Square Error, Peak Signal to Noise Ratio and Contrast.



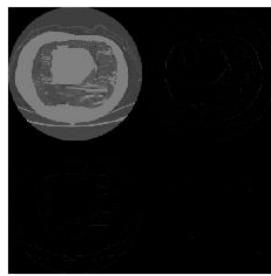
**Fig 2: Original Image**

Fig 3 is the presentation of original image from which ROI is to be extracted and compression is to be done.



**Fig 3: Region of Interest**

Fig 4 is the presentation of Region of Interest in image to detect tumor in case of medical image. Now the other parts of the image are to be compressed except the tumor.



**Fig 4: HAAR Wavelet Compression**

Fig 5 is the compression of image using HAAR wavelet. Following are the parameters that are to be calculated.

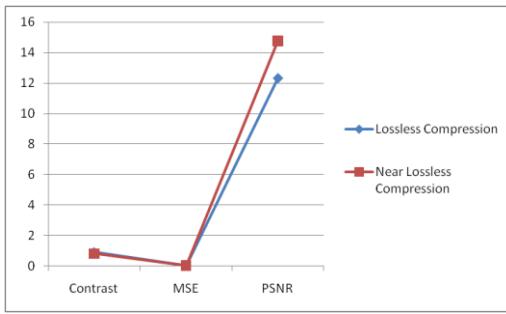
**Mean Square Error:** Mean Squared Error (MSE) or mean squared deviation (MSD) of an estimator measures the average of the squares of the errors or deviations, that is, the difference between the estimator and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate.

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2$$

**Peak Signal to Noise Ratio:** Peak signal-to-noise ratio, often abbreviated PSNR, is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel scale.

$$\begin{aligned} PSNR &= 10 \cdot \log_{10} \left( \frac{MAX_I^2}{MSE} \right) \\ &= 20 \cdot \log_{10} \left( \frac{MAX_I}{\sqrt{MSE}} \right) \\ &= 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE) \end{aligned}$$

**Contrast:** Contrast is the difference in luminance or color that makes an object (or its representation in an image or display) distinguishable. In visual perception of the real world, contrast is determined by the difference in the color and brightness of the object and other objects within the same field of view.



**Fig 5: Comparative Study for lossless and near lossless compression**

#### IV. CONCLUSION

This work has shown that the compression of image can be improved by considering frequency domain redundancy. The efficiency of lossless compression and near lossless compression are far better than that of lossy compressions. This research is based upon the comparison of lossless and near lossless compression using HAAR wavelet transformation and EZW. To

achieve the compression techniques pre-processing of images are done using filtration techniques. After filtration process classifiers are applied to support the ROI technique. Salience based ROI is calculated to maintain the quality of interested region as compare to outliers in image. From graph it is clear that the metrics in the lossless compression is better as compare to near lossless compression.

#### ACKNOWLEDGEMENT

I am thankful to my respected guide Mr.Gianetan Singh Sekhon, Assistant Professor (Computer Engineering), Yadavindra College of Engineering, Talwandi Sabo for his invaluable and enthusiastic guidance, useful suggestions

#### V. REFERENCES

- [1]. Manpreet Kaur, Vikas Wasson, "ROI Based Medical Image Compression for Telemedicine Application", Proceedings of the 4th International Conference on Eco-friendly Computing and Communication Systems, Volume: 70, 2015, pp: 579-585
- [2]. Vinayak K Bairagi, Ashok M Sapkal, "ROI-based DICOM image compression for telemedicine", Springer, Vol. 38, Issue: 1, February 2013, pp. 123–131
- [3]. Amandeep Kaur, Monica goyal, Student, "A Review: ROI based Image Compression of Medical Images", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2347-8578, Volume 2 Issue 5, Sep-Oct 2014, pp: 162-166
- [4]. J.C. Garcia,H.Fuhr, G.Castellanos-Dominguez, "Evaluation of Region-of-Interest coders using perceptual image quality assessments" , Journal of Visual Communication and Image Representation, Volume 24, Issue 8, November 2013, Pages 1316–1327
- [5]. Pratik Chavada, Narendra Patel, Kanu Patel, "Region of Interest Based Image Compression", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN(Online): 2320-9801, ISSN (Print): 2320-9798, Volume: 2, Issue: 1, January 2014, pp: 2747-2754
- [6]. Peilong Zhao, Jiwen Dong, Lei Wang, "Image Compression Algorithm Based On Automatic Extracted ROI", Fuzzy Systems and Knowledge Discovery (FSKD), 11th International Conference, Xiamen, 2014, pp: 788-792
- [7]. Kuldip K. Ade, M.V.Raghunadh, "ROI based Near Lossless Hybrid Image Compression Technique", Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference, Coimbatore, 5-7 March 2015, pp: 1-5
- [8]. Rushabh R.Shah, Priyanka Sharma, "Performance analysis of region of interest based compression method for medical images", Fourth International Conference on Advanced Computing & Communication Technologies, Rohtak, 8-9 Feb. 2014, pp: 53-58
- [9]. Sarvarinder Singh, Sarabjit Singh ,Guru Kashi University, Talwandi Sabo, "Review; ROI Based Medical Image Compression using DCT and HAAR Wavelet", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN (Online) 2278-1021, ISSN (Print) 2319 5940, Vol. 4, Issue 8, August 2015, pp: 469-472

# Improving Store and Access performance in Heterogeneous Environment using MapReduce

Kamalpreet Kaur<sup>1</sup>, Er. Gurjit Singh Bhathal<sup>2</sup>

*Department of Computer Engineering, Punjabi University, Patiala*

Email id<sup>1</sup>- kamalpreetkang8@gmail.com

Email id<sup>2</sup>- gurjit.ce@pbi.ac.in

**Abstract-** A big data term is used for large information or data sets which are gathered from government- private organizations, education- technical field, social networks, and from so many resources which generate some informational data. All these resources are produced a huge amount of data sets. There is very critical problem of a big data- is to store all the data sets on a single computer. To solve this problem Big data term is used a cluster of computers to store and process the structured and unstructured data. Hadoop framework gives a well suitable solution of big data's problem. This paper represents the Hadoop framework and its various tools that are well- designed to store and access the large data sets. This paper is also discussing about the MapReduce program.

**Keywords-** Big Data, Hadoop Framework, HDFS, MapReduce.

## I INTRODUCTION

Big data is of collection of massive amount of digital information which is collected from various social networks government organization finance and retail (stock exchange data), education field and search engine data etc. this data is not only produced by information exchange and computers software, mobile phones so on. But it also from various sources, sensors which are embedded in various environment like every virtual electric device, street cameras, temperature sensors and so many things which produced some information for others. This digital data is defined by various V's like volume- size of data, velocity- transfer rate of data, verity- types of data. Big data is most popular to describe the continuous growth of data and availability of data both structured, semi-structured and unstructured. The data in it will these types

- Structured – relation data
- Semi-structured – XML data
- Unstructured – word, pdf, text media

There are some challenges or problems in big data that are as store, searching, processing and analysis and presentation. To fulfill the above challenges or to solve the problems, big data has a solution that is Hadoop.

The traditional approach to store and process the data on computer system is that the information or data will be stored into RDBMS like Oracle, MySQL etc.

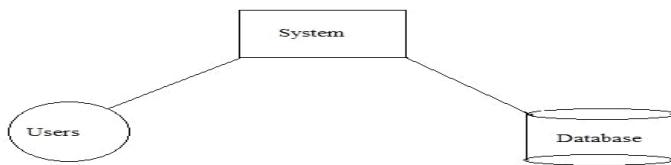


Figure 1 Traditional Approach

The main disadvantage of this traditional approach is that it is not compatible for large amount of data sets. Google give the solution of this problem. By the program of map-reduce. In map-reduce they split the data into small parts stored on cluster of a computer, which are connected through a network. From their users can access the huge amount of data very easily and quickly. They collect the result to produce a final result dataset.

## II HADOOP

Hadoop is open-source software. Hadoop framework is a registered trademark of Apache foundation which is written in java programming language. Mike Cafarella and Duge cutting started Hadoop in 2005. Apache Hadoop framework is wide popular for parallel distribution processing of large data sets. A cross cluster of computers by using simple programs. In big data compute infrastructure Hadoop is most suitable for batch processing. Because Hadoop software for reliable, scalable of distribution computing. Apache Hadoop framework is basically designed to scale up from thousands of machines. Each machine offers to process and storage of data and also implemented to detect and handle the hardware failures as well as failures at application layer. The Apache Hadoop consists of two parts. Storage part which is known as HDFS and processing part which is also known as MapReduce. In the Hadoop framework various tools are Apache Hbase, Pig, Zookeeper, Hive, Spark, Madout, Avro, Oozie and so on.

- A. **APACHE HBASE-** Apache Hbase is open source software. It is a Hadoop database which stores the huge amount of distributed into large table sets. Apache Hbase is a column oriented database which has millions of column that store the distributed data. Hbase is built on top of Hadoop distributed file system. They support the NO-SQL because there is no any relation between tables of database. Apache Hbase is same as a Google's big data. Hbase is designed to provide a quick random access of massive amount of structured data. Apache Hbase provide a high latency (time period between when an input is occurred and first response to the input event is needed) access to single row/column from billions of records therefore it provides a random real time read/write access to data. Hbase provide the consistency, availability, good speed and horizontal scale capabilities.
- B. **HIVE-** Apache Hive is a data warehouse infrastructure built on Hadoop. It is a platform to summarize the big data and make queries and analyzed. A Hive is not a relational database but is similar to relational database concepts such as tables and partitions. Hive is structured query language known as Hive QL for the unstructured data of Hadoop. Hive queries are compiled into map-reduce program using Hadoop.

- C. *PIG*- Apache Pig is a tool that provides program to analyze the large set of data. It provides a high level language program that analyze the data sets is known as Pig Latin. Pig offers run-time platform that provide a facility to use the MapReduce execute on Hadoop framework.
- D. *ZOOKEEPER*- Apache Zookeeper is open-source software built on the top level of Hadoop. Apache Zookeeper is a centralized service for maintaining distributed configuration service, synchronization service and provides a naming space to distributed system. In zookeeper platform users can read and write the data from nodes, because zookeeper uses a hierarchical structure or tree data structure.
- E. *APACHE SPARK*- Apache Spark is also an open source computing framework which is developed by the university of California, Berkeley's AMP Lab. Apache Spark is 100 times fast as compare to Hadoop MapReduce. Because Hadoop map-reduce systems are based on disk and they support batch processing, but Apache spark support streaming processing. In spark framework data reside in cache memory. Therefore, it gives faster results than Hadoop. The main aim of spark is speeding up of data analytics in runtime as well as in development. A spark is very fast and general computation engine which can access any data sources of Hadoop. Therefore, Apache spark is widely used by organizations to process a huge amount of data.

In Hadoop framework has following various modules-

- A. *HADOOP COMMON*- Apache Hadoop common provides the utilities to other Hadoop modules.
- B. *HADOOP YARN*- Apache Hadoop yarn is a resource management platform. They provide the job scheduling in clusters.
- C. *HDFS*- A distributed program files are stored and that provide the high throughput access to application data. HDFS replicates the data in various storage nodes therefore users can access the distributed data concurrently. HDFS run on top of local file system of a cluster of Hadoop and store very large data files which is suitable for streaming data access. HDFS provides high fault-tolerance and it provides master slave architecture. HDFS has two nodes- master node (name node) and slave node (data node).
  - a) A name node manages the hierarchy of file system. A master node allows file system operations like modifications, closing and opening of files, access time and so on.
  - b) Data node performs read-write operations on a file system. It allows performing block creation, deletion and updating operation to its users, but according to name node' instructions.
- D. *HADOOP MAPREDUCE*- MapReduce is a parallel programming model that was designed to process or produced large data sets on clusters of computers. MapReduce program is basically based on divide-conquer method. A map-reduce program splits the input data sets into various independent parts. Map-reduce are mostly used by Google, Yahoo and other many web organizations. In map-reduce program has two steps. In first step a complex problem is dividing into several parts in a key/ value pair and these sub-problems are solve directly then assigned to cluster of working nodes. The problems are solving and independent from each other. Final step of MapReduce program is to combine the solution of these various

sub-problems and produces a final single result of large data sets. There are two functions which are followed by map-reduce program.

- a) **MAP FUNCTION-** in a Hadoop framework map function done its job into two nodes that are master node and worker node. A programmer gives the input and master node takes the input data and spit into various sub-parts after that master node distribute these sub problems to worker nodes.
- b) **REDUCE FUNCTION-** Reduce function helps to collect the several problems with its solution and combined them in a specific way to form of a final output. In reduce function they provide a list of key/values.

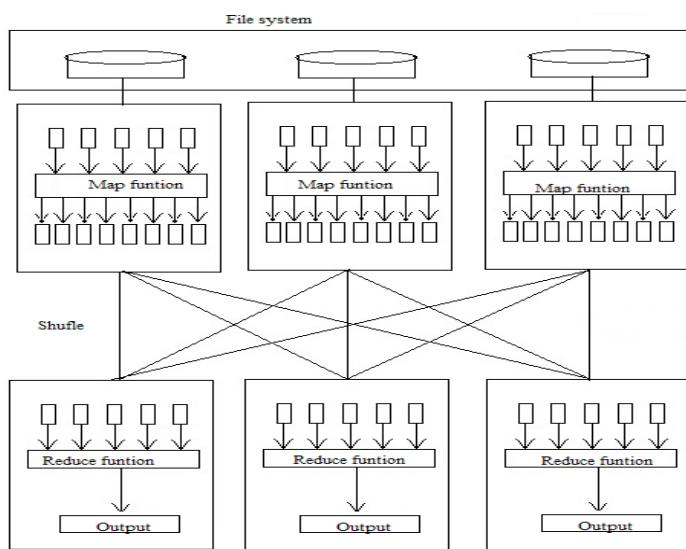


Figure 2MapReduce

We take an example of online shopping site when a customer reaches to an online shopping site then it seen many products like electronics devices, Clothing, footwear's (with different sizes and for both men and women), beauty products, daily usage products and so many. But a customer wants to buy women blue jeans and women white shirt. Then a MapReduce program is very helpful to customer. It sorts those items from a large data set which is customer wants to purchase.

### III CONCLUSION

This paper describes the big data 'problem- to storing and accessing large data sets. A massive amount of data sets is very difficult to store and process on a single computer. Hadoop framework gives the solution of this problem. All the large data sets are distributed store in HDFS. MapReduce is very useful for parallel processing of these distributed data sets. It provides a high throughput. Hadoop framework provides various tools which are useful to store and access the huge data sets.

**ACKNOWLEDGEMENT**

I sincerely thank to Er. Gurjit Singh Bhathal for their guidance and encouragement to carrying out to this research paper.

**REFERENCES**

- [1] (2016, Mrach). Retrieved from Youtube: <https://www.youtube.com/watch?v=A02SRdyoshM>
- [2] (2016, March 11). Retrieved from Webopedia : [http://www.webopedia.com/TERM/A/apache\\_hbase.html](http://www.webopedia.com/TERM/A/apache_hbase.html)
- [3] (2016, March). Retrieved from webopedia: [http://www.webopedia.com/TERM/B/big\\_data.html](http://www.webopedia.com/TERM/B/big_data.html)
- [4] (2016, Feb.). Retrieved from Inform IT: <http://www.informit.com/articles/article.aspx?p=2253412>
- [5] *Apache Hadoop Foundation*. (n.d.). Retrieved JULY 2016, from <http://hadoop.apache.org/>
- [6] B.Kezia Rani, D. R. (2015). Cloud Computing And Inter-cloud-Types, topologies annd resaerch Issuses. (S. Direct, Ed.) doi:, 24-29. doi:10.1016/j.procs.2015.04.006
- [7] C.L. Philip Chen, .. C.-Y. ( 2014, August 10 ). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. (S. Direct, Ed.) 275, 314–34. doi:10.1016/j.ins.2014.01.015
- [8] *Hadoop*. (n.d.). Retrieved july 19, 2016, from <https://opensource.com/life/14/8/intro-apache-hadoop-big-data>
- [9] Ibrahim Abaker Targio Hashema, .. ., (2015, January ). The rise of “big data” on cloud computing: Review and open research issues. (S. Direct, Ed.) 47, 98-115. doi:10.1016/j.is.2014.07.006
- [10] Ishwarappa, A. J. (2015). Periodical Computer Scinence. *A Briefe Introduction on Big Data's 5 v's and Hadoop Technology*, pp. 319-324. Retrieved july 2016
- [11] Karthik Kambatlaa, . . . (2014, July ). Trends in big data analytics. (S. Direct, Ed.) 74(7), 2561–2573. doi:10.1016/j.jpdc.2014.01.003
- [12] Katarina Grolinger, W. A. (2013, dec. 18). Data management in cloud environments: NoSQL and NewSQL data stores. *Journal of Cloud Computing: Advances, Systems and ApplicationsAdvances, Systems and Applications*. doi:10.1186/2192-113X-2-22
- [13] Marcos D. Assuncao, R. N. (2014, August 18). Big DataComputing and cloud: Trends and future directions. (S. direct, Ed.) 3-15. doi:doi.org/10.1016/j.jpdc.2014.08.003
- [14] Mohd Rehan Ghazi, D. G. (2015). Hadoop, MapReduce and HDFS: A Developers Perspective. (S. Direct, Ed.) *Procedia Computer Science*, 48, 45-50. doi:10.1016/j.procs.2015.04.108
- [15] Prantosh Kumar Paul, M. K. (2012). Cloud Computing: possibilities, challenges and opportunities with special reference . (S. Direct, Ed.) 2222-2227. doi:10.1016/j.proeng.2012.06.267
- [16] Praveen Murthy, A. B. (2014, September). Big Data Taxonomy. 1-33. Retrieved from <http://cloudsecurityalliance.org/research/big-data/>
- [17] *SAS*. (n.d.). Retrieved july 25, 2016, from www.sas.com: [http://www.sas.com/en\\_my/insights/big-data/what-is-big-data.html](http://www.sas.com/en_my/insights/big-data/what-is-big-data.html)
- [18] *Tutorial Ponits*. (n.d.). Retrieved july 2016, from www.tutorialpoints.com: <http://www.tutorialspoint.com/hadoop/>
- [19] Vora, M. N. (2011, Dec. 26). Hadoop-HBase for large-scale data. *I*, 601 - 605. doi:10.1109/ICCSNT.2011.6182030
- [20] *Wikipedia*. (2016, January 7). Retrieved from [https://en.wikipedia.org/wiki/Apache\\_HBase](https://en.wikipedia.org/wiki/Apache_HBase)

# Rician noise removal in 3D MR Images using adaptive non-local means filter

Ram Singh

Assistant Professor, Department of Computer Engineering  
Punjabi University, Patiala-147002  
[bhankharz@gmail.com](mailto:bhankharz@gmail.com)

**Abstract –** Magnetic Resonance Images (MRIs) are contaminated with Rician distributed noise during acquisition process. Unlike Additive Gaussian distributed noise, noise in Magnetic Resonance Images follows the Rician distribution. Rician noise is signal dependent. In the high contrast region of the MR images, noise follows Gaussian distribution whereas in the low contrast regions it tends to be Rayleigh distribution. Therefore, separation of the Rician distributed noise in MR image signal is a challenging and difficult task. Non-Local Means (NLM) filter has been widely used to remove noise from the 2D natural images and 2D medical image sequences. In this paper, NLM filter is used for random noise filtering and adapted to remove the noise from MR image slices using redundancy of information in the image under study to remove the noise. Due to Rician nature of noise in MR images, noise reduction method is first applied to the squared magnitude of the images. The magnitude of MR images is the square root of the sum of squares of Gaussian distributed real and imaginary parts of the image data that follows Rician distribution. Experimental results show that given methods achieves better denoising performance over the other existing noise removal methods.

**Key-words:** Rician distribution, Random noise, block-wise MRI denoising.

## I. INTRODUCTION

During image acquisition, MR images are corrupted with random noise and voltage fluctuation artifacts. Major reasons for image signal degradation is current fluctuations in the imaging modalities circuitry, thermal noise of the subject body and movement of the subject during scanning. The presence of noise in MR image not only degrade the visual quality of the image but also lowers the visibility of low-intensity objects. This degradation of the image quality severely affects the perceptual and quantitative analysis for diagnosis purpose by the medical experts. It is possible to reduce the random degradation to the image data by increasing the number of signal averages during the subject scanning, but considering the condition and comfort of the patient and requiring the speed to complete the scanning process at the earliest, this is not a suitable clinical practice in MRI imaging establishments. This is particularly true for MRI and ultrasound (US) images in which small structures are barely detectable above the noise level i.e. 2% or 3%. Therefore, post-processing image noise reduction methods are more appropriate not to increase the acquisition time and number of scanning runs. Keeping in view of it, post processing image filtering methods are more acceptable for MRI denoising. Image noise reduction activity is one of the classical problems in digital image processing for which different methods and image denoising schemes have been extensively studied and many image denoising schemes have been proposed in the existing and upcoming literature.

In section-II, a short literature review is given on image denoising methods. Section-III represents the brief description of the NLM filter. Section IV and V show the experimental details, quality measure matrices, results outcomes of the filter and their comparison with other state-of-the art denoising algorithms on Gaussian and Rician distribution of noise. Experiments are performed on real data set such as Tesla T1, T2 weighted MRI sequences of well known Brainweb database. In section V, a discussion is given about the suitability to the medical images and further improvements of the NLM filter for denoising 3D medical images.

## II. RELATED WORK

Image denoising is used as a preprocessing step in different medical image processing applications such as image registration, edge detection, segmentation and to reduce the random noise arises from the acquisition process. Gaussian filtering [1] is extensively used in MRI preprocessing. This filter is capable to reduce some image noise, especially in homogenous image regions, but also removes high-frequency image data that results in blurring edges of the objects in the image. Therefore, this filter commonly used for regularization purposes in voxel-based morphometry [1].

A major problem with a noise reduction filter is preserving edges boundary details of object in the image retaining the low frequency high intensity pixel values. A variety of filters have been developed to overcome this edge blurring effects, but no single filter is capable so far to overcome for noise reduction while preserving edge details. For example, anisotropic diffusion filter [2-3] provided the solution to remove noise using gradient details while retaining important edge structures. In [5] a new anisotropic diffusion filter is proposed which is based on a linear minimum mean square (LMMSE) error estimation that uses partial differential equations for removing Rician noise and produced good results.

Further in transform domain, Wavelet-based filters have also been used successfully to denoise the MR images using Soft-Thresholding methods [5-6]. Wavelet based denoise an image decomposing it into multiple levels and scales which represents different time-frequency components of the input image signal. At each scale level, image transform coefficients are thresholding [7] and statistical modeling [8] is performed to suppress noise and then from the thresholded components at each scale, final denoised image is reconstructed by inverse transform. Wavelet based image denoise methods introduces several visual artifacts in the filtered image. In the textured images the object edge boundaries are also become blurred in the image those further results in poor quantitative image analysis.

To overcome the limitations of wavelet transforms, a spatially adaptive principal component analysis (PCA) based denoising filters [9] [10] [11] proposed which computes the locally fitted basis to transform the image. All these transform based denoise filters derive from variations of the *transform-threshold-inverse* transform principles [12]. In accordance with the *transform-threshold-inverse* approaches, local transform approaches like sliding-window with or without overlapping produces very good results. In [19] the author has presented a DCT based weighted averaging of the transformed image coefficients to reduce Gaussian noise using linear overcomplete DCT dictionary coefficients thresholding which produced state-of-the art results.

The non-local means (NLM) filter was originally introduced by Buades et al. [14] for noise filtering from 2D images. A variety of versions of the NLM filter have been proposed to denoise 2D natural images. The NLM filter outperforms state-of-the art filtering Gaussian noise from 2D images such as Perona and Malik anisotropic

diffusion [15] and Translation Invariant wavelet thresholding [16]. In all these methods, the main limitation of the NLM filter is its computational complexity. In this paper an experimental adaptation of the NLM filter is extended to denoise 3D MR images patches called voxels, to process block-wise exploiting the self-similarity structures in the voxel regions to denoise the image.

The application of non-local means filters to 3D MR image reconstruction is pioneered by Coupe et al [18]. Weiest-Daessle et al. [19] adapted the NLM filter to remove the Gaussian distributed noise from Diffusion Weighted-MRIs. In practice, it has been observed and demonstrated that the noise contained in the MRIs follows a Rician distribution [1]. Manjon et al. [13] proposed an unbiased NLM (UNLM) filter to remove the biased deviation and to reduce the Rician noise in MR images. Naegel et al [22] extended the NLM filter to enhance to SNR of real-time cardiac MRI. All these filters improve the SNR of the MR images to a certain extent. However, when the noise becomes significant in the image, the performance of these filters is low. There are various approaches available applied by the researchers to denoise the images, but no single scheme is sufficient to fully denoise an image.

### III. NON-LOCAL MEANS NOISE FILTERING

An image  $f$  signal is corrupted with zero mean Gaussian noise  $n$  that can be modeled as:

$$\hat{f} = f_i + n_i \quad (1)$$

where  $\hat{f}$  is noisy version of the original signal  $f$  and  $n$  is Additive White Gaussian Noise with zero mean. NLM filter reconstruct the pixel values of the degraded image  $\hat{f}$  on its each voxel  $x_i$  calculating weighted average of all the voxels intensities of image  $\hat{f}$  as:

$$NLM(g(x_i)) = \sum_{x_j \in \Omega^3} w(x_i, x_j) \hat{f}(x_j) \quad (2)$$

where  $\hat{f}(x_j)$  is the voxel pixel intensity values of  $x_j$  and  $w(x_i, x_j)$  is the weight assigned to  $g(x_i)$  in the reconstructed voxel  $x_i$ . Image pixel intensities with similar values are used to decide the denoised intensity of the pixel. Based on the similar intensity of the pixel values, weights are assigned to pixels with the pixels being denoised along-with its neighboring pixels in the patch. The weight quantities quantify the similarity of the local neighborhood of the vector  $N_i$  and  $N_j$  of the voxel  $x_i$  and  $x_j$  with the assumptions that  $w(x_i, x_j) \in [0, 1]$  and sum of all neighborhood pixels under the kernel should be =1 i.e.  $\sum_{x_j \in \Omega^3} w(x_i, x_j) = 1$ . The size of the search volume  $V_i$  is  $(2M + 1)^3$  centered around the 3D voxel pixel  $x_i$ .

For each voxel  $x_i$  is calculated depending upon its  $L2$  norm distance  $\| \cdot \|_{2,\sigma}^2$  that is computed between vector  $N_i$  and  $N_j$ . This norm  $L2$  is convolved with a Gaussian kernel with its standard deviation  $\sigma$  and this distance, weight  $w(x_i, x_j)$  is calculated as:

$$w(x_i, x_j) = \frac{1}{Z_i} e^{-\frac{\|f(N_i) - f(N_j)\|_{2,\sigma}^2}{h^2}} \quad (3)$$

Here  $Z_i$  is a normalized constant that ensures that the total weight  $w(x_i, x_j)$  should limit between 0 and 1 and  $h$  works as a smoothing parameter and controls the decay of the exponential function. The value of the parameter  $h$  influence the smoothing of the denoised voxel  $x_i$ . If the value of  $h$  is very high, then all voxels  $x_j$  in the set of  $V_i$  leads to strong smoothing of the image. Setting low value of the smoothing parameter  $h$  leads to strong decay of the exponential function that results into low voxels  $x_j$  in  $V_i$  with  $f(N_j)$  very close to  $f(N_i)$  will have a significant weight.

The same NLM filter is applied to each block of the image, dividing the image volume  $\Omega^3$  into overlapping blocks  $B_{i_k}$  of size  $(2\alpha+1)^3$ . Every block  $B_{i_k}$  is centered on voxel  $x_{i_k}$  that constitutes a subset of the volume  $\Omega^3$ . The voxels  $x_{i_k}$  are equally placed at positions  $i_k = (k_{1n}, k_{2n}, k_{3n}), (k_1, k_2, k_3) \in \mathbb{D}^3$ . Here  $n$  represents the distance between the centers of  $B_{i_k}$ . To make the equal reconstruction continuity in the reconstructed image, the overlapping block support has to be non-empty. For each block  $B_{i_k}$ , non-local means filter is performed as follows:

$$NLM(f(B_{i_k})) = \sum_{B_j \in V_{i_k}} w(B_{i_k}, B_{i_j}) f(B_j) \quad (4)$$

where weight  $w(B_{i_k}, B_{i_j})$  is defined as:

$$w(B_{i_k}, B_{i_j}) = \frac{1}{Z_{i_k}} e^{-\frac{\|f(B_{i_k}) - f(B_{i_j})\|_2^2}{2\beta\sigma^2 |N_i|}} \quad (5)$$

Here,  $Z_{i_k}$  is a normalizing constant that ensures the sum of all blocks  $B_{j_k}$  in a volume  $V_{i_k}$  after convolving with the weights is = 1. For a voxel  $x_i$  that is included in several blocks  $B_{i_k}$ , number of estimates of the new intensity values are obtained for different  $NLM(f(B_{i_k}))$ . These estimations are taken from different  $NLM(f(B_{i_k}))$  for a voxel  $x_i$  are stored in a vector  $A_i$ . The final outcome of the reconstructed voxel  $x_i$  is then defined as:

$$NLM(\hat{f}(x_i)) = \frac{1}{|A_i|} \sum_{y \in A_i} A_i(y) \quad (6)$$

#### A. Application of NLM filter to 3D MRI Noise Removal:

The magnitude of the MR image is computed from real and imaginary parts of signals which contains the noise having Gaussian distribution. When the process is executed to acquire the subject image, the MR scanners provides a set of discrete Fourier data samples. These data samples are assumed to be contaminated with random Gaussian white noise during acquisition of subject data. When the image of subject is reconstructed through the inverse discrete Fourier transform, the magnitude MR image data is converted into Rician distribution. Rician noise is having a signal dependent mean and noise variance that is very difficult to separate from the useful data. In high contrast regions of the MR images, noise follows Gaussian distribution whereas in the low contrast regions of the magnitude MR image noise follows a Rician distribution. The magnitude image is constructed from real and imaginary components of the Fourier transform driven *k*-space data, pixel-by-pixel. Because this operation is nonlinear transform, the noise distribution becomes Rician from Gaussian. Following figure show these noise distribution probability functions

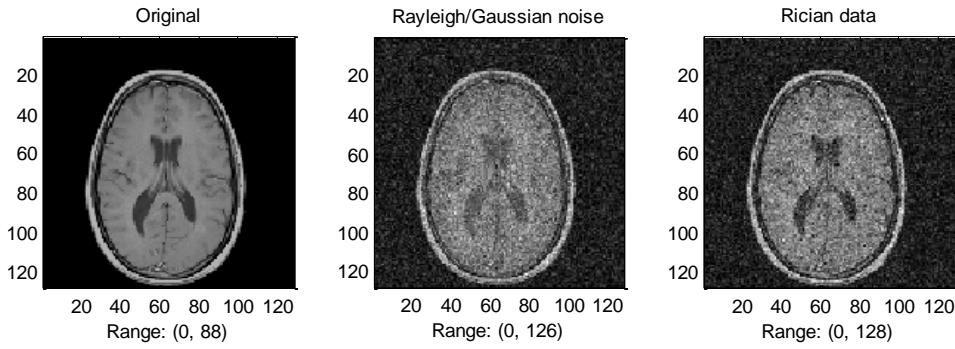


Fig: 1. From left to right (a) original input MR image (b) random Gaussian corrupted Rayleigh distribution (c) Rician noise distribution

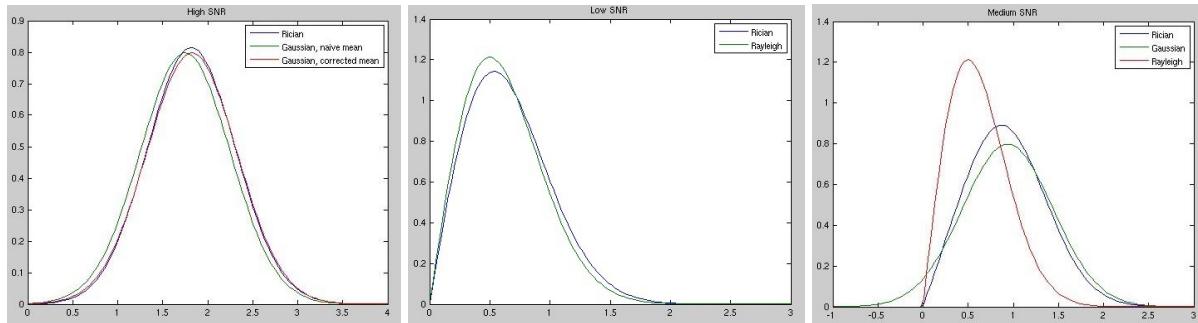


Fig: 2 (a) Normal distribution (b) Rayleigh distribution (c) Rician distribution Probability density functions (PDF) of Gaussian and Rician noise distributions in MR images.

The squared magnitude MR image (the value of each pixel in the image is square of the value of the corresponding pixel in the original magnitude image) has a noise bias which is calculated equal to  $2\sigma^2$  and then it becomes signal independent. Now it can be easily removed from the image data. Therefore, for a given MR image magnitude image  $I_n$  :

$$E(I_n^2) = I_o^2 + 2\sigma^2 \quad (7)$$

where  $I_o$  is the ground truth of image  $I_n$  and  $I_n^2$  represents the square magnitude image of  $I_n$  and  $I_o$  respectively.

Noise bias removal: To restore the true values of the pixel intensities, it is very important to remove the bias effects. As proposed by Nowak [1], the estimation of bias standard deviation is calculated by:

$$\sigma = \sqrt{\frac{\mu}{2}} \quad (8)$$

where  $\mu$  is mean value of the background regions of the squared magnitude MR image. The background regions are selected using Otsu [17] thresholding methods.

To avoid the bias effects in the reconstruction process, the estimated bias field is subtracted from the reconstructed voxel  $x_i$  pixel values to make it unbiased noise free as:

$$x_i = \sqrt{\max \left( \left( \frac{\sum_{x_j \in \Omega_i^3} w(x_i, x_j) f(x_i)^2}{\sum_{x_j \in \Omega_i^3} w(x_i, x_j)} \right) - 2\sigma^2, 0 \right)} \quad (9)$$

#### IV. EXPERIMENTS

The experimental purposes, the well-known 3D Brainweb dataset has been used containing T-1(w) MRI dataset of size  $181 \times 217 \times 181$  voxels. The input data was corrupted with the Gaussian and Rician noised in the range of 1-9% of the maximum intensity. Rician noise is added to the input voxels adding Gaussian Noise to the real part of the signal and imaginary parts and then the magnitude of the image voxels is obtained as defined in [1]. The algorithm is tested in Matlab 2012b on Dell computer having Intel Quad-Core processor machine having 4 GB RAM and Windows 10 platform. Wavelet subband Mixing, Block-wise filtering and non-local mex functions are used in this work to perform experiments, freely provided to reuse for reproduction of noise filtering techniques by [13].

##### A. Quality Measure

The performance evaluation of the given method depends on the mapping and similarity between the noise reduction and image's structural detail preservation. Each image quality measure method tends to retain the first or second perspective. The Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) both are used to measure the perceptual and structural quality of the reconstructed image from all the filtered voxels  $x_i$  by:

$$PSNR = 20 \log_{10} \left( \frac{MAX}{RMSE} \right) \quad (10)$$

Root Mean Squared Error (RMSE) is computed just squaring the Mean Square Error that is defined as

$$MSE = \frac{\sum (f(x_i) - \hat{f}(x_i))^2}{\sum f(x_i)} \quad (11)$$

and finally RMSE is computed as:

$$RMSE = \sqrt{MSE} \quad (12)$$

SSIM is the method used to assess the maximum reconstruction similarity between the original and reconstructed image and retention of the important details of the image while removing noise components. The SSIM error parameters computes the similarity in a local window by combining difference averages, variation and correlation. The value of SSIM between two images windows  $W_1$  and  $W_2$  from the original and reconstructed image is defines as in [16]:

$$SSIM(W_1, W_2) = \frac{(2\mu_{W_1}\mu_{W_2} + I_1)(2\sigma_{W_1W_2} + I_2)}{(\mu_{W_1}^2 + \mu_{W_2}^2 + I_1)(\sigma_{W_1}^2 + \sigma_{W_2}^2 + I_2)} \quad (13)$$

TABLE 1

Noise Level (%age)	Block-wise	Wavelet Sub-Mixing	Adaptive Rician NLM
1	42.91	42.82	44.26
3	37.94	37.96	38.33
5	34.92	34.98	35.37
7	32.73	32.79	33.33
9	31.06	31.07	31.82

Comparison of experimental results in PSNR.

TABLE 2

Noise Level (%age)	Block-wise	Wavelet Sub-Mixing	Adaptive Rician NLM
1	0.9928	0.9926	0.99
3	0.9773	0.9775	0.98
5	0.9552	0.9561	0.96
7	0.9258	0.9275	0.93
9	0.8921	0.8938	0.90

Comparison of denoised image reconstruction structural similarity index (SSIM) after noise removal

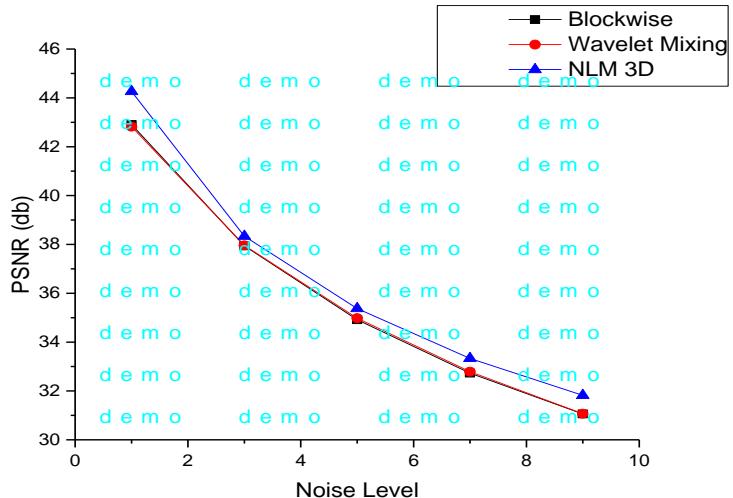


Fig: 2. Comparison of denoised image quality PSNR(db) with the existing methods

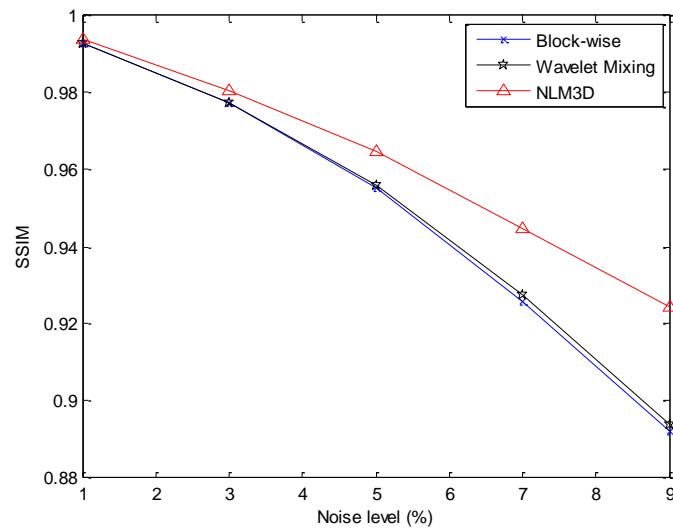


Fig: 3. Comparison of the 3D ARNLM results with the existing methods result with noise level 1-9% of the total noise intensity



Ground Truth

Noise Levels (%)	Rician Noise contaminated MR Images	Reconstructed MR Images after denoising (PSNR in db)	Noise Residues
1%	a	b (44.33)	c
3%	a	b (38.10)	c

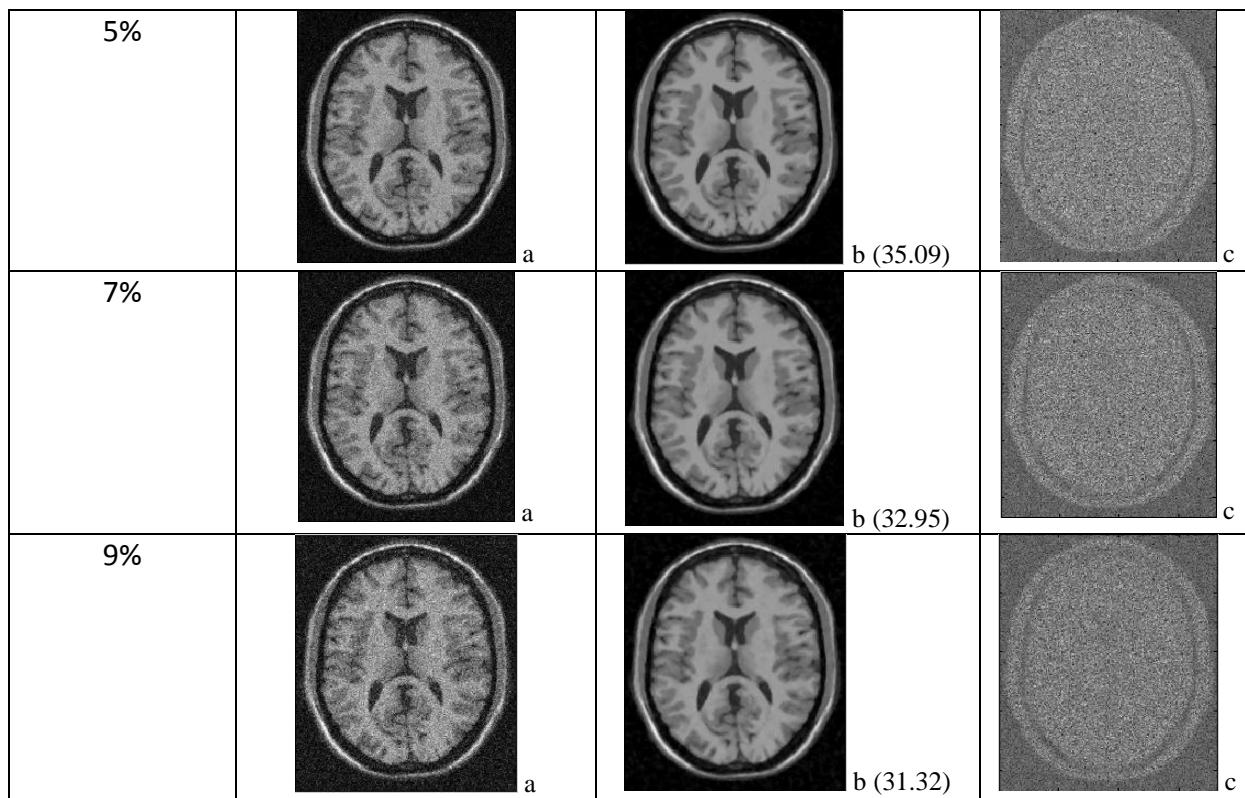


Fig: 4. From top (left to right) noise level,(a) noisy image, (b) denoised image with PSNR (db) value(s) and (c) noise residual after applying Adaptive NLM on 3D MR ground truth image voxels

## V. CONCLUSION

In this paper, experimental implementation of the non-local mean noise removal method in 3D MR image is presented. This method works on manipulation of self-similar type of intensity regions in the image and computing weighted average of the neighboring pixels to denoise the central pixel in that image. The procedure is applied to the entire image decomposing it into equal size voxels. NLM filter is adapted for 3D MRI denoising incorporating the methods given by other researchers to enhance its performance. This filter is simple to implement and produces state-of-the art denoising results. It can be combined with other denoised methods to reproduce the hybrid version of this filter to remove noise in 3D MRI more effectively and with speed. Due to its computational complexity, its computation cost is very high that requires more attention to improve its computational speed.

## REFERENCES

- [1] Nowak RD., “Wavelet-based Rician noise removal for magnetic resonance imaging”, IEEE Transactions on Image Processing, 8 (10):1408–1419, 1999.
- [2] Perona P, Malik J., “Scale-space and edge detection using anisotropic diffusion”, IEEE Transactions on Pattern Anal Mach Intell;12(7):629–639, 1990.
- [3] Carlo Tomasi and Roberto Manduchi, “Bilateral filtering for gray and color images,” in Computer Vision, Sixth International Conference on . IEEE, 839– 846, 1998.
- [4] Ashburner, J., Friston K.J., “Voxel based morphometry – the methods”, Neuro-Image”, 11, 805-821, 2000.
- [5] Gerig, G., Kubler, O., Kinkis, R., Joslesz, F.A., “Nonlinear anisotropic filtering of MRI data”, IEEE Transactions on Medical Imaging”, 11, 221-232, 1992.
- [6] Krissian K., Aja-Fernandez, S., “Noise-driven anisotropic diffusion filtering of MRI”, IEEE Transactions on Image Processing”, 18, 2265-2294, 2009
- [7] Pizzurica, A., Philips, W., Lamahiers L., Achery, M., “A versatile wavelet domain noise filtering technique for medical imaging”, IEEE Transactions on Medical Imaging”, 22, 323-331, 2003.
- [8] Donoho, D.L., “Denoising by Soft-Thresholding”, IEEE Transactions on Image Transformations Theory”, 41, 613-627, 1995

- [9] Donoho, D.L., and Johnstone, I.M., “*Thresholding selection for wavelet shrinkage of noisy data*” 16<sup>th</sup> Intl. Conf. of IEEE Engg. In Medicine and Biology Society”, 1, A24-A25, 1994.
- [10] Hari Om and Mantos Biswas, “*An Improved Image denoising methods based on wavelet thresholding*”, Journal of Signal & Information Processing, 3, 109-116, 2012.
- [11] Zhang, X.P., and Desai, M.D., “*Adaptive Denoising on SURE Risk*”, IEEE Signal Processing Letters, 5, 265-267, 1998.
- [12] Buades A., Coll B., Morel J.M., “*A non-local means algorithm for image denoising*”, IEEE International Conference on Computer Vision and Pattern Recognition, 2, 60-65, 2005.
- [13] Manjon JV, Carbonell-Caballero J, Lull JJ, Garcia-Marti G, Marti-Bonmati L, Robles M. “*MRI denoising using non-local means*”, Medical Image Analysis, 12:514–523, 2008.
- [14] Coupe, P., Manjon Jose V., G. Alieas, Arnold D., Robles, M., and Collins D.L., “*Robust Rician Noise Estimation for MR Images*”, Medical Image Analysis, 14(4), 483-493, 2010.
- [15] Manjón,J.V., Coupé, P., Martí-Bonmatí,L., Robles,M., Collins, D. L., “*Adaptive Non-Local Means Denoising of MR Images with Spatially Varying Noise Levels*”. Journal of Magnetic Resonance Imaging, 31,192-203, 2010.
- [16] Coupé, P., Manjón,J.V., Robles, M., Collins, D.L., “*Adaptive Multiresolution Non-Local Means Filter for 3D MR Image Denoising*”, IET Image processing, 6(5): 558-568, 2012.
- [17] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., “*Image Quality Assessment from error visibility to structural similarity*”, IEEE Transactions on Image Processing, 13, 600–612, 2004.
- [18] Otsu, N.A., “*Threshold selection method from gray-level histograms*”, IEEE Transactions on system, man and cybernetics,9, 62-66, 1979.
- [19] Guleryuz, O.G., “*Weighted Averaging for denoising with overcomplete dictionaries*”, IEEE Transactions on Image Processing, 16 (12), 3020-3034, 2007.

# Pattern and Color Descriptors for Retrieval of Images.

Shilpa, *Student*, Er. Navdeep Singh, *Assistant Professor*, Dept. of Computer Engineering, Punjabi University Patiala.

**Abstract**— For the retrieval of images, CBIR (Content Based Image Retrieval) technique is used for extraction of features from the images. Various approaches exist for the feature extraction from the images in the database, but they lack in accuracy. To solve this - a hybrid technique is proposed in this paper for efficient feature extraction of images. Local Binary texture patterns such as LBP (Local Binary Pattern) and CS-LBP(Center-Symmetric Local Binary Pattern) are used with color descriptors such as Color Histogram and Color correlogram to form a hybrid approach to make retrieval more accurate and efficient. Euclidean Distance classifier is used for the similarity matching between the query image and the images stored in the database. The proposed technique has an accuracy of 0.86 with precision and recall factors of 0.92 and 0.93 respectively which is fairly large then the techniques used so far in isolation.

**Keywords**— Content Based Image Retrieval, Color descriptors, Euclidean Distance for similarity measure and LBP patterns.

## I. INTRODUCTION

In a large database, it is very difficult to search for a particular image. To overcome this limitation CBIR approach came into existence to make retrieval of images more accurate and efficient. Fig.1 represents the working of CBIR process. CBIR approach deals with the features of images such as color, texture and shape rather than the image itself.

Color descriptors like Color moments, Color Histogram, CCV (Color Coherence Vector) and Color Correlogram are used for extracting features from the images. The uniqueness of color distribution in an image is represented by Color Histogram. Color correlogram focuses on both pixel distribution and spatial correlation of pair of pixels. To make feature extraction more accurate color moment and color coherence vector are used.

For texture feature extraction besides coarseness, regularity and directionality, LBP (Local Binary pattern) has emerged as a powerful tool for classification and retrieval procedures. LBP originally expressed by Ojala et al. [11] has gained popularity because of its property of illuminance. Various extensions of LBP came into existence to reduce the histogram bins for retrieval. CS-LBP advised by Marko et al. [12] extracted the features by reducing number of histogram bins from 256 to 16 bins which is half of what is obtained by LBP which is easy to use in region descriptor. Color histograms work on RGB (Red Green Blue) and HSV (Hue Saturation value) color spaces .Color correlogram provides information of pixel distribution and spatial correlation of pixels, by focusing on how spatial correlation changes with respect to distance in an image.

Another color descriptor called color moment is used for feature extraction from images in the form of mean, variance and skewness of the images.

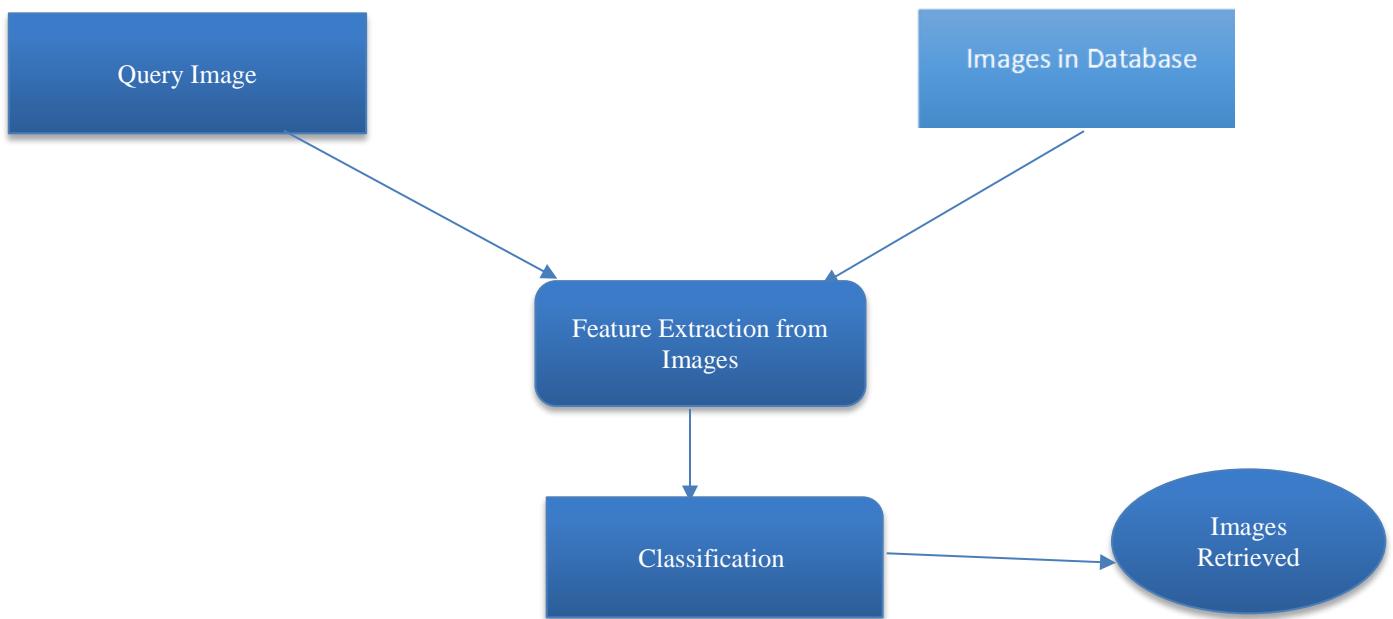


Figure 1. CBIR Process

The research paper is organized as: Section II describes the related Survey, Color and LBP patterns are discussed in Section III, Section IV describes the proposed work and Section V describes the results followed by conclusion in Section VI.

## II. RELATED SURVEY

Pengyu Lin et al. [1] used weight assignment operators for retrieval system .On the basis of image content, weight was assigned to texture and color features. This method provides more accuracy as compared to traditional Content Based Image Retrieval process because in this retrieval, results were efficient and accurate as compared to other methods which were based on single feature vector

Tri Huynh et al. [2] introduced a novel descriptor called Gradient-LBP for encoding information of facial depth that was inspired by 3DLBP.The method was used for face recognition. For evaluation Kinect and Range seamer databases are used.

Greg Pass et al. [3] introduced a technique called Histogram Refinement which split the pixels into several classes based on local property. Color Coherence Vector was used as split histogram that partitioned the histogram of images whose color histograms are same.

Lidiya Xavier et al. [4] used Wavelet Transform for the extraction of texture features from images. This method was tested on DRD (Diabetic Retinopathy Database).Precision and Recall measures obtained from this method gave more accuracy as compared to other retrieval techniques.

S.Mangijiao Singh et al. [5] proposed an integrated approach of color moment and color histogram for feature extraction from images. To overcome the limitation of color histogram lacking efficiency in spatial information, color moment was combined for efficient retrieval. HSV (Hue Saturation value) quantization and 256 bins were used to represent the image.

Hari Hara et al. [6] introduced an approach for Content Based image Retrieval. Color moment primitives were used where image was partitioned into four segments and moment were extracted from all segments and clustered into four classes. Variance moment considered as primitive gave more accuracy than existing method. Histogram intersection and Euclidean distance was used for similarity measurement.

Ahmed Talib et al. [7] presented two compact representations of color correlogram for color feature extraction. Colors from images were compressed and correlogram descriptor's distance was generalised and dominant color-based correlogram was used for representation. These two compact representations were combined to achieve higher accuracy and feasibility.

T. Raja Lakshmi et al. [8] gave CBMIR (Content Based Medical Image Retrieval) System for medical purposes. Haralick features, Run-Length features, Histogram Intensity length features and Zernike moments were used for CBMIR. To reduce the dimensionality problem and to improve the performance of CBMIR, a hybrid approach of CBMIR with 'branch and bound algorithm' and 'artificial bee colony' was used to improve accuracy parameters.

Roshi Chaudhary et al. [9] proposed that visual content of the image, extract efficient information as compared to text-based approach. Hybrid approach of color moment for color feature extraction and LBP (Local Binary Pattern) for texture feature extraction was used for retrieval process from a large database.

### III. BINARY PATTERNS AND COLOR DESCRIPTORS

#### 1. Local Binary Pattern

The LBP operator (local binary operator) given by Ojala et al. [11] is a pattern that calculate the value of pixel by comparing with its (3\*3) neighbours.

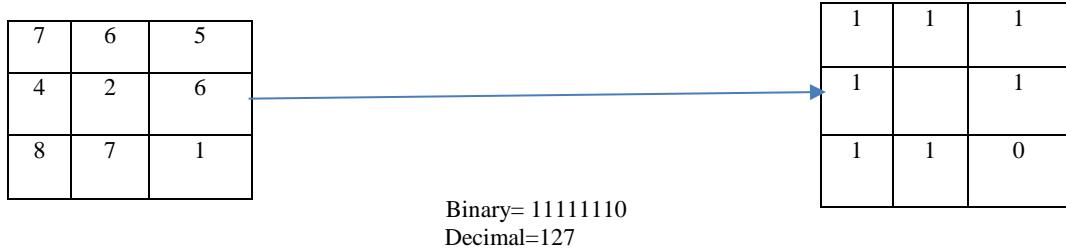


Figure 2. LBP operator

In this binary pattern, gray level value of each pixel is compared with its neighbours and result is obtained by calculating the difference between the value of neighbouring pixels and center pixel. If the value of neighbouring pixel is greater than center pixel value, then it is encoded as "1" otherwise "0".The binary pattern is obtained by combining all these values in anticlockwise direction and then the pattern is further converted into decimal form which is used for labelling the pixel which is shown in Fig.2. The mathematical representation of LBP is as follows:

$$LBP_{M, N} (x, y) = \sum_{j=0}^{N-1} s(g_j - g_c)2^j,$$

$$W(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The notation  $(M, N)$  represents sample points  $M$  over radius  $R$  on pixels  $(x, y)$ .

Where  $g_c$  represents gray level pixel value of center pixel and  $g_j$  represents the gray value of  $N$  equally spaced pixels of radius  $R$  [2] and  $W(x)$  represents the threshold function.

## 2. Center-Symmetric LBP

To overcome the limitation of long histograms in region descriptors, Center-Symmetric LBP came into existence [6]. In this approach, each pixel is compared with the center pixel and value is calculated for binary pattern. The pattern obtained from this technique contains 16 ( $2^4$ ) bins which is half the number obtained from simple LBP. The feature vector obtained is used for analysis.

The pattern is represented mathematically as follows:

$$\begin{aligned} \text{CS-LBP}_{R,S,T} &= \sum_{j=0}^{N-1} m[g_j - g_c + (\frac{S}{2})] 2^j \\ W(x) &= \begin{cases} 1 & \text{if } x \geq t \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (2)$$

$g_j$  and  $g_c + (\frac{S}{2})$  corresponds to the gray level pixel values and  $S$  represents total pixels which are equally spaced on radius  $R$ .  $W(x)$  represents the threshold function.

## 3. Color Histogram

Color is an important factor for extraction which was used in image retrieval. Feature extraction in the form of color is as easy as compared to shape and texture features.

Color Histogram is defined as number of pixels that have colors in each of fixed list of ranges of color that span image's color space, the set of possible colors [14]. This approach represents the uniqueness of color distributions in an image which is based on certain color spaces like RGB and HSV. The color space is divided into certain number of small intervals called bins. The mathematical representation of color histogram is given as [2].

$\text{Hist} = \{\text{Hist}[0], \text{Hist}[1], \text{Hist}[2], \text{Hist}[3], \dots, \text{Hist}[i], \dots, \text{Hist}[n]\}$  where histogram color bins are represented by  $i$  and  $\text{Hist}[i]$  describes pixels of color in an image and number of bins are represented by  $n$ .

## 4. Color Correlogram

Color histogram has a disadvantage of spatial correlation. To solve this problem, color correlogram came into existence [7] that describes correlation statistics of an image. The color correlogram is represented by a table indexed by color pairs  $(C_{c_j}, C_{c_k})$  where  $k^{\text{th}}$  entry describes the probability of finding color  $C_{c_j}$  at distance  $s$  from color  $C_{c_k}$  in an image,  $j$  and  $k$  are the indexes of color but still there is problem of memory space and time computation. To solve the problem of space and time complexity, autocorrelogram came into existence in which correlation with same color is kept and other colors are ignored.

## 5. Color Moment

Color moment is a color descriptor used for extraction of features from the images. This factor is calculated by taking mean, standard deviation and skewness of the image.

Mean is defined as the color average in an image. Standard deviation is calculated by taking square root of variance of color distributions in an image and skewness is calculated by taking the cube root of the variance of the image.

$$\text{Mean: } \mu_j = 1/N \sum_{k=1}^P f_{jk} \quad (3)$$

$$\text{Standard Deviation: } \sigma_i = \left( \frac{1}{N} \sum_{k=1}^P (f_{jk} - \mu_j)^2 \right)^{1/2} \quad (4)$$

$$\text{Skewness: } S_i = \left( \frac{1}{N} \sum_{k=1}^P (f_{jk} - \mu_j)^3 \right)^{1/3} \quad (5)$$

Where N represents the total number of pixels in the image and p represents the last image pixel. The color value of j-th color component of the k-th image pixel is given by  $f_{jk}$  and  $\mu_j$  represents the first color moment for j-th color value of the image.

#### IV. PROPOSED WORK

In this work, a hybrid approach of texture feature techniques (LBP and CS-LBP texture patterns) and of color feature extraction techniques (color histogram, color correlogram and color moment) is proposed. The proposed approach is more efficient and accurate than the techniques used in isolation.

#### Algorithm

For the query image ( $P_1$ ):

1. Pre-process the input image (query image) and database images by converting RGB to gray level.
2. Apply LBP and CS-LBP to obtain texture feature vector M1.
3. Apply color moment, color histogram and color autocorrelogram to obtain feature M2.
4. Combine M1 and M2 to obtain new feature vector M\_N1.
5. Use Euclidean Distance to calculate similarity measure between query image ( $P_1$ ) and images in database.
6. Images containing similar features as that of query images are retrieved from the dataset.

#### V. EXPERIMENTAL RESULTS

To evaluate the texture and color features performance Wang database [10] is used. This database contains 1000 images which are further divided into ten different classes each containing 100 images.



Figure 3. Query Image.



Figure 4.Retrieved images using Color Features with its confusion matrix.

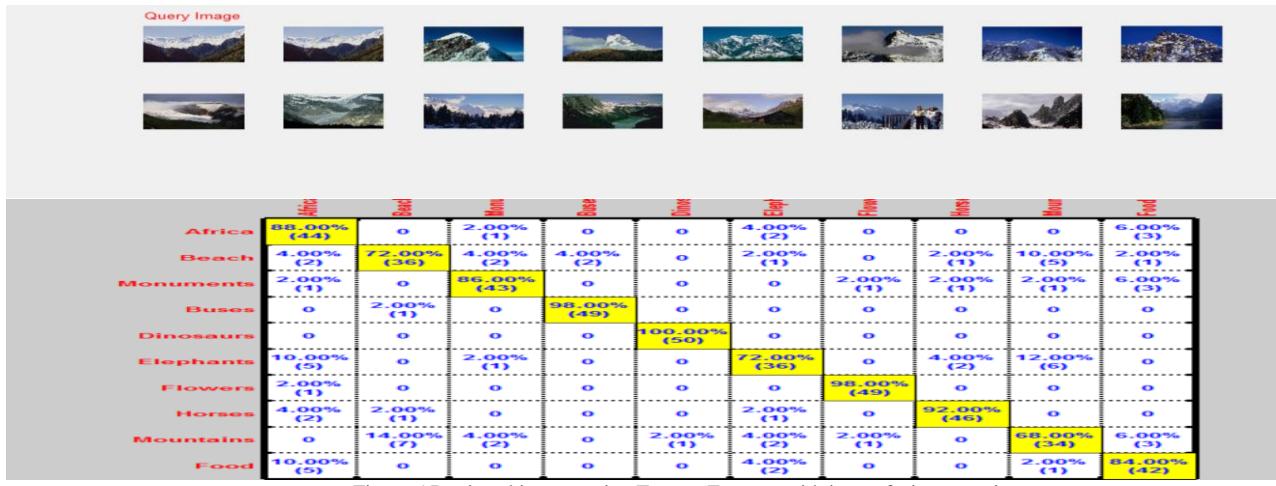


Figure 5.Retrieved images using Texture Features with its confusion matrix.



Fig 6.Retrieved images using Proposed Approach with its confusion matrix.

TABLE I shows accuracy of different techniques. Fig.7 shows the graphical representation of accuracy values. The Accuracy is calculated as follows:

$$\text{Accuracy} = \frac{\text{Number of correct Assumptions}}{\text{Total number of classes}}$$

TABLE I: Accuracy of Different Techniques

Classes	Color Features	LBP Texture Features	Proposed Technique
Africa	0.86	0.88	0.90
Beach	0.60	0.72	0.70
Monuments	0.74	0.86	0.88
Buses	0.88	0.98	0.96
Dinosaurs	0.98	0.95	0.90
Elephants	0.84	0.72	0.84
Flowers	1.00	0.98	0.98
Horses	0.90	0.92	0.94
Mountains	0.64	0.68	0.70
Food	0.92	0.84	0.86
Overall Accuracy	0.83	0.85	0.86

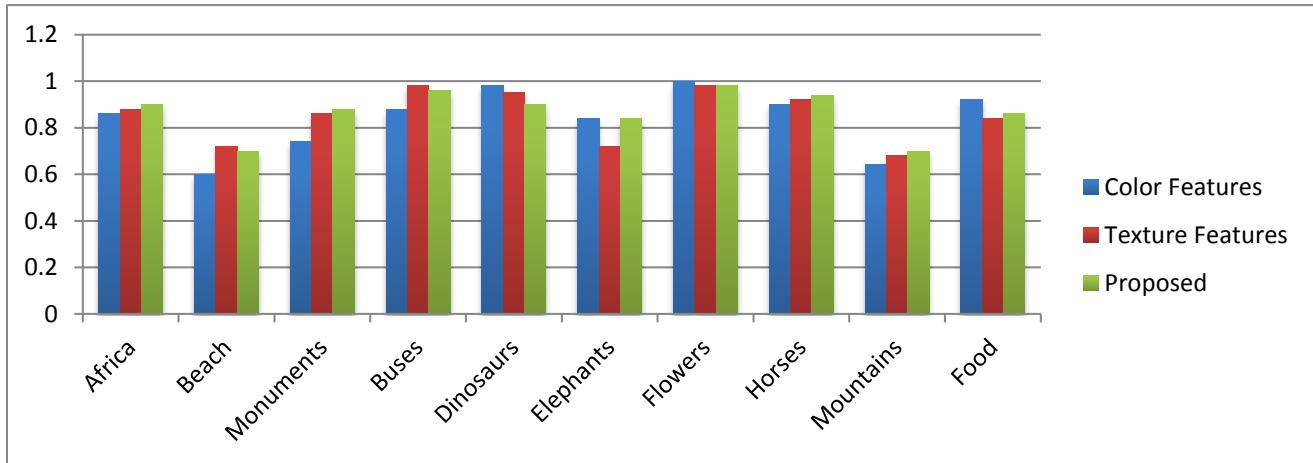


Figure 7. Graph showing color features, texture features and proposed features accuracy.

TABLE II shows Precision measure of different techniques and Fig. 8 represents the precision values graphically. The precision value is calculated as follows:

$$\text{Precision} = \frac{\text{Number of correct images retrieved}}{\text{Total number of images retrieved}}$$

TABLE II: Precision of different techniques

Classes	Color Features	LBP Texture Features	Proposed Technique
Africa	0.74	0.73	0.81
Beach	0.76	0.80	0.76
Monuments	0.69	0.93	0.81
Buses	0.97	1.00	1.00
Dinosaurs	0.95	0.86	0.97
Elephants	0.89	0.87	0.93
Flowers	0.93	0.98	1.00
Horses	0.94	1.00	1.00
Mountains	0.96	0.94	0.92
Food	1.00	1.00	1.00
Overall Precision	0.83	0.91	0.92

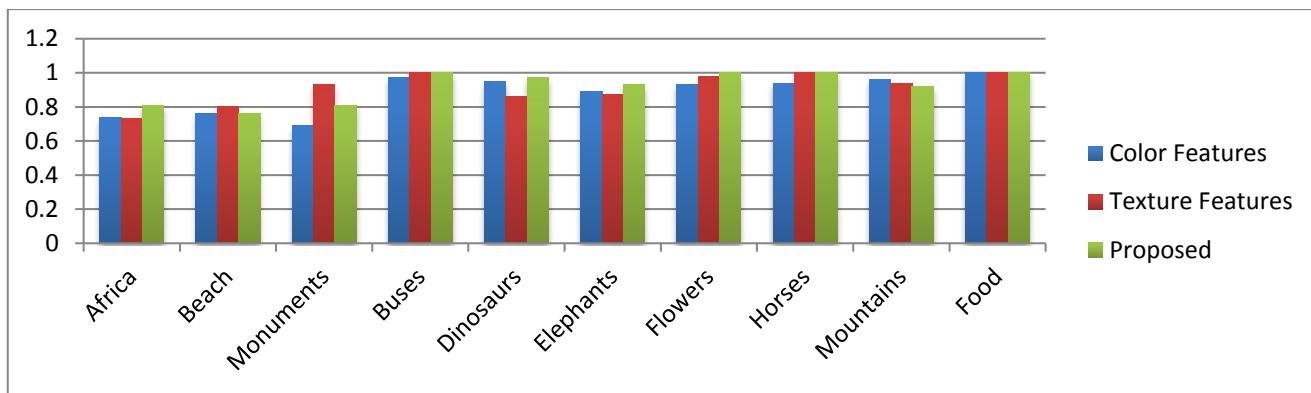


Figure 8. Graph showing color features, texture features and proposed features precision.

TABLE III shows Recall measure of different techniques and Fig.9 shows the graphical representation of recall values. The recall is calculated as follows:

$$\text{Recall} = \frac{\text{Number of correct images retrieved}}{\text{Number of correct images in the Database}}$$

TABLE III: Recall of different techniques

Classes	Color Features	LBP Texture Features	Proposed Technique
Africa	0.86	0.88	0.90
Beach	0.63	0.75	0.72
Monuments	0.88	0.87	0.93
Buses	0.97	1.00	1.00
Dinosaurs	1.00	1.00	0.93
Elephants	0.95	0.81	0.93
Flowers	1.00	1.00	0.93
Horses	0.95	1.00	0.97
Mountains	0.94	0.91	1.00
Food	1.00	1.00	1.00
Overall Recall	0.91	0.92	0.93

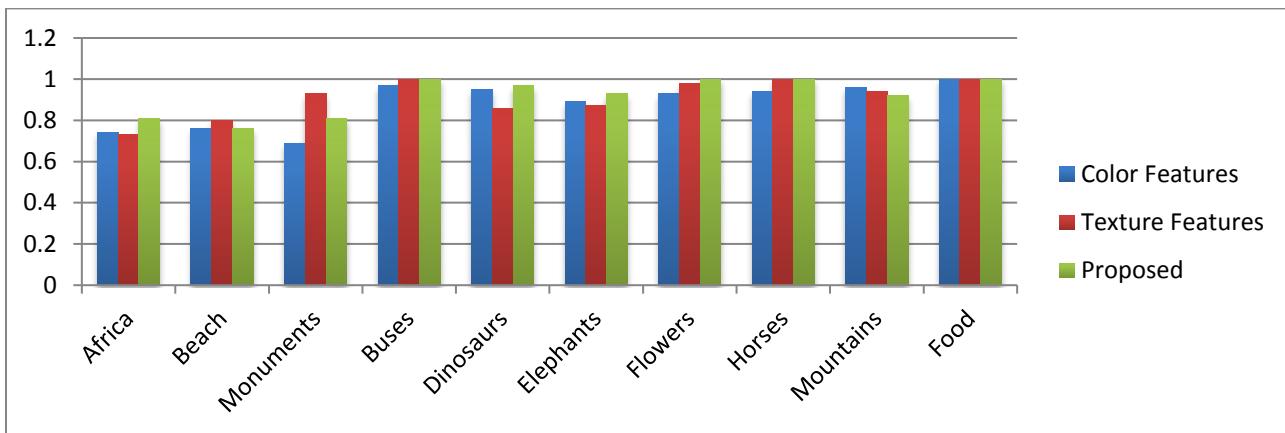


Figure 9. Graph showing color features, texture features and proposed features recall.

## VI. CONCLUSION

In this paper, we have studied various LBP texture patterns and color feature extraction techniques for retrieval of images from huge database, but they lack in efficiency and accuracy. A hybrid technique of texture and color descriptors is proposed to overcome the limitations of these techniques when used in isolation. This technique is more accurate and efficient in terms of precision and recall measures as compared to the existing techniques as far as we know.

## REFERENCES

- [1] Pengiu Liu , Kebin Jia , Zhuozheng Wang and Zhou Lv , “A New and Effective Image Retrieval Method Based on Combined Features.”, IEEE, 2007.
- [2] Tri Huynh,Rui Min, Jean-Luc-Dugelay, ”An Efficient LBP-based Descriptor for Facial Depth Images applied to Gender Recognition using RGB-D, Face Data”, 2008.
- [3] Greg Pass, Ramin Zabih "Histogram Refinement for Content –Based Image Retrieval", 2010.
- [4] Lidiya Xavier and I.Thusnavis Bella Mary ,”Evaluation of Retrieval System Using Textural Features Basedv on Wavelet Transform”,CCIS pp-481-484), 2011.
- [5] S.Mangijiao Singh, "Content Based Image Retrieval using Color Moment and Gabor Texture Features". IJCSI. (vol.9), 2012.
- [6] Hari Hara Pavan Kumar Bhuravarjula, V.N.S.Vinay Kumar,” A Novel Image Retrieval Method Using Color Moment”, IJECSE, 2013.
- [7] Ahmed Talib, Massudi Mahmuddin,Husniza Husni,”Efficient , Compact and Dominant Color Correlogram Descriptors For Content Based Image Retrieval”.Fifth International Conference on Advances in Multimedia, 2013.
- [8] T.Rajalakshmi and R.I.Minu,”Improving Relevance Feedback for Content Based Medical Image Retrieval), IEEE,2014.
- [9] Roshi Chaudhary, Nikita Raina, Neshu Chaudhary and Rashmi Chauhan, “An Integrated Approach to Content Based Image Retrieval.” IEEE, 2014.
- [10] O.mich, R.a, “Histograms Analysis for image Retrieval” Pattern Recognition, 2014.
- [11] T.ojala ,” Multi resolution Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns”, Pattern Recognition.
- [12] Marko Heikkila, M. P, "Description of interset region with center-symmetric local binary pattern operator. Elsevier ".
- [13] <https://www.google.co.in/search?q=lbp+histograms+images&espv=2GZQKHZMTB9A> accessed on 26-05-2016.
- [14] <https://www.google.co.in/webhp?sourceid=chromeinstant&ion=1&espv=2&ie=UTF-8#q=coor+histogram> accessed on 25-07-2016.

# A Survey of Various Malware Detection Techniques

Er. Kuldeep Singh

Department of Computer Engineering  
Patiala, India  
E-mail [sidhu.kuldeep89@gmail.com](mailto:sidhu.kuldeep89@gmail.com)

Dr. Lakhwinder Kaur

Department of Computer Engineering  
Patiala, India  
E-mail [mahal2k8@yahoo.com](mailto:mahal2k8@yahoo.com)

**Abstract**—In this paper number of techniques have been reviewed that are used for malware detection. Any harmful computer program which performs the undesirable action without the consent of user is called malware. It may be virus, worm, spyware, backdoor, Trojan horse etc. Malware plays the main threat in the computer security and increasing day by day. With the evolutions malware becoming more strong i.e. changing its signature and behavior dynamically so it is difficult to detect them. Numerous techniques have been developed to detect malware. All have some advantages and disadvantages. Signature based detection; Anomaly based detection, Heuristic based and artificial immune system based. Most antivirus used the signature based detection and it used for the signature of known malware which can't detect unknown malware. To overcome this shortcoming behavioral based detection is used, which find the behavior of the program and identified that is it benign or malware. Data mining and machine learning methods are used by heuristic malware detection. Artificial immune system is just like biological immune system that provides the protection to system by differentiating between self and non self particles.

**Keywords**—Malware detection; Artificial immune system; benign; Signature; Heuristic.

## I. INTRODUCTION

Malware is any unwanted program that alters the benign program or slows down the speed of the computer system. Malware [3] also called malicious code that performs the action on system without the consent of the user. Malware includes virus, worm, spyware, Trojan horse and other harmful codes.

**Virus:** Virus is a harmful code that executes by itself and copies their code into any other host file [4]. Another property of virus is that it replicates itself. It mainly used to alter the files, damages the system and deletes the important information.

**Worm:** Worm also replicate and execute by itself. But unlike virus it need not the host file to execute, it can execute and spread by itself. It can replicate through networks also that why mainly used to infect email and word processing file. Mainly worm is used for resource eating purposes.

**Spyware:** Spyware is a computer program that installed on user's computer without the permission of user and also user is unaware of spyware program. Spyware mainly used the capture the activity performed by the user and send it to their intended owner.

**Trojan horse:** Trojan horse is the program that appears to be legitimate program or we can say that it is useful program with certain condition defined. Malicious function is performed only when that condition met, after that it performs their harmful action on computer system.

**Backdoor:** Backdoor is the unauthenticated path, which mainly use by the programmer to provide the fast access for maintenance. But hacker can use this path without the permission of user and damages and steel the important information. Backdoor is mainly used for remote access of computer system.

To prevent from such type of attacks first need to detect the malware. For this purpose malware detector program is required. Malware detector is implementation of one of the method of malware

detection that protects the system from different type of attacks and vulnerabilities present in the computer system.

Figure 1 showing the various malware attacks on the computer system. Worm can be spread in shared computer that affect all the systems. USB also infected with virus and whenever inserted into other system virus can copy itself to other system and infect the system. Vulnerabilities may be occurs in the system that have without security updates and hacker can also used various hacking techniques to infect the secure computer system.

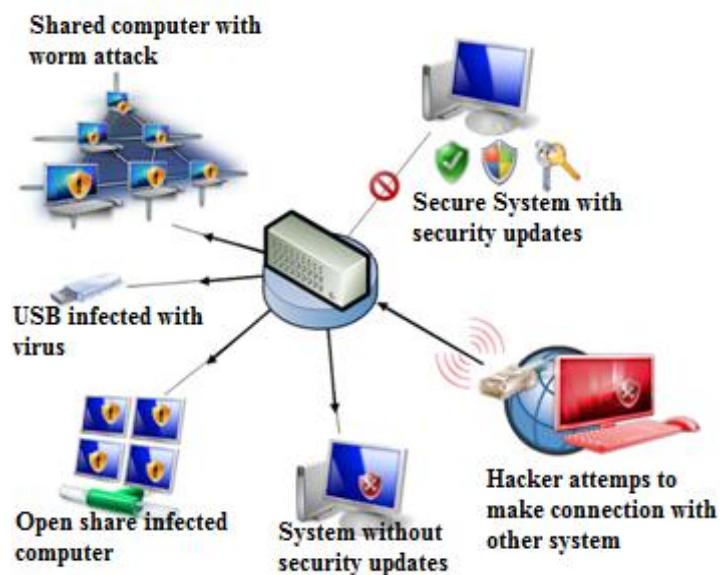


Fig. 1. Various Malware attacks on computer system.

## II. MALWARE DETECTION TECHNIQUES

There are many malware detection techniques available that are reviewed in the literature. But based on some methods and statics used in the techniques are classified into four types: Signature based detection, Anomalies based detection, Heuristic based detection and artificial immune system based detection [1, 2, 3, 4,].

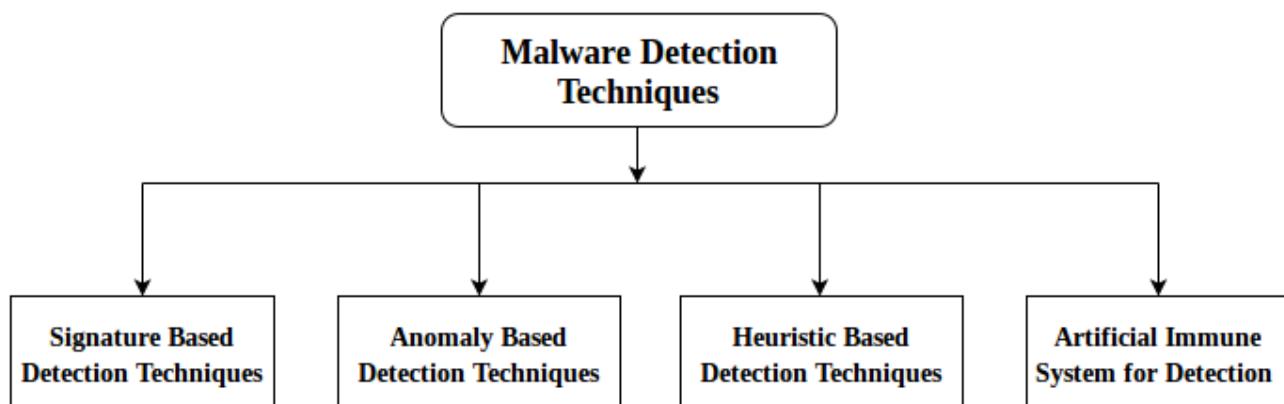


Fig. 2. Classification of various types of malware detection techniques.

All techniques have their own advantages and limitations. Due to the changing environment attacker and hacker are also becoming strong and making the malware that are changing their identity and code dynamically, so that's why it becomes difficult to detect malware completely by single individual approach.

**A. Signature based detection:** In this method the signature of the known malware are added to the repository, by using hash or secure hash algorithm, signature is generated [1]. Signature of detected malware that are saved in repository is used in the detection process. If the signature matches then that program is malware otherwise it is benign program. This technique is efficient only in static environment where the program remain same. It cannot detect metamorphic and polymorphic types of malware.

**Metamorphic malware:** In metamorphic types of malware after each iteration, the code of the program changes, each time it is giving new code thus it becomes very difficult to detect by signature based techniques.

**Polymorphic Malware:** In this technique the main body of the code is encrypted with cryptographic key algorithm. The main body of the code does not change but encryption key is changing and thus the signature of the program also changed every time and it become difficult to detect these types of malwares by signature based malware detection technique.

**TABLE I.** Benefits and limitations of signature based malware detection techniques.

Signature based techniques	
Benefits	Limitations
<ul style="list-style-type: none"> <li>1. Simple to implement and fast technique.</li> <li>2. More accurate for static types of attacks.</li> <li>3. Low false positive rate and high true positive rate for known attacks.</li> </ul>	<ul style="list-style-type: none"> <li>1. Can detect only known malware.</li> <li>2. Can't detect zero day or one day attacks.</li> <li>3. Repository increases with the increase of number of signatures.</li> <li>4. Cannot detect metamorphic and polymorphic type malwares.</li> </ul>

**B. Anomaly based detection:** Anomaly based detection works in two parts. In first part the system trained with the normal behavior of the computer program. Second phase is the testing phase in which in incoming traffic are compared with the trained engine [9]. System compares the behavior if the incoming program showing different or abnormal behavior form trained engine data, then it is identified as malware. It is somewhat difficult to train the every normal behavior so that why when some exception occurs, system identifies it as anomalies.

**TABLE II.** Benefits and limitations of anomaly based malware detection techniques.

Anomaly based techniques	
Benefits	Limitations
<ul style="list-style-type: none"> <li>1. Can detect zero day and one day attacks.</li> <li>2. System is continuously leaning the network activity and updating profiles so less maintenance is required.</li> <li>3. It gives the more accurate result if the system is being used from longer time.</li> </ul>	<ul style="list-style-type: none"> <li>1. Difficult to identify the behavior of complex program.</li> <li>2. Complexity Increases because it is quite difficult to define the rules.</li> <li>3. High false alarm rate.</li> </ul>

**C. Heuristic based detection:** These techniques overcome the limitations of previous discussed techniques [2]. It can detect the unknown attacks or malwares based on the data mining and machine learning techniques. Many methods are used in heuristic based detection techniques, some of them explained below.

**File Emulation:** Also known as sandbox method. In this method programs are tested in virtual environment method. If program behaves like malicious code then it is identified as a malware.

**API/System calls:** Every program uses application program interface to communicate with OS. Based on the sequence of system call the behavior of the program can be identified. Normal programs have short sequences of system call and malware will carry larger sequences. Threshold values and Hamming distance can be used to identify benign and malware programs.

**Control flow graph:** Control flow graph (CFG) is used to represent the flow of the program, from statement to statement what action is performed by the program is illustrated in CFG.

**TABLE III.** Benefits and limitations of Heuristic based malware detection techniques.

Heuristic based techniques	
Benefits	Limitations
1. Can detect unknown attacks. 2. Can detect metamorphic and polymorphic types of malwares. 3. It provides the generic signature detection methods.	1. Take more time for scanning and slow down the system speed. 2. Still false positive rate not removed completely.

**D. Artificial Immune system based detection:** The study of artificial immune system started in mid 1980s by Farmer, Packard, and Perelson [11]. Their study suggested that computer security can be increased by mimicking the behaviour of biological immune system that is self-regulated, highly adaptive, self-organizing nature and have memory to store the past behavior and based on the past behaviour and ability to learn from what is happening currently makes it the best protective system.

Immune system is the defensive system for any biological system. It provides the protective resistance to the foreign challenges that affect [6, 7] the body's normal function. Biological immune system is mainly two types: Innate immune system and adaptive immune system [8]. In innate immune system the immunity is gained from parents and it is non-specific type means it can provide protection against any types of foreign molecules and pathogens. Adaptive immune system is specific type i.e. it operates on foreign challenges only when that specific type of antigen enters in the body.

Artificial immune system (AIS) is inspired from the biological immune system. It is the emerging area in the field of computer security. It can provide automated self-protection from malwares. AIS works based on the self and non-self. The benign program is called self and any other harmful code is known as non-self. Four main algorithms are used till now in AIS like: Negative selection algorithm, artificial immune networks, Clonal selection algorithm and Danger theory [7]. All algorithms are evolutionary algorithms that are inspired from nature.

**TABLE IV.** Benefits and limitations of artificial immune system based malware detection techniques

<b>Artificial Immune System</b>	
<b>Benefits</b>	<b>Limitations</b>
<ol style="list-style-type: none"> <li>1. Can provide automated self immunity to computer system.</li> <li>2. Can detect new unknown attacks.</li> <li>3. Self learning capabilities are present and thus find out the malware and produce the antibodies to cure the problem.</li> </ol>	<ol style="list-style-type: none"> <li>1. Some time auto immunity kills the benign programs.</li> <li>2. Still not implemented at commercial level.</li> <li>3. Somewhat difficult to identification of self and non self.</li> </ol>

### **III. CONCLUSIONS**

Large numbers of malware detection techniques are available, but all of them have their own limitations. None of them alone can provide the complete solution. With the evolution, malwares are also becoming strong and performing huge destruction on computer system. So malware detection is very crucial research topic, in this area lots of research needs to be done, to develop the efficient techniques for malware detection. Artificial intelligence and neural network are the new research area, which can be explored in heuristic malware detection techniques, to can enhance the detection process. The Artificial immune system which is not still developed completely is also a new area which can provide automated self immunity to the computer system.

### **References**

- [1] B. Ismail, D.M. Salvatore, and S. Ravi, "A Survey of Intrusion Detection Systems in Wireless Sensor Networks," IEEE Communication Survey & Tutorial. Vol. 16, No. 1, First Quarter 2014
- [2] B. Zahra, H. Hashem and H. Ali, "A Survey on Heuristic Malware Detection Techniques," 5th Conference on Information and Knowledge Technology (IKT) 2013.
- [3] J. Aycock. "Computer Viruses and Malware," Springer, 2006.
- [4] P. Szor, "The Art of Computer Virus Research and Defense," Addison Wesley for Symantec Press, New Jersey, 2005.
- [5] M.B. Susan, B.V. Rayford, "Fuzzy Data Mining And Genetic Algorithms Applied To Intrusion Detection," In Proceedings of the National Information Systems Security Conference (NISSC), 2000.
- [6] J. Elizebeth, G. Mark, "A Review of Artificial Immune System Based Security Frameworks for MANET," Int. J. Communications, Network and System Sciences, 2016, 9, 1-18
- [7] D. Dasgupta, " An Overview of Artificial Immune Systems and Their Applications," Springer-Verlag, 3-1 1998.
- [8] H.K. Paul, D. Paul, G.H. Gunsch, L.B. Gary, "An Artificial Immune System Architecture for Computer Security Applications," IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 6, NO. 3, JUNE 2002.
- [9] D. Sanjeev, L. Yang, Z. Wei and C. Mahintham, "Semantics-Based Online Malware Detection: Towards Efficient Real-Time Protection Against Malware," IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 11, NO. 2, FEBRUARY 2016.
- [10] S.M. Shraddha, and P.G. Pranit, "Trust-based Voting Method for Efficient Malware Detection," 7th International Conference on Communication, Computing and Virtualization Procedia Computer Science 79 ( 2016 ) 657 – 667.
- [11] J. D. Farmer, N. H. Packard, and A. S. Perelson, "The immune system, adaptation, and machine learning," Physica D: Nonlinear Phenomena, vol. 22, no. 1–3, pp. 187–204, 1986. View at Google Scholar · View at Scopus

# Optimal Selection of Factors Influencing Student Academic Performance in Educational Data Mining

Gurmeet Kaur(Student) and Dr. Williamjeet Singh (Assi. Professor)

Department of CSE, Punjabi University, Patiala

[grmtkaur76@gmail.com](mailto:grmtkaur76@gmail.com) , [wiliamjeet@gmail.com](mailto:wiliamjeet@gmail.com)

**Abstract-** Data mining is widely used in educational field to find the problems arise in this field is called Educational data mining (EDM). The higher institutes aimed to provide quality education to its student and to provide excellent graduates for the country growth. Best way to achieve this level quality education is to analyze the student academic performance. This analysis will help to identify students who are at risk and needs special attention. But student performance depends upon various factors. Number of studies has been carried out on these influencing factors. In this paper, we explore five categories of factors which positively or negatively affect the student performance. These categories are: Family information, Academic information, Personality, Management skill and Miscellaneous. This study will help the students to know their weak points which degrade their performance and also helpful for faculty and institutes management to counsel and motivate the students to improve their abilities who would be found at high risk of failure.

**Index Terms-**Data mining, educational data mining, Academic performance, Family information, Academic information, Personality, Management skill.

## I. INTRODUCTION

In recent years, there has been increasing interest in the use of data mining (DM) to investigate educational field. Educational Data Mining (EDM) is concerned with developing methods and analyzing educational content to enable better understanding of students' performance. It is also important to enhance teaching and learning process. Students are the most important part of any educational institute or university or school. The aim of educational institutes and universities is to produce well educated and intelligent students for society who will be increase the literacy rate of the country. Intelligent students play important role for social and economic growth of any country. So, growth of any country somehow linked with academic performance of students. But the student's academic performance affected by various factors. These factors may social, psychological, economical and personal. The influence of these factors may vary from student to student. The poor performance of institute and high rate of failure affect institute image, students and also to parents of students who have to pay repeated expenses.

Till now, number of studies carried out by researchers to explore the factors like age, gender, communication skills, environmental issues, family background, township and many more which affect the student performance directly or indirectly. The result and finding of these studies are varies from student to student, city to rural area, country to country and culture to culture.

This paper is aimed to identify and explore the factors that may affect the performance of students and to find the relationship between those selected factors and performance of student. This study will help to 1) Find out

weak parameters of students, 2) Helps to provide appropriate help to weak students, 3) To decrease the failure and dropout rate of institute.

The remainder of this paper is organized as follows: Section II) Discusses related previous work, Section III) Discusses contribution of this paper in research world, Section IV) Explain Important factors influencing student's performance, Section V) Show Theoretical framework and Section VI) Discuss conclusion and future work.

## II. LITERATURE SURVEY

Number of study has been done on student's academic performance. Different researchers have presented their different factors that affect performance. Some of them are discussed below and in TABLE1:

Syed Tahir Hijaz, et al. [13] carried out a study to explore factors affecting college student's performance. The aim of their research was that student performance in examination is linked with student's attitude towards attendance, time for self study, family income, mother's age and mother's education. The research is the bases on information and data collected from students of a group of private colleges. Regression statics was used for analysis which shows positive and negative influence of these factors on student performance.

Jedsarid Sangkapan, et al [6] presented a study aimed to find the factors affecting academic performance of students who were put in risk status at Prince of Songkla University. They took sample group of 390 students from Prince of Songkla University with Yamane's sample size. The results conclude that one of the parameters gender was affecting academic achievement, with the statistical significance value .001, while other affective factors such as anxiety, responsibility, and environmental factor such as instruction quality also affected academic performance with statistical significance value .05. Binary logistic regression was applied to analyze factors affecting academic performance of students under certain conditions.

Victor Mlambo [15] proposed a study to investigate and analyze some parameters of academic performance in an introductory biochemistry course. He took a random sample of 66 registered students of AGRI 1013 and collect data on demographics, learning preference, and entry qualifications. The effect of learning preference, age, gender, and entry qualifications on academic performance was determined. Relationships between all these parameters were analyzed with Pearson's chi-square test.

Julio G. Soto, et al.[14] carried out a study on Student's performance in two semesters of Cell Biology course. They analyzed Teaching strategies, behaviors, and pre-course variables for student's performance. Pre-semester and post semester surveys were conducted to know student's perceptions about class difficulty, work load and effort put into the course, and professional goals. Chi-square ( $\chi^2$ ) tests showed that completion of chemistry requirements, passing the laboratory component of Cell Biology, homework, and attendance were related to passing course. Logistic regression revealed that perfect attendance followed by GPA, were the most important factors linked with passing the course.

Herminio Rodriguez, et al.[4] presented a study conducted in public and private universities in Puerto Rico conducted for high failure rate in the first accounting course. This study also analyzed student's performance on internal and external classroom factors that might have affect on their performance. The sample 1721 students were

selected from different 3 universities. A chi square test revealed that public universities have higher number of failure rate as compared to private universities.

Irfan Mushtaq, et.al [5] carried out a study to identify factors influencing college student's performance. The goal of their research was that student performance in examination is associated with communication, learning facilities, proper guidance and family stress. For this research, information and data collected from students of private colleges to create student profile.

TABLE I  
 SUMMARY OF RELATED WORK

<b>Goal</b>	<b>Author</b>	<b>Factors</b>	<b>Technique used</b>
Factors Affecting Student's Academic Performance	Irfan Mushtaq, Shabana Nawaz Khan[5]	Communication skills Learning facilities Proper guidance Family stress	Regression analysis
Factors Affecting Students Academic Achievement into Probation Status at Prince of Songkla University	Jedsarid Sangkapan[6]	Study habit Anxiety Responsibility Environmental factor	Binary logistic regression
An Analysis of some Factors Affecting Student Academic Performance in an Introductory Biochemistry course at the University of the West Indies	Victor Mlambo[15]	Demographics Learning preference Entry qualifications	Chi square test
Factors Influencing Academic Performance of Students Enrolled in a Lower Division Cell Biology course	Julio G. Soto Sulekha Anand[14]	Class difficulty Amount of study Effort put into the course Professional goals Teaching strategies Behavior Attendance	Chi square test Logistic regression
Factors Influencing Student's Academic Performance in the First Accounting Course	Herminio Rodriguez[4]	Internal class room s (class size, schedule, environment, course material etc) External class room (Family, extra- curricular, work etc)	Chi square test
Factors Affecting Student's Performance: A Case Of Private Colleges	Syed Tahir Hijazi S.M.M. Raza Naqvi[13]	Attendance Time allocation for studies Family income Mother's age Mother's education	Simple linear regression

### III. CONTRIBUTION

As discussed above, number of researches has been done on this topic. Different researchers have explored various factors with different findings and results. Each study explains various factors that affect student's performance with different effect on each student.

The contribution of this paper in research field is to explore the five categories of factors which directly or indirectly affect student performance with positive and negative impact. These five categories are family information, academic information, management skill, personality and miscellaneous. How these factors can affect performance of student will be discussed in next section. This study will help students, teachers as well as management of institutes to take proper and adequate steps to improve the student's performance and success of institute.

#### IV. IMPORTANT FACTORS

This section will explain five categories with their different indicators which effect student performance. Each category includes number of indicators which have different impact for students. Fig.1 shows each category with their corresponding indicators or factors.

##### *1. Family Information*

Family information includes factors like family income, parental education, family stress, time given by parents, correlation of parents with children, family responsibility etc [3],[5],[6]. Each factor has impact on student. Student with high Family income status provided with full facilities, fee for study instead of low family income student. Student with low family income leave study in between due to financial issues. Well educated parents can help their children in their study, can provide proper guidance at each and every step instead of illiterate or less educated parents. With high demand of expectation, responsibilities and suffering from family confliction stress student performance will be poor. Poor interaction of parents with their children and institute activities can negatively affect the student academic performance.

##### *2. Academic Information*

Academic information include past and present academic record of students like marks of tenth, marks of secondary education, attendance in class, marks of previous semesters, interest for course, time devotion for self study, learning style, assignments, attendance etc. some students choose course with their wish or some under the pressure of their parents wish, so those students may have less interest in course and that will degrade their performance [7],[15]. What learning style is follow in the class because all students have different learning style preference like read/write, visual and aural. Sufficient time devotion for self study is must after class will also give positive impact for good performance. Regular students get more marks than irregular students.

##### *3. Management Skill*

Management skill includes two types of management that time management and stress management:

*Time management:* Time is vital resource for everyone but everyone does not utilize it at the same level. Ability to do own work on time and use time effectively is a perfect time management. Student who complete their assignment on time or before deadlines, give priority to important activities, follow a revision timetable and prepare for exam daily can perform better than those students who do not follow any time table and do last minute preparation for exams which considered as prime source of poor academic performance [2]. Effective use of time helps the students to balance their academic work load with other activities, and also they learn to manage tight schedule effectively.

*Stress Management:* Stress is a challenge to students in institutes and the way it is managed may reflect in their academic performance. The challenges include handle excess workload, effective use of time, managing situations and managing emotion, developing necessary competence. The effects of stress can be positive or negative [11]. Positively used, stress can be a motivator by which if student take stress in positive way, he/she can accomplish target in given time. Stress however can be a barrier in student's academic achievements. If he/she perceives stress in negative way, this may cause mental illness and leads to failure. Ability to handle or mange above challenges in effective way is called Stress management skill.

Stress management skill and time management skill both play vital role in student's academic performance and also related with each other. Balancing the workload leads to good time management which lower down the stress of student.

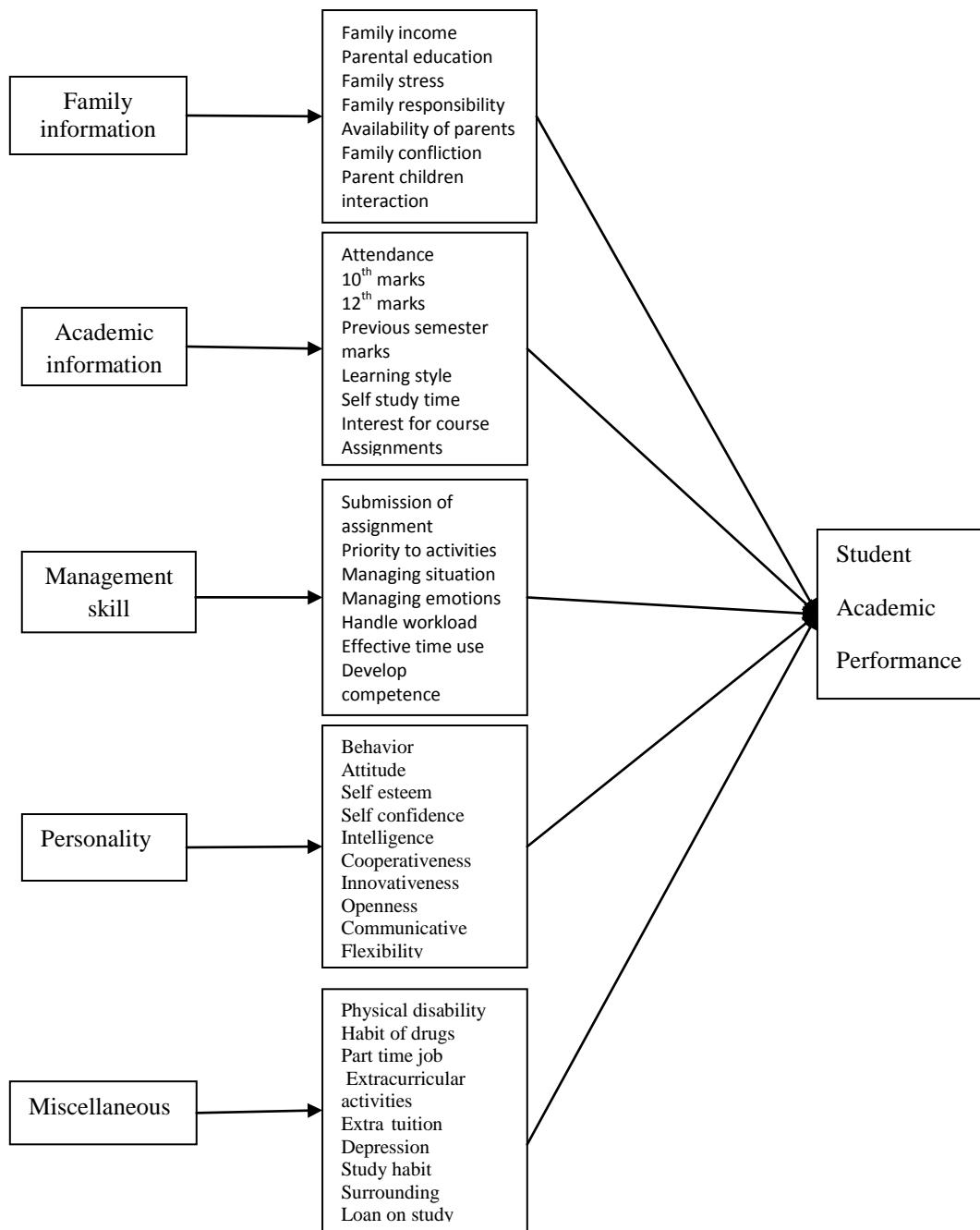


Figure1: Factors influencing Student Performance

#### 4. Personality

Personality is defined as behavior, attitude, self esteem, self confidence, intelligence, cooperativeness, innovativeness, openness, organized and determined, way to adapt the things, communicative, aggressive, fear, nervousness, sadness characteristics of persons [1],[10]. Every person has different personality with these characteristics. The student who can work in group, share their ideas, have courage to express their views and clear doubts openly will get more knowledge and grades. Self confident and intelligent students take part in debates, seminar or presentation share their knowledge and get new knowledge from others easily and also can evaluate their performance. But students with careless, less self esteem and fear characteristics are not able to communicate with others which can be the reason for poor performance. Open students are curious, interested and insightful to get new things and to update their knowledge. These students are organized complete their work on time and determined to achieve goals instead of careless and non openness students. Positive personality characteristics ensure the high academic performance of students.

#### 5. Miscellaneous

This category includes factors like any physical disability, habit of drugs like alcohol, smoke, etc, spend more time on phone and social sites, part time job along with study, extracurricular activities, take part in games, extra tuition, depression due to any reason and any accident happened in life[2],[9],[8]. Student may have habit of any type drugs or if they consume drugs daily can affect their problem. Some students have to do any part time job due to any reason or ant financial crisis so they cannot give sufficient time to their study and their mind can be distracted from study. It can negatively affect their performance. Depression can cause loss of concentration, mental illness leads to poor performance. These all factors have more or less affect on the student's academic performance. Some of them have positive impact and some of them have negative impact on academic performance.

#### V. PROPOSED METHODOLOGY

The proposed work is to perform an analysis of the student performance with above explained factors which influence the student achievements, their grades in exams. Figure2. Shows the procedure which will be follow to achieve the required valuable results.

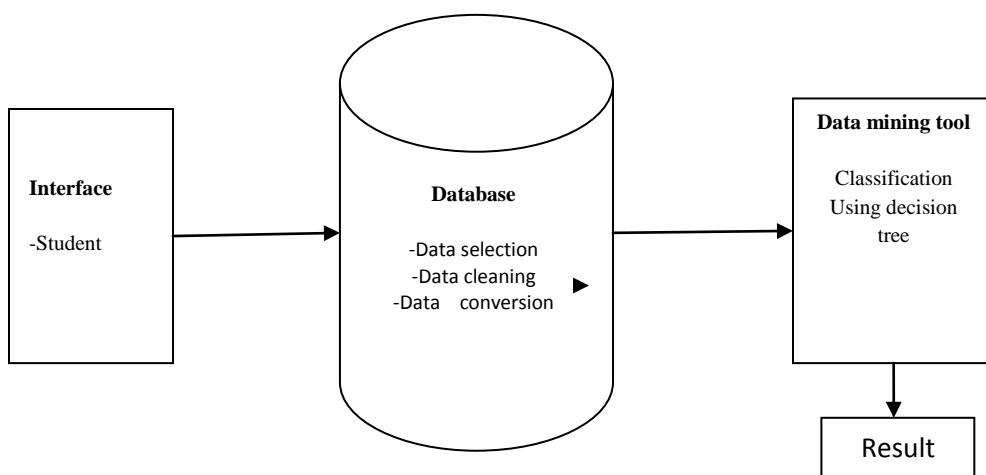


Figure2: Architecture of proposed methodology

The useful result will be produced by applying optimal data mining techniques on data that will be helpful for the students and faculty to improve student's performance.

Data will be collected from computer engineering student of colleges through questionnaire which contain question about their academic records, family, skills, personality etc. Collected data will be go through three processes as shown in figure2:

Data Selection

Data cleaning

Data conversion

Data selection and cleaning will be done for errors, missing value and data inconsistency. Then data conversion will be done by transferring the data into excel spread sheet.

For analysis we will apply decision tree classification technique using WEKA tool. Today various data mining tools are available but WEKA is free available tool and can easily run on almost modern computers.

From analyzed result we will define high or impact of parameters and this result can be used for decision making to improve performance by reducing the influence of parameters.

#### VI. CONCLUSION AND FUTURE WORK

As we know, Excellence is the important part of student life and also for the growth of higher institutes and universities. So, in this paper we studied the research area of student domain to know student performance and factors that may affect their performance. So that appropriate measures can be taken to improve the performance and success of institutes. We discussed such factors which affect the student's performance in various manners. Our goal is to decide which factors are more valuable and which are less valuable in student performance.

In our future work we will analyze the student performance with these factors to know the value of their effect on performance. For this analysis first we will collect data from computer engineering students of different colleges through questionnaire. Then we apply Classification technique decision tree using WEKA tool. From the analyzed result we will define which parameter is positively or negatively affect the performance and which is more valuable and which is less valuable.

#### REFERENCE

- [1]. Bi Zhu, "Individual differences in false memory from misinformation: Personality characteristics and their interactions with cognitive abilities", Elsevier, Personality and Individual Differences 48, 2010.
- [2]. Çigdem Mercanlioglu, "The Relationship Of Time Management To Academic Performance Of Master Level Students", International Journal Of Business And Management Studies, Vol. 2, 2010.
- [3]. Gillian Considine, "Factors Influencing the Educational Performance of Students from Disadvantaged Backgrounds", National Social Policy Conference, 2001.
- [4]. Herminio Rodriguez, "Factors influencing student's academic performance in the first accounting course", November 2005.
- [5]. Irfan Mushtaq, Shabana Nawaz Khan, "Factors Affecting Students' Academic Performance", Global Journal of Management and Business Research, 2012.
- [6]. Jedsarid Sangkapan, Kasetchai Laeheem, "Factors Affecting Students Academic Achievement into Probation Status at Prince of Songkla University", The 3rd International Conference on Humanities and Social Sciences, April 2,2011.
- [7]. K.Rajeswari, Suchita Borkar, "Predicting Students Academic Performance Using Education Data Mining", IJCSMC, Vol. 2, Issue. 7, July 2013,
- [8]. Lawrence C. Caranto, Sunshine B. Alos, "Factors Affecting the Academic Performance of the Student Nurses of BSU", International Journal of Nursing Science, 2015.
- [9]. Lindsay S. Ham, "College students and problematic drinking:A review of the literature", Clinical Psychology Review 23, 2003.
- [10]. Meera Komarraju, The Big Five personality traits, learning styles, and academic achievement", Elsevier, Personality and Individual Differences 51, 2011,

- [11]. Owoyele, Jimoh Wale, "Relationship between Stress Management Skills and Undergraduate Students' Academic Achievement in Two Nigerian Universities", An International Multi-Disciplinary Journal, April 2009.
- [12]. Pauline Smith, "Intelligence and educational achievement", Elsevier Intelligence 35, 2007.
- [13]. Naqvi, Syed Tahir Hijazi, "Factors Affecting Students' Performance: A Case Of Private Colleges", Bangladesh E-Journal Of Sociology, January 2006.
- [14]. Sulekha Anand and Julio G. Soto, "Factors influencing academic performance of students enrolled in a lower division Cell Biology core course", Journal of the Scholarship of Teaching and Learning, Vol. 9, No. 1, January 2009.
- [15]. Victor Mlambo, "An analysis of some factors affecting student academic performance in an introductory biochemistry course at the University of the West Indies", Caribbean Teaching Scholar Vol. 1, No. 2, November 2011.

# Automated Approach for Checking Consistency in UML Use Case and Sequence Diagram

Ramanpreet Kaur, Dhavleesh Rattan

Department of Computer Engineering,  
Punjabi University, Patiala

**Abstract:** Unified Modeling Language (UML) is a language in which pictorial representation makes a great significance. There is a disadvantage of inconsistency in the language. This is a very vast problem of the language which is not resolved yet. A lot of work has been done to remove the problem by the number of researchers but still more work is going on the language. The user has to suffer a lot if the problem get occurred. Consistency is termed as the compatibility of the system in which elements of different diagrams are combined and make a satisfied system which meets the required expectations of the user. In this paper, an automated approach is presented in which parsing technique and consistency rules are used. The consistency of two diagrams, i.e. use case and sequence diagram has been checked. Object Constraint Language (OCL) constraints are used for validation.

**Keywords:** UML diagrams, OCL constraints, consistency rules, parsing technique, automated approach.

## I. INTRODUCTION

UML is a language which is used to design a model. It is used for diagrammatic representation of the UML model. It is easy for any human being to recognize a picture more clearly as compared to other mathematical terms or codes. UML contains two types of diagrams: static and dynamic [1]. The static diagrams are those which presents only the structure of the diagram whereas dynamic diagrams present both the structure and the operations [4]. This language is simple due to which large and complex systems become easy to understand. Numbers of diagrams are used in the language which makes a model on combining with each other [5]. But as the diagrams are combined, the elements are also combined due to which there are more chances of inconsistency. The inconsistency reduces the quality of the system up to a great extent. The user can't get the required results because of the inconsistency. So the model should be consistent. Consistency means that all the elements contained in the diagrams must compatible with each other and system should be error free. Consistency of diagrams means excellent system that depicts the higher quality of the system. Consistent diagrams will reduce the errors and bugs in the system.

In this paper, two diagrams are compared to check their consistency. If diagrams are consistent, code is generated else code is not generated. OCL constraints are used only for validation. Time of consistency checking, error efficiency and design efficiency of diagrams are the factors that are to be calculated.

## II. RELATED WORK

The proposed work is related to different fields like parsing technique, consistency rules and OCL constraints. The quality of software development is very important to make the software error free. Changes in software may cause inconsistency in the project [2]. Poor quality has high chances in generation of failure. Inspection and testing are the main approaches used in the development of software. M. Thirugnanam et al [6] used parsing technique for checking consistency. The use case diagram was selected to convert the diagram into XML file format because UML cannot remove inconsistency from design diagram. After this conversion, XML tags and relevant information were extracted by using parsing technique. This technique divides the tags into tokens and extracts some relevant information. If the relevant information satisfied all the consistency rules, then the model was consistent otherwise inconsistency occurred between the diagrams. Consistency rules are used to identify inconsistencies in UML diagrams [7]. These rules are statements that give the way to the user with some restrictions. It guides the user that how to work in UML diagrams. Thirteen consistency rules were proposed for six UML diagrams i.e. use case diagram, activity diagram, state machine diagram, sequence diagram, class diagram and communication diagrams [7]. Consistency rules can be different for each diagram or it can be used mutually. Three mutual consistency rules were applied to use case and activity diagram using logical approach [3]. There are several consistency rules but some of them are ambiguous which are not adopted by Object Management Group (OMG) [8]. A. H. Khan et al [9] used class diagram and apply OCL constraints on it. OCL constraints were used to apply some restriction on the diagrams. The diagram followed only those instructions which were allowed by the OCL constraints. Translation time was also computed and showed that if number of elements were increased then it consumed much time.

## III. TOOLS AND TECHNIQUES USED IN PROPOSED WORK

### A. *My Eclipse*

Eclipse is a tool that has an Integrated Development Environment (IDE) for programming languages like java, C, C++ etc. Java programming language is used to develop My Eclipse. Hence, rich client application, IDE and other tools are developed using this tool.

### B. *NetBeans*

NetBeans is a software development platform written in Java. Modules are a set of modular software components that are used to develop applications allowed by NetBeans platform. The NetBeans is developed in Java, but many other languages like PHP, C, C++ and HTML are also supported by this platform. It is a cross-platform which runs on Microsoft Windows, Mac OS X, Linux, etc.

### C. *Parsing Technique*

Parsing technique is applied on the XMI tags that convert tags into tokens and the relevant information of diagrams will be extracted.

### D. *Consistency Rules*

Consistency rules for use case and sequence diagrams are given below:

1. Each object should be connected to at least one other object.
2. No element should be missing, i.e. diagram should be fully complete.
3. Objects in both the diagrams should be equal, i.e. number of elements in use case diagram matches with that of the sequence diagram.
4. In both the diagrams, same elements should be linked together.
5. The operation should be in order in both the diagrams.
6. If in use case diagram one objects calls any other object, in sequence diagram the same object should be called by that object.
7. Same examples should be used in both the diagrams.

#### E. OCL Constraints

OCL constraints are generated from OCL language. These constraints are conditional statements and have some restrictions.

#### IV. FLOWCHART OF THE PROPOSED WORK

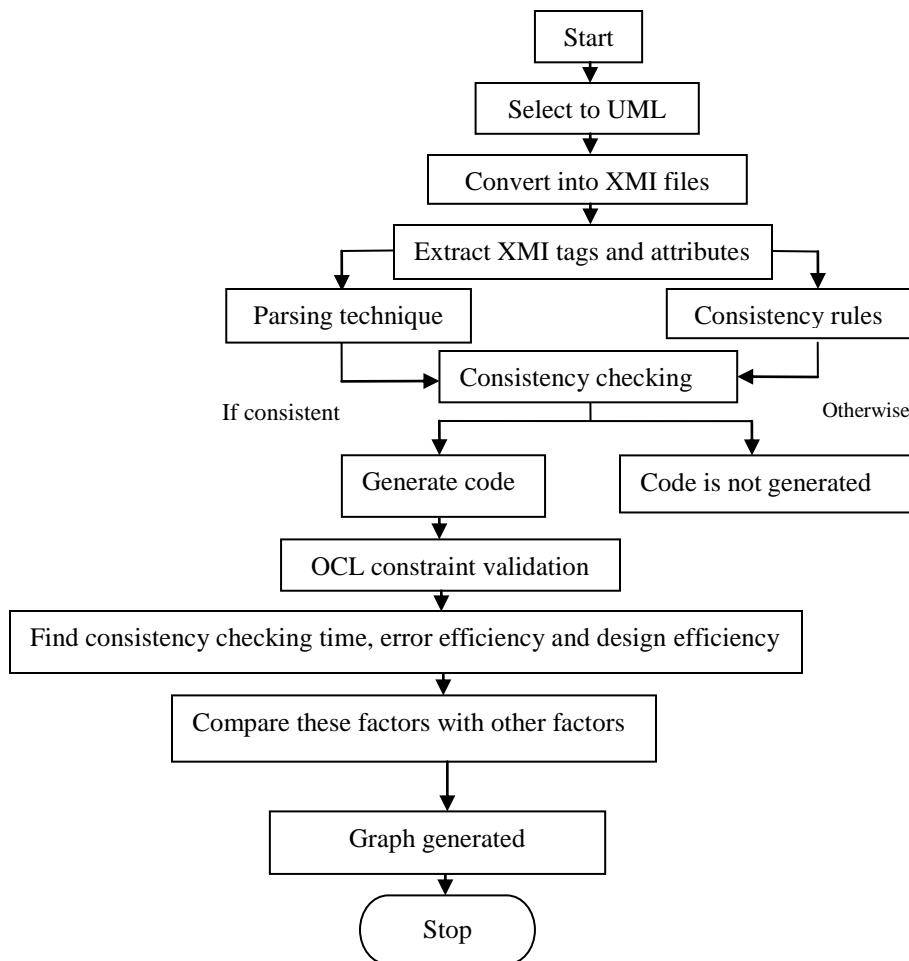


Figure 1. Flowchart representing the working of the system.

## V. DESCRIPTION OF PROPOSED WORK

*Step 1:* Make UML diagrams in My Eclipse tool.

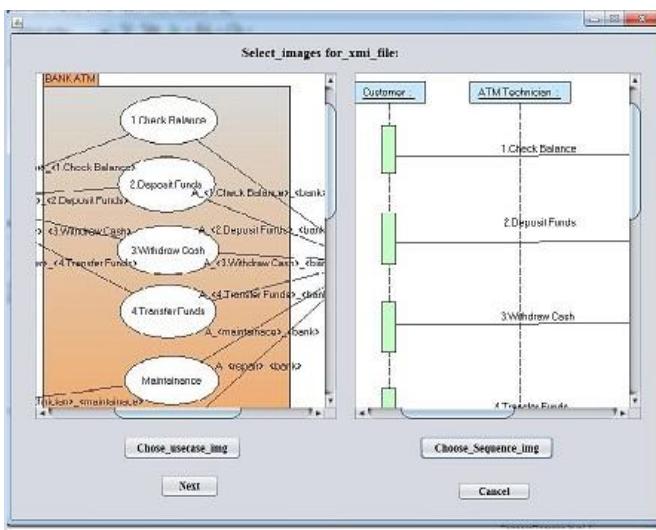


Figure 2. Use case and Sequence diagram in My Eclipse tool.

First select two UML diagrams i.e. Use case and Sequence diagram. After this make them in My Eclipse tool as shown in fig. 2. My Eclipse tool is java based tool. Both use case and sequence diagrams are dynamic diagrams.

*Step 2:* Export these diagrams in XML Metadata Interchange (XMI) file. XMI is a file format which is adopted by Object Management Group (OMG).

*Step 3:* Choosing the XMI files of Use case and Sequence diagram.

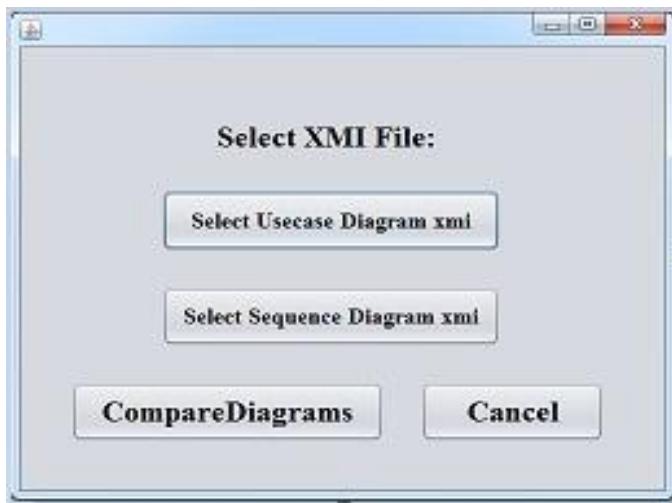


Figure 3. Select XMI files of Use case and Sequence diagram.

A dialog box is open which consist of selection of XMI file of both the UML diagrams named as Use case diagram and Sequence diagram. Select the XMI files of the use case and sequence diagram and then select the button of

compare diagrams which is given in the fig. 3. Apply parsing technique and consistency rules to check the consistency of both the diagrams. This technique also extracts the relevant information of the system.

*Step 4:* Extract the XMI tags and attributes from the XMI files.

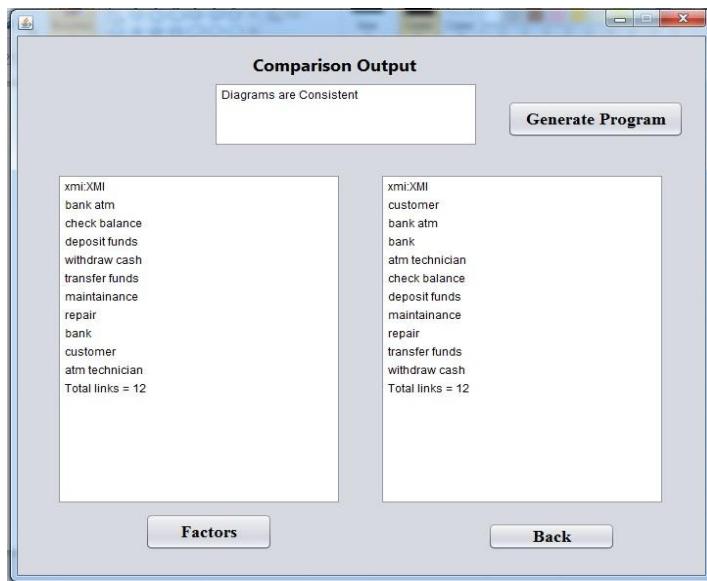


Figure 4. Comparison output box.

If diagrams are consistent, then it generates code else code is not generated. Fig. 4 shows the output of consistent diagrams which means that it is able to generate code.

*Step 5:* Dialog box of the Java program.

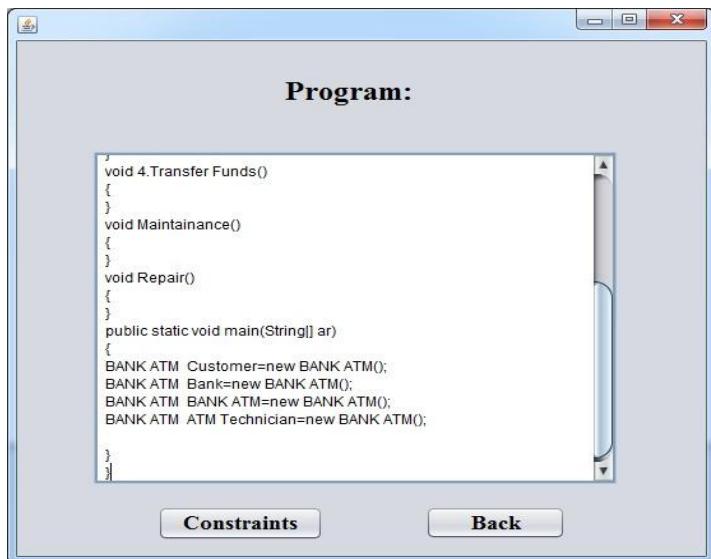


Figure 5. Dialog box of Java program.

After comparison it generates the dialog box of the Java program in which classes and attributes are declared in the given fig. 5.

*Step 6:* After the generation of code, validation of OCL constraints has been done.

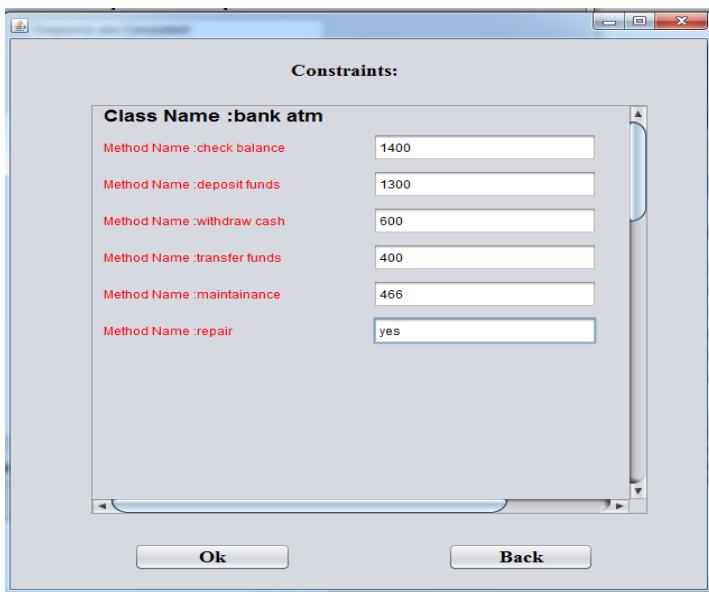


Figure 6. Constraint dialog box.

After generation of program fig. 6 shows the box of constraints which have some values. We put different constraints on the methods according to the user requirement, some of the them took numerical values and other have character and float values.

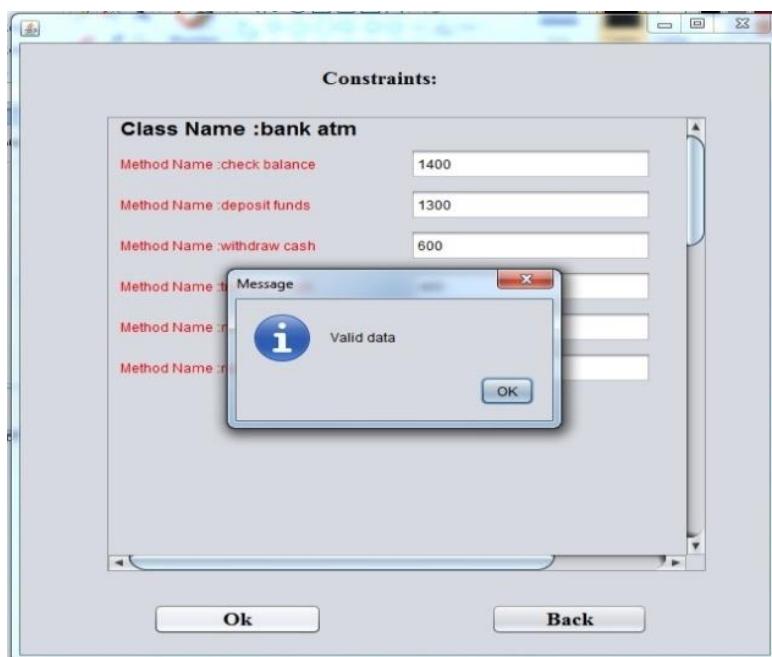


Figure 7. Message regarding valid data.

If all fields contain valid data then it will show the validation message. Fig. 7 shows the message of valid data.

*Step 7:* Check consistency time and error efficiency, design efficiency and number of elements in each diagram.

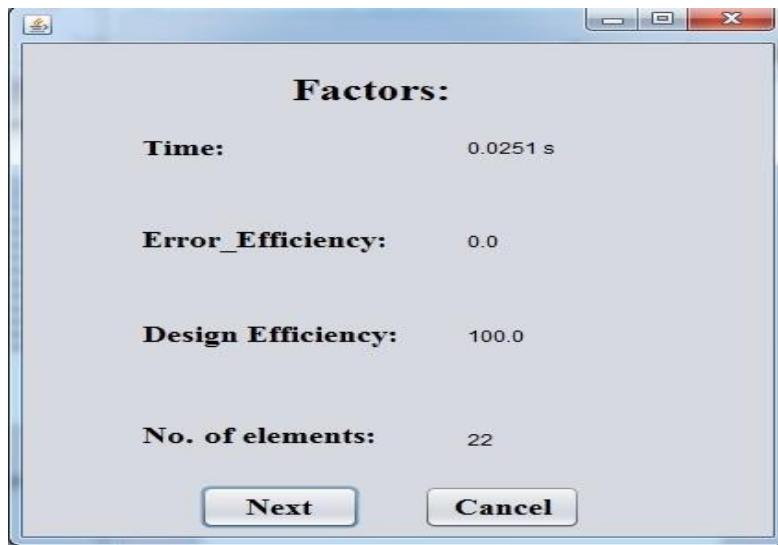


Figure 8. Factors of consistent diagrams.

Fig. 8 presents the factor of consistent diagrams in which the value of error efficiency 0.0 and design efficiency 100.0. For consistent diagrams error efficiency should be 0% and design efficiency should have percentage 100%.

*Step 8:* Compare the consistency checking time of the Use case and the Sequence diagram with other time factor which is using other technique.

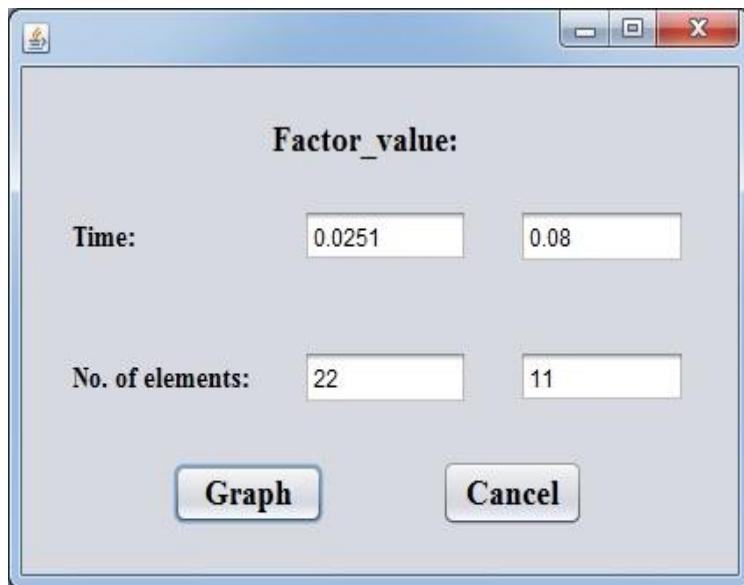


Figure 9. Factor value box.

Fig. 9 have the values of time factor and elements of the diagram. It also shows that the proposed technique takes less time than the other one.

## VI. RESULTS

*Time Graph:* Graph shows the comparison of two different techniques.

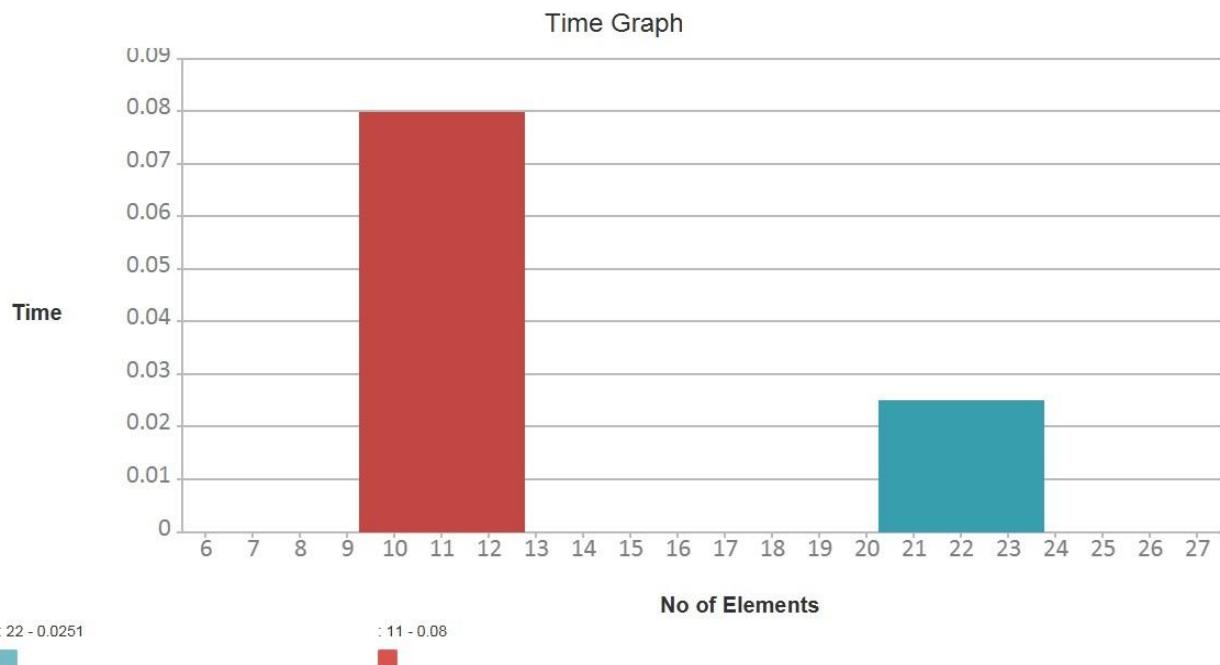


Figure 10. Time vs. No. of Elements.

Fig. 10 represents the execution time of the consistent diagrams. In this it shows that the execution time of our technique is very less as compared to the other technique.

TABLE I  
 FACTORS OF INCONSISTENT PROJECTS

Project	No. of elements	Consistency checking time (sec)	Error efficiency	Design efficiency
Bank- ATM	22	0.03	0.86	99.14
Library	15	0.0161	0.93	99.7
Restaurant	18	0.0195	0.83	99.17

Table I contains the factors of inconsistent projects in which values of error efficiency and design efficiency got change according to the inconsistencies but the value in consistent diagrams will always remain constant. The value of error efficiency is 0.0 and design efficiency is 100.0 if the diagrams are consistent.

## VII. CONCLUSION

It is very difficult for the novel user to detect the inconsistency. So we presented this technique in such a way that novel user can easily find out the inconsistency. In this method parsing technique and consistency rules are used. The technique takes less time as compared to other techniques. Moreover it provides the design efficiency. OCL constraints are used only for validation and further work can be done based on it. There are three parameters i.e. consistency checking time, error efficiency and design efficiency based on the performance of the algorithm is computed. The proposed technique is also simple than the other techniques.

## REFERENCES

- [1] H. Rasch and H. Wehrheim, "Checking Consistency in UML Diagrams: Classes and State Machines," *International Federation for Information Processing*, vol. 2884, pp. 229-243, 2003.
- [2] G. Thimm, S. G. Lee, Y. S. Ma, "Towards unified modelling of product life-cycles," *ELSEVIER, Computers in Industry*, vol. 57, pp. 331-341, 2006.
- [3] N. Ibrahim, R. Ibrahim, M. Z. Saringat, D. Mansor and T. Herawan, "Consistency Rules between UML Use Case and Activity Diagrams using Logical Approach," *International Journal of Software Engineering and its Applications*, vol. 5, pp. 119-134, 2011.
- [4] L. Mia and K. Ben, "A Method of Software Specification Mutation Testing Based on UML State Diagram for Consistency Checking," *ELSEVIER, Advanced in Control Engineering and Information Science*, vol.15, pp. 110-114, 2011.
- [5] A. Egyed, "UML/Analyzer: A Tool for the Instant Consistency Checking of UML Models," *IEEE, International Conference on Software Engineering*, pp. 793-796, 2007.
- [6] M. Thirugnanam and S. Subramaniam, "An Efficient Design Tool to Detect Inconsistencies in UML Design Models", *International Journal of Computer Science and Business Informatics*, vol. 9, 2014.
- [7] X. Lui, "Identification and Check of Inconsistencies between UML Diagrams", *IEEE, Computer Sciences and Applications*, pp. 487-490, 2013.
- [8] R. Dubauskaite and O. Vasilecas, "Method on Specifying Consistency Rules among Different Aspect Models, expressed in UML," *ELEKTRONIKA IR ELEKTROTECHNIKA*, vol. 19, pp. 77-81, 2013.
- [9] A. H. Khan and I. Porres, "Consistency of UML class, object and statechart diagrams using ontology reasoners," *ELSEVIER, Journal of Visual Languages and Computing*, vol. 26, pp. 42-65, 2015.

# Introduction to 2-D Barcode and QR-code with Applications

Er. Abhi Sidana<sup>#1</sup>, Dr. Lakhwinder Kaur<sup>#2</sup>

<sup>#</sup>Department Of Computer Engineering  
Punjabi University Patiala, Ucoe, India

<sup>1</sup>abhisidana95@gmail.com

<sup>2</sup>mahal2k8@gmail.com

## Abstract

Various kinds of barcodes have been used in emerging fields of m-commerce. Two dimensional barcode gain rapid achievements over different types of barcode like Datamatrix, Beetag, PDF-4417 codes. With the advancement of technology one dimensional barcode is shifted to two dimensional barcode for their large storage capacity and error correction level. Mobile industry gaining more attention to present diverse commerce data and enhancing the user experience by storing more data and reduced input. This paper introduces 2-D barcode concepts, their types and QR-code structure. The process of mobile tagging to connect to real world web along with different applications in m-commerce. The aim is to analyze the QR-code so that it can be prevented from outside attacks. Paper also represents different security levels of QR-code.

**Keywords:** QR-code structure, mobile tagging, 2-d barcode types, QR-code security.

## 1. INTRODUCTION

With the rapid development of 2-D bar codes more number of bar codes are gaining popularity in our daily lives. For product monitoring and goods tracking 1-D bar codes were used to store different product information, found on the packaging, retail sale and buy. Today digital barcodes are effective means for mobile communication system because of reasons:

1. It is simple method to present enormous data in m-commerce as it includes product information, payment and purchasing.
2. It improves the experience of a mobile user by reducing inputs as number of built in cameras deployed on mobile phones easily capture the QR-code and extract information.
3. The merging of 2-D barcode technology and mobile phone technology in many countries allows linkage to digital web world from physical world.

Emerging needs of mobile based QR-code applications can be meet by understanding the concept of QR-code, more research and technology work are needed in m-commerce services. Design pattern and good understanding of encoding in QR-code help engineers to make 2-d barcode applications for mobile device. Many of research papers showing different classifications, technologies are available but there is lack of survey papers focusing on m-commerce application system and implement reports on structure of QR-code having threat models in it. This paper first discusses about general concept of 2-d barcode showing their types then it examines the QR-code behavior and structure. Applications in m-commerce based on recent study and survey. Then our paper examines the possible attacking parts of QR-code to provide it maximum security.

This paper is organized as following: II section discuss brief overview of concept of 2-D barcode including different types of 2-D barcodes. III section explores QR-code structure. IV section describes mobile tagging and applications in m-commerce including some challenges in real world for creation and design. V section describes the attack strategies which review the mask, character count indicator, character encoding and threat models in QR-code. Finally this paper remarks the conclusion and future work proposed.

## 2. CONCEPT OF 2-D BARCODE

In 2007, Gao, Prakash presented importance of 2-D barcode, their usage and application in m-commerce. They improve the experience of a user in mobile commerce by reducing the inputs. Barcode appeared decades ago when 1-D barcode were used first for railway transportation and tracking of goods and services in USA (1). Traditionally 1-D barcode stored data in parallel lines of varying widths and encode numbers only. Linear barcodes also known as 1-D barcode can be found virtually on different products for storing limited amount of information. Nowadays it has grown many needs to fulfill the government, business cards, health care, library and small scale enterprises. 1-D barcode is the process of encoding numbers into a sequence of bars of different widths so that it can be used for extracting information later which can be read by using different scanners and mobile camera equipped phones. As information is limited in 1-D barcode, large amount of storing information as compared to linear barcode can be introduced with the development of 2-D barcodes in 1980. They store enormous data in small area and require decoding device such as mobile device. 2-D barcode can encode numbers, letters, punctuation marks, all the alphanumeric characters, numeric characters. It became popular with large data capacity storage. The protocol which defines standard for arranging spaces and modules is called symbology of barcode. It also defines technical details such as character set, encoding, error correction.[1]

## 2.1 Types of 2-D barcode

In 2008, Schmidmayr, Ebner, Kappe discussed different types of 2D barcodes for mobile tagging. According to their overview barcodes can be divided in two ways a) Matrix 2-D barcode e.g. include QR and Datamatrix b) Stacked 2-D barcode e.g. Code49 and PDF-417. Several barcodes were developed in mobile tagging.

**a) Datamatrix:** This code was developed by department of defense and its standardization is done[2] by NASA. Recent study shows it is most space efficient type of barcode. Comparison of datamatrix and QR lead to bright scope of datamatrix. It is used in marketing of electronic components in industrial production.

**b) Beetag:** This type of barcode allows a placement of logo for branding in center of symbol. It was optimized to address specific issues of western world mobile market where autofocus functionality is equipped in mobile phones so easily could be encoded without macro lens.

**c) Aztec Code:** It was invented in 1995 and do not require any quiet zone around code. Data in this code is put in spiral form around center mark and fits up to 3832 digits, 3067 letters and 1914 bytes. Popular for european rail components.[6]

**d) QR Code:** In 2015, Gaikwad, Singh expertise of the developers introduce concept of QR-code. It stands for quick-response code as it decodes the content at a very high speed. It is basically a type of 2-D barcode introduced by Denso Wave in 1994, a Japanese company. Features of QR code are high data density, ability to encode no, text, kanji characters. It stores information in both vertical and horizontal direction as compared to linear barcode which only stores data in one direction. It can store 7089 characters, 4296 alphanumeric characters, 2953 bytes of binary data[3].One of the main feature of this code it can be scanned from any angle by decoder.

## 3. QR-CODE STRUCTURE

In 2010,Schrittwieser, Kiesberg and Sinha examines the QR-code and present his structure in a broad way. Structure of QR-code consists of different patterns and designs that have well defined functions. The information is encoded in square black and white modules.

**1) Finder pattern:** It plays a significant role in success of decoding. It consist of 3 identical square located in all corners except at lower right bottom corner of code. This pattern have equivalent ratio of black and white pixel so allowing decoder software to determine its center regardless of position scanned.

**2) Separators:** Finder pattern are surrounded by guard zone of 1 QR 10 module wide called separators. They have one pixel width and separate the finder pattern from the actual data in QR-code.

**3) Alignment Pattern:** Moderate distortion of an image is compensated by alignment pattern. Sampling grids are also determined with the help of this pattern.

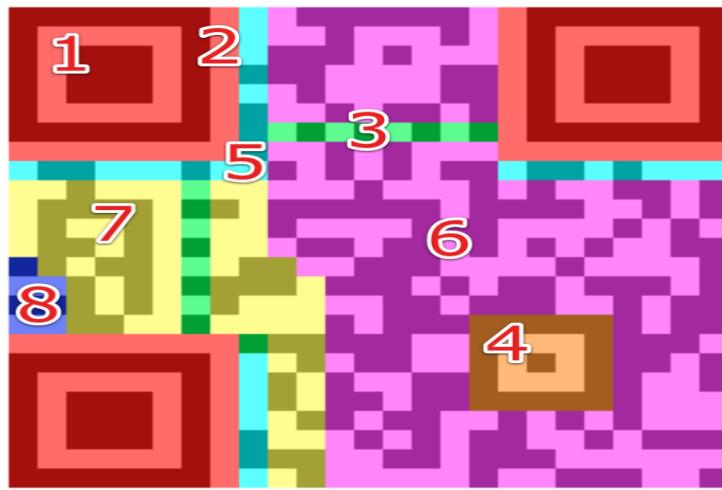


Fig. 3 1 Basic Structure of QR-code (5)

**4) Format information:** It consists of 15 bits after separators and stores information of error correction code in 8-bit long code words. It also reports mask information.

**5) Data:** The actual data is stored in 8-bit parts and converted into a bit stream. 8-bit part is also called codeword.

**6) Remainder bits:** If error correction and data bits cannot be divided into codeword's without remainder they will be empty bits.

**7) Timing pattern:** It consists of on black and white module located between finder patterns. Software decoder determines the width of a one module with timing pattern.

The outside of QR-code is a zone called as quiet zone after which decoder do not recognize anything.

#### 4. MOBILE TAGGING

In 2008, Schmidmayr and Ebner and Kappe provided the process of mobile tagging and applications related to print media. The process of capturing the barcode image with embedded camera on cellular phone, decoding the image and then linking to real world web to get enormous data information from it is called as mobile tagging[3]. In Japan 75 percent of users have built in barcode reader in phones and are using to decode or extract specific information depending upon nature of application. E.g. a sms can be sent just by scanning, URL decodes, automatically dialing instead of typing. In print media its use is at most. The advantage of ad tracking are common to physical world.

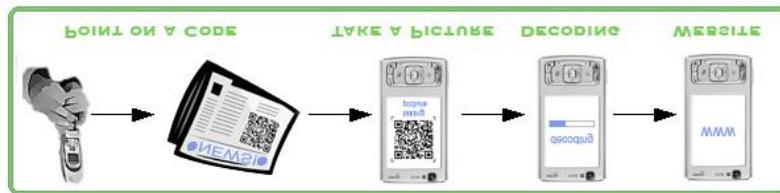


Fig. 3.2 Graphical illustration of process mobile tagging[5]

##### 4.1 Applications in M-commerce and challenges:

1. Commercial use of QR is in telecommunication where many smart phones are largest driver for its popularity.
2. With recent technology development 2D barcodes are used in buy-sell, post sale activities wireless trading for commerce transaction.
3. QR-code now appears in advertisement, coupons and product covering materials that can be decoded afterwards with mobile devices.
4. For wireless users it can be used to access a site using phone by just scan instead of typing it manually and for marketing.
5. Use of QR code is an effective and efficient technique without entering keyboard entry to do something action.
6. For product monitoring and tracking it can be used.
7. Many mobile payment systems, now gained popularity through use of QR-code.

##### 4.2 Improvements Needed:

There are certain issues which can be resolved by providing security to QR codes. These issues are:

- 1) Datamatrix and QR code are standardized codes but there is lack of internal structure of data encoded.
- 2) Older versions of devices with cameras don't allow encoding smaller 2D codes. Coming smart phones are a solution to this.
- 3) In advertisements barcodes can be replaced by others ones.
- 4) It is necessary to find a specific path so that entering events and other data would be easy to enter.

#### 5. PROVIDE SECURITY TO QR-CODES

QR-code security presented by Narayanan discussed various attacks and their consequences. As these codes are readable by machine only so author explored different kinds of strategies from attacker point of view.

**5.1 Threat Models:** There are basically two types of threat models for manipulating the QR codes. First model is in which an attacker can invert any module of code from either white to black or vice versa. Second model deals with changing white module to black and not vice versa. These are explored as follows:

**5.1.1 Both colors threat model:** In this an individual can copy the style of existing code and place their own on the original by making a sticker. This can be easily done with the help of adobe designs. In this way attacks can be done on the QR code.

**5.1.2 Single color threat model:** In this individual modify the single color only and can attack the QR code.

**5.2 Attacking Strategies:** Structure of QR code contain a lot of information such as data encoding, Function pattern, mask, version information, which can be used by many attackers to manipulate the original code. Since our eye can't distinguish between these codes quickly and properly it can only be understood by decoder or a scanner so to provide security to QR we need to analyze the different attacking parts of codes to make them more secure.[4]

**5.2.1 Masking:** Modules which are used to create QR consist of black and white pattern collectively known as mask. Mask help decoding devices and also increases pictures contrast. When we generate a QR code every mask of 8 specified ones is applied is rated according to that mask will be chosen. One mask in every QR code is necessary and version is encoded with mask. One can target the mask and can change data so strong error and correction algorithm is used to protect is from attackers.

**5.2.2 Mode:** It is basically a character encoding through which information is contained in code. We can have numeric mode (contain only digits), alphanumeric mode (which can contain lower case uppercase letters \$ % etc), Kanji characters. Mode is defined at the start of data part by leading 4 bits. Attacks to these modes can be done by mixing the modes in one QR code.

**5.2.3 Character Count Indicator:** It is just after the mode indicator and it mainly depends upon version high version will have longer character count. Attacks in character count indicator can be buffer overflow and buffer underflow. In buffer underflow decoding device only decodes the first few letters and next data part is seen as a new segment. In buffer overflow decoder decodes the filler part and adds more amounts of data in that space.

## 6. CONCLUSION

QR-codes gain more popularity with the advancement of mobile technology to encode large information. In this paper, the general concepts of 2D barcode, structure, their types and their applications have been discussed. The emerging need of QR-code in Wireless trading and M-commerce applications mobile tagging is presented. This paper reviews the possible attacks in manipulating QR code. The work presented is an attempt to gain the security information in QR-code to prevent it from attacks.

## REFERENCES

- [1] J.Z.Gao, Lekshmi Prakash and R.Jagatesan,"Understanding 2D-Barcode technology and applications in m-commerce-design and implementation of a 2d-barcode processing solution", In COMPSAC IEEE Computer Society, pp.49-57, 2007.
- [2] Akshara Gaikwad, K R Singh, "Information Hiding Using Image Embedding in QR-codes for color images", IJCSIT, vol.6 (1), pp.278-283, 2015.
- [3] Paul Schmidmayr, Martin Ebner, Frank Kappe,"What's Power behind 2D Barcodes? Are they the Foundation of the Revival of Print Media?" 6th International Conference on Knowledge Management and New Media Technology, pp234-242, 2008.
- [4] A. Sankara Narayanan,"QR Codes and Security Solutions", International Journal of Computer Science and Telecommunications, vol.3, pp.69-72.July 2012.
- [5] [https://en.wikipedia.org/wiki/Mobile\\_tagging#/media/File](https://en.wikipedia.org/wiki/Mobile_tagging#/media/File): Mt\_process\_english.jpg.
- [6] Peter Kieseberg,"QR Code Security", SBA research, Vienna Austria.

# A Review: Biometric Security Mechanism in E-commerce

Sumit Bansal <sup>#1</sup>, Er. Supreet Kaur <sup>\*2</sup>

<sup>#1</sup>Student, M.tech Scholar, Computer Engineering Deptt.

Punjabi University, Patiala-147002

<sup>#1</sup>sb002@ymail.com

<sup>\*2</sup>Assistant Professor, Punjabi University, Patiala-147002

<sup>\*2</sup>supreetgill13@gmail.com

**Abstract-**Internet is an amazing tool and it has changed the lifestyle of the people considerably. It is the ease of use, efficiency and time factor, which have contributed to the growth and popularity of E-commerce [1]. Some of the issues that are occurred during payment are speed, time, durability, security etc. In these days, online transactions using internet enabled mobile phone, Tablets and Laptops has been widely used. Debit/ Credit Cards are used for payment for online business or E-payments and online banking is also practically used. As the technology is rising day-by-day misuse of this technology has increased. Cards frauds like using phone calls, phishing attack, card cloning, hacking the internet enabled devices [2]. Due to these type of problems in the E-commerce, biometric system can be used. Biometric have proved to be good security mechanism and has been implemented in some areas. This review paper reviews use of biometric system to reduce these types of the unauthorized accesses and misuse of information. Various techniques used by researchers include biometric models, PIN, Password can be used to overcome the problem. An increasing number of people are overwhelmed by the efficiency and convenience of internet for making web-based transactions globally, as it is the technology of the 21st century [3].

**Keywords:** - E-commerce, CIA, Fingerprint, PIN, Password, Face Recognition, Cryptography, Authentication

## I. INTRODUCTION

### A. E-commerce

In wireless trading (Pre scale, post scale, buy) m-commerce pays an important role. It satisfies the growing needs of organization, business purpose etc. in our daily lives. E-commerce is selling or buying goods or services with the help of internet enabled device. E-commerce is a fast and easy to use method in these days. E-commerce provide many services like ticket booking, mobile banking, mobile purchasing, online transactions, online shopping etc. Due to these types of services it provides an opportunity to users for universal devices to be used anywhere freely [4].

### B. Issues

1. Password hacking is one of the main problem in these days. Hackers hack a password of a person and misuse its services.

2. *Phishing Attacks:* - Phishing refers to the process of acquiring login details password, user name in an electronic communication. Various types of phishing attacks are following: -

- Fake calls are one of the latest trending problem. An unauthorized person calling and want to know your password, PIN or other important information. This information is used to withdraw money from user accounts or shop from the victim money.
- Spam E-mails is the online mails that comes into any electronic mail to lure the user information trap where user gives confidential details. These details are then misused by attackers. Spam Mails gives you Lottery opportunity, Job or Foreign Dreams Opportunities.
- *Messages Phishing:* - Many types of forges are done by sending false messages to different users. It traps the user to give his confidential details which are misused by attacker later.

3. *Card Cloning:* - when we swipe the card on machine it captures the cards details like number, pin code, CVV number which can lead to misuse of card.

4. *Sniffing Devices:* - There are devices that monitor the data travelling over the network. Sniffers are very difficult to detect and very dangerous in capturing the data.

## II. BACKGROUND

### A. *The CIA Principle*

The basic principal for security analysis is CIA triad. It stands for confidentiality, integrity & availability. In any secure system these three parameters are necessary. These principals are applicable starting from the user's internet history to the whole encrypted data of internet [11]. These are elaborated as follows: -

- *Confidentiality*

It involves set of rules of promises that limit access on certain type of information. Cryptography and encryption are common to the confidentiality of transferring data from one device to another.

- *Integrity*

In simple words integrity means data in model is accurate and representation of original information in encrypted. Incorrect information cannot be entered into the system data will always be accurate.

- *Availability*

A means information is easily accessible to all the authorized viewer all the time. Only authorized person can view the data, can modify or delete the data at any time.

### *B. Biometric*

Biometric consists of methods for uniquely recognizing humans based upon one or more intrinsic Physical or Behavioral Traits [4][5]. In computer science biometric is used for access control and identification/authentication. In biometric unique characteristics of a person is identified and solved and later in verification phase are re-checked upon the result i.e., whether the person's traits match or not then the access is given.

- *Fingerprint*

This method is one of the widely used identification of a user or an individual. Fingerprint is used for security in biometric. Fingerprints of each and every person is unique.it is not possible to match finger print of two different persons. Fingerprint is also one of the good security way. We can extract the useful information by using minutiae in fingerprint.it is a cheap and easy to implement [6].

- *Face Recognition*

Face Recognition is the another idea that is used in biometric devices for identification.in face recognition all the faces match with database and it is accepted or rejected by biometric device. Face recognition is a fast and commonly used biometric idea [6].

- *Iris Recognition*

Iris contain pupil and outer responsible for identification of an individual. Pupil is dark color centered surrounded by another in which into is stored and extracted afterword by a ring and matched with database [6].

### *C. Advantages of Payment by Biometric Methods: -*

Some of the advantages over the biometric payment as compare to the other methods of payment defined below: -

- *Security*

In the other security techniques, the data is easily copied but in biometric payment systems data or keys cannot be copied easily.

- *Uniqueness*

All the biometric data is unique. In other security methods password, PINs may be same (like A=12345678, B=12345678 etc.) for two or more than two persons but in the Biometric authentication method it is not same. So Unique identification in this method.



- *Performance*

Performance of biometric data is good as compare to others methods of payment. In this user first enter the password or PIN then system granted but in biometric just touch the screen and access granted. Due to this the performance of the system is also increased.

- *No Need to Change*

In Biometric security no need to change password or keys time to time. In PINs and passwords security user change their security codes/keys to protect their system from hackers or unwanted problems that comes in the system but in biometric no need to change the security PINs of the system [12].

### III. LITRATURE SURVEY

Use Different researches carried out and different methods were used in E-commerce using biometrics. This paper represents some more techniques which are more secure at the same time reliable like secure e-payments or E-commerce methods are used. These methods are elaborated in different section of this paper.

A. In this research paper authors Simon Liu and Mark Silverman proposed different type of authentication *techniques* for e commerce/Payments [10].

In these techniques some personal number or password that is set by the user is entered by completing the transaction. In other methods/techniques a Password that may be a number or word sent to user's mobile phone and after receiving the PIN the Transaction is Completed. One of the other techniques that is discussed is Biometric technique. In this different Biometric methods are used to complete the transactions. Biometric technique provides a unique authentication for payments in the E-commerce. Many biometric techniques like face, finger, palm, iris, Signature, Voice are used in the E-commerce.

In 2001 silverman and liu discusses about the working of biometric system firstly capture the biometric which we have chosen, extracting the biometric template then store the template in repository, then match the template against stored template and record a secure audit with respect to system use.

B. In this research paper the author 'Rajendra Reddy Vangala Sreela Sasi Ph.D. used web based architecture to use encrypted iris Patterns as biometric attribute for authentication of a customer for e commerce transaction. In this approach need high level resolution images. Also discuss very effective method for preventing credit card frauds the iris images processed normalized and enhanced using component analysis to find the pattern in data. [2]

C. In Conclusion of "Security role of biometric in Electronic Transactions" author Mr. Anshuman Tyagi, Mr. Piyush Bhustyhan Singh, Mr. Vikash Singh Yadav, Mr. Sang harsh Kumar Singh, Dr. Amod Tiwari says electronic communication is very commonly used in all security requirements such as authentication, privacy/ confidentiality, integrity and non-repudiation. Cryptography is not helpful only for thefts but also helps in user authentication [7] [8] [9]. It also uses PIN, password and biometric for authentication. All the other type of the security ways is not easy to remember by any person but the biometric is the permanent or unique identification of user. In biometric password, token, PIN, remember problems are removed or finished. It also discusses fraud detection in which every individual who performs the transection registered in any device the information of individual is stored in database and it can lead to the misuse of store information further.

D. Author's Nikita Gupta, Devendra Mani Tripathi from International Journal of Computer Applications discussed the possible biometric identifiers. In this paper the accuracy of face print, voice is high but key stroke dynamic is medium. The survey reveals that two level authentication is better than single level authentication. First one is PIN based authentication and second is biometric authentication [5].

E. In research paper "Enhancement of security with the help of real time authentication and one-time password in E-commerce transactions" the authors Z. Zareh Hosseini & E. Barkhordari use three different types of authentication for providing security in E-commerce transection. From the authors point of view three identification are PIN based, OTP (one-time password) & biometric authentication.

In this research paper the researcher says we can provide robust identification and authentication to the E-commerce transection by biometric sensibility techniques like iris, fingerprint authentication with one-time password [7] [8].

TABLE I  
 PROGRESS IN USE OF BIOMETRICS

Sr. No.	Paper Year	Paper Title	Authors	Techniques	Conclusion
1.	2001	A Practical Guide to Biometric Security Technology	Simon Liu and Mark Silverman	PIN, secure ID, Biometric	Biometric more secure as compare to others.
2.	2004	Biometric Authentication for E-commerce Transection	Rajendra Reddy Vangala, Sreela sasi Ph.D.	Encrypted iris patterns	Iris biometric is very premature but need high resolution images
3.	2012	Security Role of Biometrics in Electronic Transections	Mr. Anshuman Tyagi, Mr. Piyush Bhushan Singh, Mr. Vikash Singh Yadav, Mr. Sang harsh Kumar Singh, Dr. Amod Tiwari	Cryptography	Cryptography also helps in user authentication
4.	2012	Comparative Study of Different Biometric Authorization for Mobile Payment System	Nitika Gupta, Devendra Mani Tripathi	PIN, Second Level Authorization	First level is PIN based and second level is Biometric Authorization
5.	2013	Enhancement of Security with the help of real time authentication and one-time password in E-commerce transections	Z. Zareh Hosseini & E. Barkhordari	PIN, Second Level Authentication, Face, Voice, Biometric	Fingerprint Iris biometric and OTP use for more security

#### IV. CONCLUSION

In all types of the security mechanisms in E-commerce some type of errors or security issues are involved. So it is important to improve its security problems.

There are many security issues as discussed above in E-commerce. Some researchers have provided the solution and some researchers are still being developed to resolve these issues and increase the performance, integrity, security and accuracy of the system. The biometric security mechanism the security in the E-commerce and cloud computing can be improved at a high level.

#### REFERENCES

- [1] R. Sato, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Web Site: [www.ijettcs.org](http://www.ijettcs.org) Email: editor@ijettcs.org, editorijettcs@gmail.com Volume 2, Issue 2, March
- [2] Biometric Authentication for E-commerce Transaction Rajendra Reddy VangalaSreelaSasi Ph.D. Cannon University, 109 University squareErie,PA16541sasi001@gannon.edu,vangala002@gannon.edu
- [3] Randy C. Marchany, Joseph G. Tront, "E-commerce security issues", Proceedings Of the 35th Hawaii International Conference on System Sciences, January 7-10,2002.
- [4] J. Biometric Security Mechanism in Mobile Payments Michael Gordon Mona Institute of Applied Science, University of West Indies, Kingston, Jamaica& Dr. Suresh Sankaranarayanan Department of Computing, University of West Indies, Kingston, Jamaica e-mail:pssuresh@hotmail.com
- [5] Comparative Study of Different Biometric Authorization for Mobile Payment System "International Conference in Recent Trends in Information Technology and Computer Science (ICRTITCS - 2012) Proceedings published in International Journal of Computer Applications® (IJCA) (0975 – 8887) Nikita Gupta, Department of Computer Engineering, Army Institute of Technology Pune, India Devendra Mani Tripathi, Department of Computer Engineering, Army Institute of Technology Pune, India.
- [6] Proceedings of the World Congress on Engineering 2011 Vol II, WCE 2011, July 6 - 8, 2011, London, U.K. "A Review of the Fingerprint, Recognition, Face Recognition and Iris Recognition Based Biometric Identification Technologies", Tiwalade O. Majekodunmi, Francis E. Idachaba.
- [7] Security 2013 5th Conference on Information and Knowledge Technology, Enhancement of security with the help of real time authentication and one-time password in E-commerce Transactions Z. Zareh Hosseini Department of Engineering & Technology Payame Noor University PO BOX 19395-3697 Tehran, I.R of IRAN Z.zhosseini@yahoo.com E. Barkhordari Department of Engineering & Technology Payame Noor University PO BOX 19395-3697 Tehran, I.R of IRAN Ehsanbarkhordari@yahoo.com.
- [8] Federal Financial Institutions Examination Council, authentication in an internet banking environment.
- [9] Security Role of Biometrics in Electronic Transactions Mr. AnshumanTyagi1, Mr. Piyush Bhushan Singh2, Mr. Vikash Singh Yadav3, Mr. Sangharsh Kumar Singh4, Dr. Amod Tiwari5 1 Research Scholar, Sai Nath University, Ranchi, Jharkhand. 2 Research Scholar, Sai Nath University, Ranchi, Jharkhand. 3 Research Scholar, CMJ University, Silong, Meghalaya. 4Research Scholar, CMJ University, Silong, Meghalaya. 5 Asst. Prof., Dept. of Computer Science &Engg., PSIT- Kanpur. 978-1-4673-1344-5/12/\$1.00 ©2012 IEEE.
- [10] "A Practical Guide to Biometric Security Technology" Simon Liu and Mark Silverman, 1520-9202/01/\$10.00 © 2001 IEEE.
- [11] Securing Online Shopping Using Biometric Personal Authentication and Steganography, "Hussam Ud-Din" A. Ihmadi, Prof. Ahmed Al Jaber and Dr. Amjad Hudaib, 0-7803-9521-2/06/\$20.00 ©2006 IEEE.
- [12] International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), "Biometrics in Secure e-Transaction", Ms. Swati S Bobde1, Prof. D. N. Satange2, Volume 2, Issue 2, March – April 2013.

# Review on machine learning

**Baljinder Kaur<sup>#1</sup>, Dr. Himanshu Aggarwal<sup>\*2</sup>**

**Baljinderk2221@gmail.com**

**Post Graduation Student**

**Department of computer engineering,Punjabi University Patiala, Punjab,India**

**himanshu.pup@gmail.com**

**Professor**

**Department of computer engineering,Punjabi University Patiala, Punjab,India**

***Abstract-*** In this paper an prevalent review of the study of machine learning, its types and techniques are provided. Machine Learning is the computational methods which are studied for improving performance by mechanizing, and the acquisition of knowledge from experience and convert it into expertise. In the field of Machine Learning one considers the important question that how we can make machines able to “learn”. The paper is describing the classification, regression and their techniques of supervised learning and overview on unsupervised learning. In this paper we will review some techniques of supervised learning like k-nearest neighbor classification, decision trees, Neural networks.

**Keywords:** Machine learning, regression, supervised learning, unsupervised learning.

## I. INTRODUCTION

The Machine Learning field get from the broad field of Artificial Intelligence, which aims to copy the intelligent abilities of human by machines. The impecunious performance results generated by statistical estimation models have drown the estimation area for over the last decade. The aim of machine learning is to provide more efficient and effective knowledge engineering processes to replace much time consuming human activities [3].

Machine learning has set of methods which can detect pattern in data, create new pattern, and can estimate future data. There are mainly two types of machine learning supervised and unsupervised. Supervised learning use labeled data and in unsupervised learning no labeling for data is used, and it also create its own interesting patterns. Machine learning has vast area, that's why we need much knowledge to develop machine learning applications successfully [1].

There are many applications of machine learning such that data mining, neural network, robot locomotion, anomaly detection, software engineering, stock market, search engines, online advertising, game playing, Text classification and categorization, network intrusion detection, Bioinformatics, recognition of hand writing and speech, brain computer interfacing, monitoring of electric appliances, drug discovery etc [6].

## II. LITERATURE SURVEY

This paper provides a broad review on estimating software development process using machine learning techniques. Machine learning is new phase, which is trying to provide correct estimates regularly. Machine learning system excellently “learns” how can we take estimates from training set of previous projects. The main motive of the review is to base the research on expert estimation and to provide other researchers satisfactory information about machine learning techniques. Techniques used in this paper are classification and regression tree, neural network, CBR expert estimations for software development [6].

Supervised classification is the task which is mostly and easily implemented by Intelligent systems. Many techniques have been established based on the Artificial Intelligence and Statistics, like Logic-based techniques, Perceptron-based techniques, Bayesian Networks, Instance-based techniques etc. The aim of supervised learning is classification (provide discrete output) and regression (provide continues output) [7].

In this paper comparison is given between unsupervised and supervised learning models, and their pattern classification evaluations according to which they are applied to the higher education scenario. Classification has important role in machine based learning algorithms and in whole Artificial Intelligence system. The error back-propagation learning algorithm which is accommodated by the supervised learning is good for many non linear real-time problems and in unsupervised learning model, KSOM provide powerful results and classification at this moment [2].

## III. TYPES OF MACHINE LEARNING

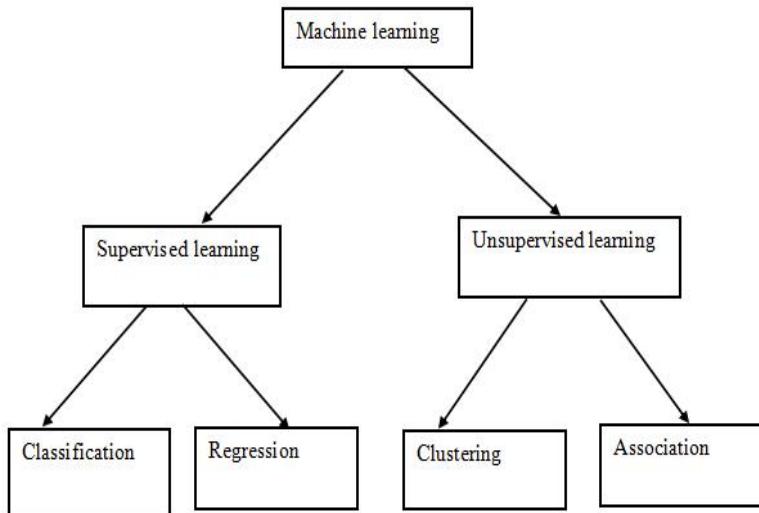


Figure 1 Types of machine learning[4]

## 1. SUPERVISED LEARNING

Supervised learning approach, give us mapping from input(x) to output(y). In supervised learning there are labels associated with each input-output pair  $D = \{(x_i, y_i)\}_{Ni=1}^N$ .  $D$  = training set, and  $N$  = number of training examples. Input( $x_i$ ) may be a complicated structured object like an image, an email message, a graph, a sentence, a time series and any other [2].

Labels are associated with each example in supervised learning. Label is assumed as the answer of the question for the given example. Further two categories of supervised learning are classification and regression, if the label is distinct the problem is called as classification problem and if the label is real value then it is known as regression problem. With the help of these examples, anyone is usually excited to figure out the answers of unseen problems, which are not still observed. Thus learning is not only a way of memorizing, it is also observation to known problems [1].

## 2. UNSUPERVISED LEARNING

Unsupervised learning is second type of machine learning. It is also called descriptive learning. In unsupervised learning we give inputs only,  $D = \{x_i\}_{Ni=1}^N$ , there is no labeling in unsupervised learning. We need to find “interesting patterns” in the given data. Unsupervised learning is also known as knowledge discovery. In unsupervised learning we try to find anomalies and hidden regularities in data [2].

*Classification:* Now we will discuss classification. Our aim is to study a mapping from input(x) to output(y), in which  $y \in \{1, \dots, C\}$ ,  $C$  is the number of classes. When  $C = 2$  the classification is known as binary classification. If  $C > 2$  then it is known as multiclass classification. Where the term “classification” is used, that mean multiclass classification with a single output, if not we estate else ways. Function approximation is the approach to explain the problem. Suppose  $y = f(x)$ , for new function  $f$ , and our aim of study is to measure the function  $f$  for given a labeled training set, after that form prediction. For this our main motive is to do predictions on strange inputs(means that we have not seen that problems previously), it is also known as Generalization. In that way predicting the response on training set is easy [5].

## IV. SUPERVISED LEARNING TECHNIQUES

*k-Nearest Neighbor Classification:* k-Nearest classification is an uncomplicated method. In this techniques k points are found for training set, which are closest for test point and after that label is assigned to the test point from k points. This method is popular for its low classification errors and its simplicity but on other hand it is expensive and need large memory to store the training data [4].

*Decision Trees:* Decision tree gives clear-cut true/false queries about the training document in the tree structure form. Where leaves present image of identical category of text document and branches presents the image of association of features which spot to categories. Decision trees can applied to any type of data. This technique is

also applicable when data is massive and have large attributes. The main risk is to apply this technique is, that it gives more alternative trees for training data [3].

*Neural Network:* Neural network is a part of artificial intelligence. It is based on human/animal neurons, however in the brain there are millions of neurons interconnected. In the neural network there are large number of computing devices, they form a complex network for make communications. There are many application of Neural Network like robotic, data mining, geology, engineering etc. It is also represents data in the form of graph in which nodes represents the neurons and edges represents the links between the nodes. With each input, some weights are associated [6].

It create a feed forward network which does not create cycles. There are three main types of layer in this network input layer, hidden layer and output layer. The output layer always gives a single output. In feed forward graph the flow of data is always in one direction, from input to hidden layers and to output layer.

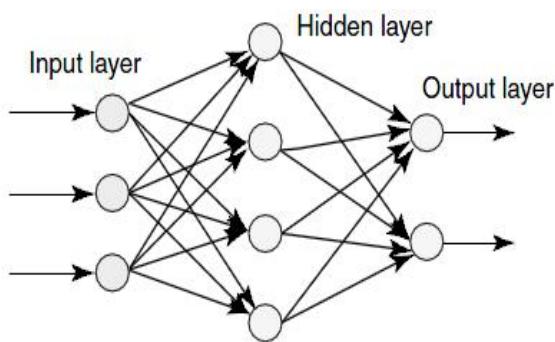


Figure 2 Feedforward network.[6]

## V. CONCLUSION

This paper is describing the study of machine learning and types of machine learning. The prime purpose of this paper is to provide some basic knowledge about, how machines are helping us to solve complex problems. In Machine learning, knowledge is extracted from examples or input, which are given by the experts. Supervised learning is most widely used now a days, for classification or pattern recognition. On the other hand, unsupervised learning gives interesting pattern of the data/input. This is also called knowledge discovery. In this paper, techniques of supervised learning are also reviewed and described briefly. In future, we will study about algorithms associated with these techniques and some more concepts of machine learning. Classification and regression are already described in this paper, we will focus on clustering and association of unsupervised learning. This is vast area of Artificial Intelligence, so we will try to cover as many as topic which are necessary for our work area.

REFERENCES

- [1] Dy.G.J,Brodley.E.C,"Feature Selection for Unsupervised Learning",Journal of Machine Learning Research,pp. 845-849,April 2014.
- [2] R.S,A.A,"Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification"International Journal of Advanced Research,February2013.
- [3] Dharmadhikari.s.c,M.I,P.K,"Empirical Studies on Machine Learning Based text Classification Algorithms",International Journal,November 2011.
- [4] G.Ratsch,"A Brief Introduction into Machine Learning",Friedrich Miescher Laboratory of the Max Planck Society,April 2011.
- [5] Murphy.k.p,"Machine Learning A Probabilistic Perspective",The MIT Press Cambridge,May 2013.
- [6] Y.S, Bhatia.P.k,O.S,"A Review of Studies on Machine Learning Techniques",pp.70-84,April 2012.
- [7] Kotsiantis.S.B,Zaharakis.I.D,Pintelas.P.E."Machine Learning: a review of classification and combining techniques",Springer Science Business Media B.V,pp.159-190,November 2007.

# Comprehensive Analysis of Web Usage Mining and its Tools

Anmol Kaur, Dr.Raman Maini

Student, Department of Computer Engineering, Punjabi University, Patiala, Punjab, India  
annolkaur1320@yahoo.com

Professor, Department of Computer Engineering, Punjabi University, Punjab, India  
research\_raman@yahoo.com

**Abstract---**Web Mining can be defined as highly expressive tool which is used to fetch the required information. The distilled data can be used to ameliorate the Web Usage Mining. It helps in the prediction of next accessed by the user, financial data analysis, intrusion detection. There are different steps to clean the data and then it is analyzed through different techniques to find relevant results. To find the results Web SIFT tool is used. Web Usage Mining consists of three parts which Web Usage Mining, Web Content Mining and Web Structure Mining. From the studies it has been analyzed that Web Usage Mining deals with the how to predict the behavior of the users to increase efficiency of the website by collecting the log files from the server. Web Content Mining is used to find out the relevant information from the various documents of the web, while Web Structure Mining comprises of web pages as nodes and hyperlinks as edges.

**Keywords---**Pattern Analysis, Pattern Discovery, Log file, Web Serve, Client, Visualization and Data Preparation.

## I. INTRODUCTION

In today's era of science and technology the World Wide Web has become a necessary medium to disseminate the information each and everywhere. Due to its giant nature its scope is very vast. The wide level has resulted into abundant amount of information is available for all those who use internet for their different purposes [2].

The various techniques are applied on the source to obtain the related information because it consists of structured and unstructured documents. To meet the satisfaction level of the user of the different fields it is necessary to differentiate the web documents in terms of text file, images and multimedia files [2]. There are various ways and tools to extract the knowledgeable data. Every user from the different professions needs efficient tools to find the related data in a very accurate manner. Thus the approach of the web service provider is to predict the different user behavior to minimize the load of the traffic and web site is designed in such a manner that it could be effective for different profession users [12]. For example student needs are regarding their study, business people want data about their market shares and bank manager want the details of the accounts of the different users etc.

## II. WEB MINING CATEGORIES

The Web Mining Categories can be broadly classified into the following manner (A) Web Content Mining (B) Web Structure Mining (C) Web Usage Mining.

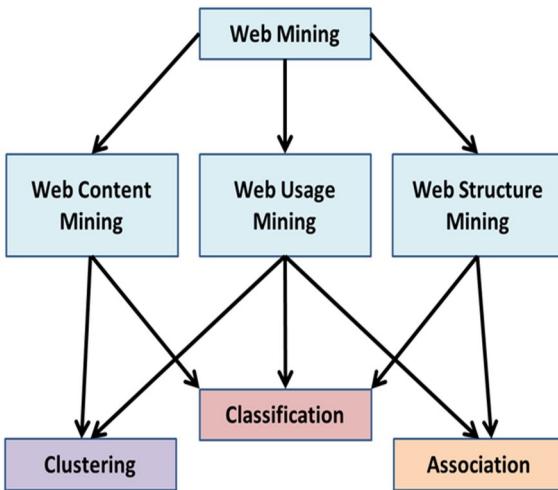


Figure 1.Web Mining Categorization [1]

- A. *Web Content Mining* -- Web Content Mining can be defined as the deduction of knowledgeable data and fruitful information from the various web document sources. To examine the relevant data it is necessary to scan the required the data whether it is text document, video or graph. When the whole clustering is done then it is said that scanning is fulfilled [3]. As large quantity of data is available on the internet, the search engines provide result in order of high preference. This is only possible with the technique of Content Mining. This mining has two different strategies which are Retrieval of information view and Database view. Free texts which are unstructured in nature, HTML documents which are semi-structured all are defined in the Web Content Mining [4].
- A. *Web structure Mining*-- Web structure Mining with identification of the relationship among pages that are linked by data or connection which are direct. There are two problems which are reduced only with the help of structure Mining which are (i) Irrelevant search results because search engines prefer low precision method (ii) incapability of indexing of the huge amount of the data which is available on the world wide web[3]. Structure of the web hyperlink helps to reduce the above problems which are possible with Web structure Mining [11].
- B. *Web Usage Mining* -- Web Usage Mining is helpful in finding the required and relevant data by learning the behavior of user activities called web log. Network logs, server logs and web logs are some of the web logs. These logs help us in deciding what kind of interest an user have. The history of logs is collected by the organizations that who had visit which site and how much time he had spend on that [3] [11].For example if somebody is looking for study related sites he is predicted as good learner, if somebody visits music sites he can be judged as a music lover.

### III. NEED OF WEB USAGE MINING

Web Usage Mining is one of the important categories of the Data Mining process which is helpful in extracting the pattern which is generated through the activities of the various users i.e. client-server interaction [10].

The growth of E-Commerce is rapidly increasing day by day. The trend of doing business and payment of goods is also changing called online payments, transactions and building of the relationship called the client-server relationship. In this work, more stress is given on the various steps which are involved in the Web Usage Mining process to identify the behavior of the user and the patterns involved e.g. dynamic content. The web sites perform mass customization and personalization by discovering the clusters of users with similar access patterns. From the information and generalizations obtained, helps in management of the site in the better way [10] [11].

### IV. DATA SOURCES REQUIRED FOR THE WEB USAGE PROCESS

The applications of Web Usage Mining are formed on the collection of the data from the following three important sources , which are defined as (i)Web Servers (ii)Proxy Servers (iii) Web Clients.

These are explained in brief as under:

- A. *Web Server*-- The abundant of information is collected through the web servers in form of log files. The remote host IP, name, request time, dates all the information which is necessary to collect these log files. There is a standard way to express this whole collected information: Common log Format, Extended log Format, LogML [10].
- B. *Proxy Side*--The enhancement in the speed of the navigation which is done by caching is possible when the Internet Service Provider give their customer the facility of the Proxy.
- C. *Server Client Side*--The use of the Java applets, Java Scripts is done to trace out the data usage on the side of the customer [11].

### V. PREPROCESSING OF THE DATA

In the field of the Web Usage Mining the preprocessing of the data has a sound importance. But, this process is somewhat time consuming. The different steps involve are discussed as under:

- A. *Cleaning of the Data*--In this step filtering is done to eliminate the items of the web log which are not necessary for the purpose of the mining. Most of the times suffixes are removed like “JPEG”, “JPG”, “GIF” etc are removed which define what kind of file this is [11]. These all unwanted files are eliminated to make the process easy to find the relevant results [8].

- B. *The Identification and the Reconstruction of Session*--(i)In this step from the log files which are not enough rich in information ,the identification of the session of different users is done.(ii)after identifying the session navigation path of the user is reconstructed [10] [13].
- C. *Retrieval of the Content and the Structure*--The URL's are considered as the rich source of information to mine the data. URL's are mainly used in the applications of the Web Usage Mining. If the proper classification is judged in pre -steps, the help of Web Structure is taken. On the other hand, many times in search engines the Web Content Mining is also used [8].
- D. *Formatting of the Data*--It is the last step. After the completion of all these steps, the proper format of the data is made to apply the techniques for mining purposes and the data is stored in the repository. For the sequencing of the web pages WAP-trees are considered [10].

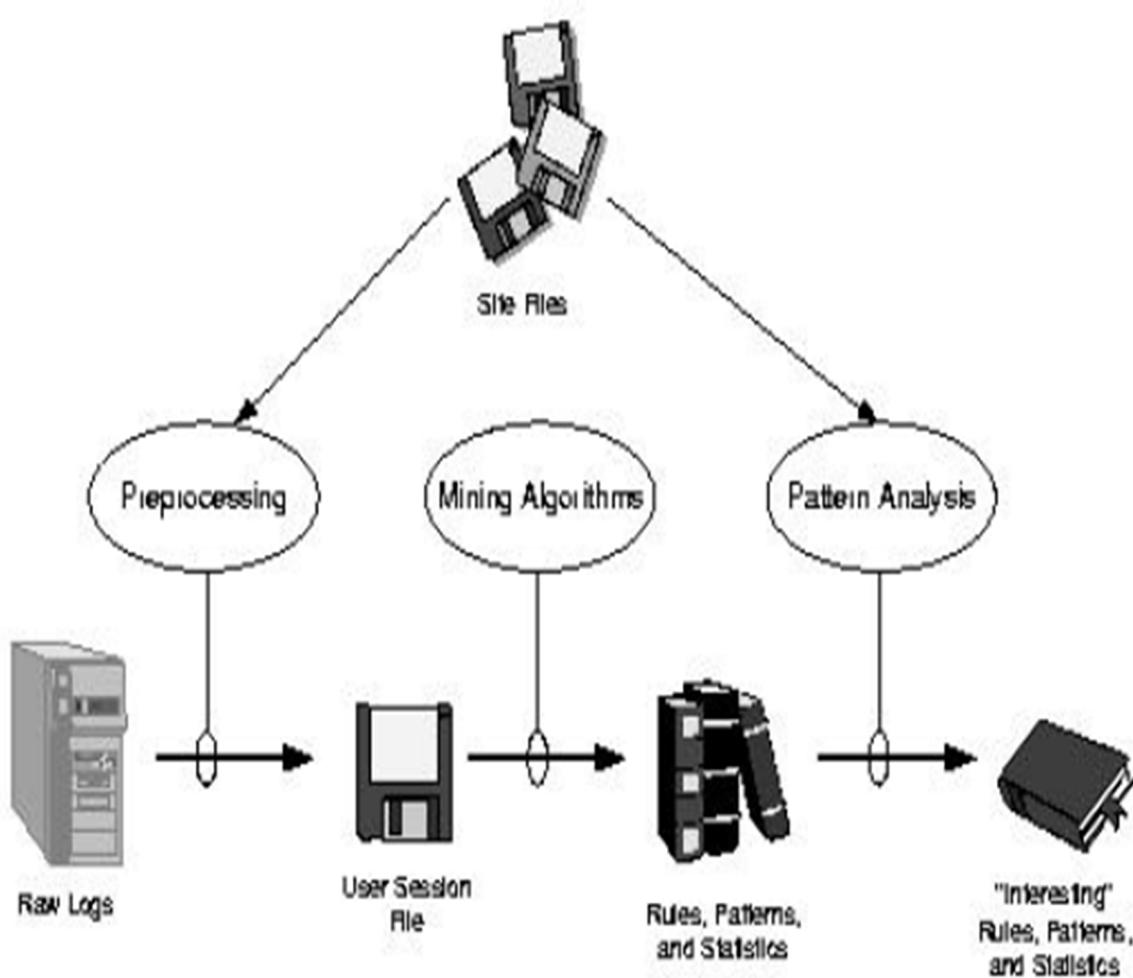


Figure 2.Web Usage Mining Process [1]

## VI. DISCOVERY OF PATTERN TECHNIQUES

It is one of the important parts in the mining applications. It consists of different stages like machine learning, identification of patterns etc. are applied on the information which is available. Here are some techniques which help to find the patterns from the data which is already get processed. Here, brief introduction to these techniques are discussed.

- A. *Association Rule*--This technique of mining is mainly used to find relation which is not in order. Association rules are used to find associations among web pages that frequently appear together in users' sessions. By using the association rule the designers of the web site can enhance their sites [3].
- B. *Clustering*-- It can be defined as way of collecting all the data items which share the common properties. The use of the clustering defines which group of customers has same kind of nature. Clustering technique play major role in the applications of the E-Commerce [3].
- C. *Classification*--The technique of the Classification deals with mapping of the data items into set of classes that are already predefined. Supervised learning which includes K-nearest neighbor, Decision trees, Vector support etc [12] is helpful in generating the results through the use of the classification technique.
- D. *Sequential Patterns*--It is the method defined to generate the inter-session pattern. Temporal analyses which include detection of the changing of the points, trend analysis are also defined under the sequential pattern [9] .The sequential patterns play major role in the prediction of the future trends. These predictions define the placement of the advertisements.

## VII. ANALYSIS OF THE PATTERNS

This is considered as the last step of the Web Usage Mining. After the analysis of the patterns only the useful patterns get extracted and the other remaining ones which are irrelevant get eliminated by using pattern analysis method. There are generally two methods for the analysis of the pattern. The first method is the SQL query concepts and the second method is before applying the operations of OLAP, the multidimensional data cube should be build. New emerging technique like visualization is also in use [11]. This is also fertile area of research. Yet there are many applications, which are used for the commercial analysis but these not much liked by the users due to inflexibility and slow nature. For the efficient, suitable, highly adaptable and flexible tools to be developed are in great demand.

## VIII. WEB USAGE MINING TOOLS

There are some various tools which are used for the Web Usage Mining.

A. *Web Site Information Filter System (Web SIFT)*--The three categories of domain information named content, structure and usage which are used for the web usage mining process. To find the relevant results, content and structure data from the website are used by the Web SIFT. Web Miner prototype is the foundation of the Web SIFT system. The Web Miner prototype is divided into three parts named preprocessing of the data, discovery of the pattern and the analysis of the pattern. Java, Procedural SQL and relational database are used to implement the Web SIFT [3].

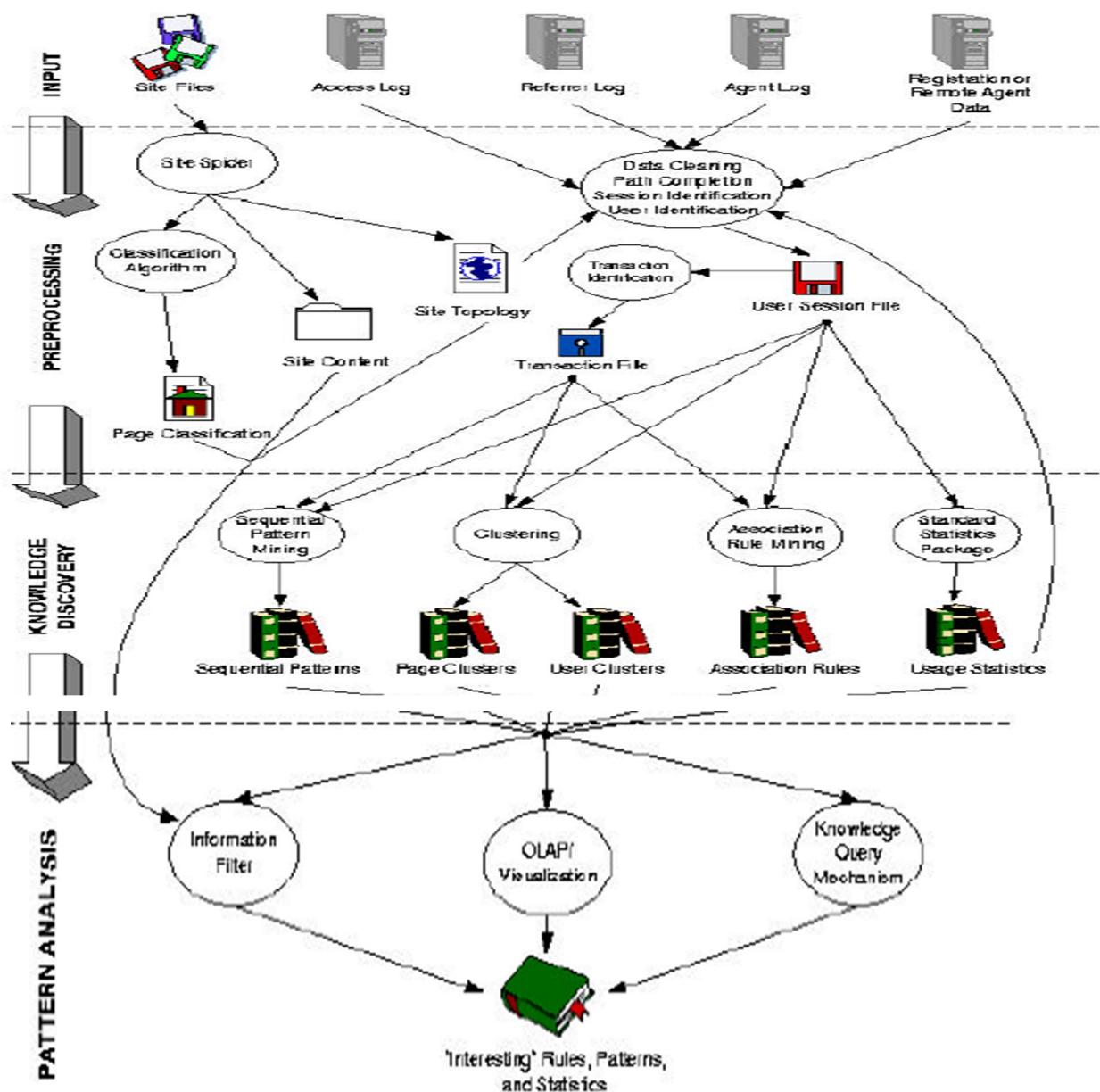


Figure 3. Architecture of the Web SIFTS [3].

- B. *KOINOTITES:Used for the personalization--*The role of this technique is to build the communities of the user on the world web which show the same kind of interests. Adaption according to the needs of every user makes the web more user friendly. Through the collection of the log files KOINOTITES predict the behavior of the different users and organazises it into more knowledgeable form [8]. The major components of KONOTITES (i)*Mining component*—It contains the konwledge about the preprocessing of the data,identification of the sessions ,representation and the knowledge about the various pattern of the different users.(ii)*Graphical User Interface*—This is helpful in the interaction between the user and the system.The Java progmming language is used to implement these components [3].
- C. *Web Utilization Miner(WUM)--*This is also used for the recognition of the patterns with the help of using the MINT language called the mining language. The textual, structural and statistical behavior is also supported by MINT.The proper representation and it's knowledge about the recognised pattern helps the developer to increase the quality of the site. Only the interested patterns are exracted and the other ones are ignored in the sequential steps which are used in the mining process. There are two types of this (i) The MINT Processor mines the whole information in consonance with those who have proficiency in the language.(ii) The aggregation Service mines the data by continuosly analysing the log activities of the users [3].

## IX. CONCLUSION

In this work, an overview of the Web Usage Mining is done. The description about that how to prune the useful information after collecting the data from the various web log files. It has been analysed that data is cleaned to eliminate the unuseful items. Identification of the session of different users is done.Onlythe useful patterns get extacted. Patterns are discovered through techniques like Association which is used to find the relation which amond web pages. Classification define the classes are predefined. Decision trees, K-nearest neighbor are helpful in generating the patterns.Different tools which are used for the mining purposes such as Web SIFT which helps to find the structured data and KOINOTITES is used to build communities on the web is discussed. As the demands of the users are changing day by day, the trend of the new techniques are emering to improve the design of the websites .

## REFERENCES

- [1] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data",*SIGKDD Explorations volume-1, Issue-2*, Jan 2000.
- [2] Raymond Kosala and Hendrik Blockeel, " Web Mining Research: A Survey",*ACM SIGKDD*, July 2000
- [3] Yan Wang," Web Mining and Knowledge Discovery of Usage Patterns",*CS 748T Project (Part I)*, February, 2000.
- [4] R. Cooley, B. Mobasher, and J.Srivastava," Web Mining: Information and Pattern Discovery on the World Wide Web",*Ninth IEEE International Conference, IEEE*, Nov 1997.
- [5] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava , "Data Preparation for Mining World Wide Web Browsing Patterns",*Supported by NSF Grant*, Oct 1998.
- [6] R. Kosala, H. Blockeel,"Web Mining Research: A Survey", in *SIGKDD Explorations 2(1)*, ACM, July 2000.
- [7] B. Masand, M. Spiliopoulou, J. Srivastava, O. Zaiane, ed. Proceedings of "WebKDD2002 –Web Mining for Usage Patterns and User Profiles", Edmonton, CA, 2002.

- [8] R. Kohavi, “Mining E-Commerce Data: The Good, the Bad, the Ugly”, Invited Industrial presentation at the ACM SIGKDD Conference, San Francisco, CA, 2001.
- [9] M. Spiliopoulou, “Data Mining for the Web”, Proceedings of the Symposium on Principles of Knowledge Discovery in Databases (PKDD), 1999.
- [10] Shahnaz Parvin Nina, Md. Mahamudur Rahaman, Md. Khairul Islam Bhuiyan, Khandakar Entenam Unayes Ahmed, “Pattern Discovery of Web Usage Mining” International Conference on Computer Technology and Development,2009.
- [11] Anitha Talakkula, “A Survey on Web Usage Mining, Applications”, Computer Engineering and Intelligent Systems ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online) Vol.6, No.2, 2015
- [12] R.Kaur, S.Kaur, A.Kaur, R.Kaur, A.Kaur, “An Overview of Database management System, Data warehousing and Data Mining”. IJARCCE, Vol.2, issue.7, July 2013.
- [13] K. Maheshwar and D.Singh, “A Review of Data Mining based instruction detection techniques”. International Journal of Application or Innovation in Engineering & Management.Vol.2, Issue 2, Feb.2013.
- [14] A. N. Mahanta “WebMining:Application of data mining” pp. 111-116 Proceedings of NCKM-2008.
- [15] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, and A.S.K.Ratnam, “A Study of Data Mining Tools in Knowledge Discovery Process”. International Journal of Soft Computing and Engineering, Vol.2 ,Issue 3, July.2012.
- [16] Sawan Bhawsar, Kshitij Pathak, Sourabh Mariya and SunilParihar, “Extraction of Business Rules from Web logs to Improve Web Usage Mining” International Journal of EmergingTechnology and Advanced Engineering ,ISSN 2250-2459,Volume 2, Issue 8, August 2012.
- [17] M. Rao, M. Kumari, and K. Raju, “Understanding UserBehavior using Web Usage Mining”, International Journal of Computer Applications, 1(7), 2010, 55–61.
- [18] Sawan Bhawsar, Kshitij Pathak, Sourabh Mariya and SunilParihar, “Extraction of Business Rules from Web logs to Improve Web Usage Mining”, International Journal of EmergingTechnology and Advanced Engineering ,ISSN 2250-2459,Volume 2, Issue 8, August 2012.

# IMPLEMENTATION OF BLACK HOLE NODE DETECTION USING INTRUSION DETECTION SYSTEM

Gurpreet Singh<sup>1</sup>, Dr. Raman Maini<sup>2</sup>

<sup>1</sup>*Department of Computer, Punjabi University Patiala, Punjab, India*

<sup>2</sup>*Professor Punjabi University Patiala, Punjab, India<sup>2</sup>*

**Abstract**—Mobile ad-hoc networks are known by their continuously changing topology, no permanent infrastructure, resource restrictions and multi-hop scenario. Owing to this, they are more exposed to security attacks. In these networks, one of the hazardous attacks is packet dropping attack differentiated into 2 types: 1. Black hole and 2. Gray hole. Within both of them, an attacker replies to source node incorrectly that it is encompassing the shortest path to the destination. In former, attackers drops the entire volume of packets sent through it and in latter, attacker drops packets selectively. The following study implement to detect and isolate these attacks in network. The performance of the network has been compared without black hole node and with black hole node on the basis of PDR, throughput and energy consumption of the network. It has been observed that the actual performance of the network without black hole node is better as compare with the network with black hole node.

**Keywords**— Mobile ad hoc networks, black hole attack, grey hole attack, throughput.

## I. INTRODUCTION

A wireless network which is managed by itself and possessing the potential of real time network is referred to as MANETs. In these networks, the nodes can exchange information with each other exclusive of the centralized authority. On account of their features for instance wireless links, dynamicity and distributedness, these are exposed to numerous security attacks which can be Wormhole attack, Packet Dropping Attack, Black hole attack, Gray hole attack, Flooding attack, jellyfish attack, Sybil attack etc. [12] MANET can alter the network path due to congestion'. There is a range of security Attacks in Mobile Ad hoc networks. One of them is the Attack named as Packet Drop attack which is occurring into the Network Layer.

MANET is prone to a variety of kinds of attacks described earlier, all of them are considered as denial of service attacks [13]. In the black hole attack, a compromised node drops all packets sent through it by faking to sender that it possesses a valid shortest path. Eventually receiver node never gets any packet from the sender node. For this reason, the performance of the system is hampered.

This paper presents a brief study about the past work that has been to identify and avoid black hole attack in networks. Then the proposed scheme has been described and the results have been presented in this paper.

## II. RELATED WORK

IN [1] THE AUTHORS ILLUSTRATE THE FEATURES, FUNCTIONS, AND VULNERABILITIES OF MOBILE AD HOC NETWORK. THE STUDY ALSO PRESENTS A GENERAL IDEA OF THE ATTACKS AND THEIR ALLEVIATION ORDETECTION IN ROUTING PROTOCOLS. DUE TO INCREASE USE OF WIRELESS NETWORKS, MOBILE PHONES, SMART DEVICES ARE AMASSING ATTRACTIVENESS MAKING THE AD HOC NETWORK A RISING FIELD. EACH UNIT OR NODE OR HUB IN A MANET IS OPEN TO MOVE AUTONOMOUSLY IN ANY ROUTE OR DIRECTION THUS CONNECTING TO OTHER UNITS REGULARLY. EVERY ENTITY MUST FORWARD PACKETS DISTINCTLY, AND FOR THAT REASON IT PLAYS ROLE OF A ROUTER. THE DESIGNED ROUTING PROTOCOL MUST BE ABLE TO HANDLE WITH THE NEW CHALLENGES CREATED FOR INSTANCE NODES MOBILITY, SECURITY SAFEGUARDING, AND QOS, INADEQUATE BANDWIDTH AND RESTRICTED POWER SUPPLY ETC.

The authors in [2] presents introduction about MANET and brief description of attacks in MANET. With the augmentation in application of MANETS, safety has become a critical necessity to make available protected communication among mobile nodes. The researchers in this illustrate a technique to identify and avoid black hole attacks by warning other nodes. To defeat the challenges, there is a necessity to make a security solution that accomplishes both protection and wanted network performance. It has been explained about ANT NET, where ACO system and pseudo code of it has been proposed.

In [3] the authors put forward various promising answers to trouble caused by Black hole attack. The major test in creating a MANET is preparing each device to constantly keep up with the information compulsory to correctly forward traffic. AODV is a reactive routing protocol for such networks. It is intended to sustain in an surroundings of mobile nodes, coping up with a range of network behaviors for instance mobility of the nodes, link failures etc. Black hole attack is a sort of DoS attack wherein a router that is believed to relay packets rejects them. This more often than not happens from router acting illegitimate from a variety of unusual reasons. One of this declared in research, is with the support of denial-of-service attack. Since packets are regularly dropped from a network, such kind of attack is very tough to distinguish and put off.

In [4] the previous ways to black hole attacks on AODV protocol are described. The paper states that the crucial objective of the safety solutions for system is to give services, for instance verification, privacy, reliability, secrecy, and accessibility, to mobile users. Black hole attack is the brutal safety threats in wireless ad-hoc networks that can be effortlessly used by exploiting weakness of on- demand routing protocols namely DSR or AODV.

In [5], the authors have detected the malicious black hole node present in the system by intrusion detection system using fuzzy logic. The authors have combined the fuzzy rules in AODV routing protocol where they have used four modules namely fuzzy parameter mining, fuzzy calculation, fuzzy authentication part and alarm packet creation part. The trustworthiness level of the node in the network is calculated on account of the historic patterns observed in the network Further the authors have used the concept of threshold value to verify the trustworthiness of the node. The node is assumed to be faithful only if its trustworthiness factor surpasses the threshold value. The broadcasting of the alarm packet by the nodes is used to prevent the communication with black hole node in the system.

In [6], the authors have implemented their algorithm in AODV with the intention to identify and avoid the black hole nodes present in the network. The approach is found useful to detect and prevent both forms of the black hole attack namely single black hole attack where barely individual node launches the attack in the system cooperative

black hole attack where the attack is done by extra black hole node in the system. They have used Fibonacci series pattern to broadcast the RREQ packets in the system and then in order to detect the malicious node, the communication interval is weighed against the threshold value. If the value goes beyond the threshold, at that moment the malicious node is detected by comparing the D\_Seqno with the received sequence number. Their algorithm is also able to detect the node even if it is in idle state.

In [7], the authors have designed and intrusion detection system to protect the network from black hole nodes present in the system. Since black hole node replies to the source node with higher seq. no., so the intrusion detection system checks the updates made in the routing table by the nodes. The node which has updated the routing table and sent higher seq. no. is considered to be malicious node. The path where the malicious node is present is rejected by the source node and the source node then looks for new path to the destination.

In [8], the authors have evaluated the performance of two ad hoc routing protocols under black hole attack, namely AODV routing protocol and AOMDV routing protocol. The authors have utilized the advantage of AOMDV which forms multiple paths, if black hole attack occurs in single path then data from source to destination can be sent using some other available path. Due to this multipath option, AOMDV showed better performance against AODV routing protocol in terms of both PDR and number of packet drops.

In [9], the authors have focused on designing IDS for mobile networks which also reduces overhead in the network. For this, they have proposed the concept of mobile agent which calculates the amount of packets forwarded by the node and amount of packets received by the node. This ratio is called to be confidence ratio. The specialized mobile agent calculates this ratio for all the intermediate nodes present in the route from source to destination. If the confidence ratio for any node is not as much of the predefined threshold value, in that case that node is suspected to be malicious node in the network.

In [10], with the purpose of detecting the black hole node the authors have modified the existing DSR routing protocol. During the routing phase the source node sends the RREQ in the network so as to find the shortest route to the destination. The authors have used the fact to distinguish the malicious node, that whenever any node in the network replies back to the source node more than one times alleging that it has shortest route to the destination, then that node is suspected to be misbehaving node in the system. After finding the malicious node, the source node sends out a packet containing information about the malicious node so that other nodes do not communicate with the malicious node.

### III. BLACK HOLE ATTACK

In AODV for the duration of the path discovery stage when a node acquires a Route Request (RREQ) message, if it has route to destination then it responds to the sender node by means of the route reply (RREP) message. In this it contains the seq. no. to the destination. On account of RREP message the sender decides which one will be the most excellent path for transferring the data to the destination node. [11]

Black hole node i.e. the malicious node forwards a false RREP packet to a source node with the purpose of pretending itself containing a path to destination node. When a source node receives multiple RREP, it finds the route reply with greatest sequence number as considers it as the most new routing information and chooses the path included in that Route Reply packet. When the Seq. No. are equivalent it picks the path possessing lowest hop count.

In black hole attack, RREP is sent with D\_seq\_no. bigger than the genuine destination node to the source node, so that the data transfer will be sent through the intruder. Along these lines, upon receiving the data, it drops the packet and hampers the communication among source and destination node. [11]

#### IV. IMPLEMENTATION OF BLACK HOLE NODE DETECTION USING IDS

The intention of this scheme is to detect and avert the malicious black hole node in mobile ad hoc network. Once the source node has to transmit data to target node, it begins with sending the route request packets to destination. And upon accepting the route request packets the destination replies with route reply messages and a path to send data is created. In this scheme the IDS works in the following manner:

First the performance of the network is analyzed in the normal scenario, i.e., in the absence of any malicious node. After that if any malicious node has entered the network with the aim to drop the packets, and then the anomaly occurs. For example previously if the destination was receiving packets in a normal way, then after the entrance of malicious node the destination would not receive same amount of packets. IDS detect the nodes which causes such anomalies in the network. As a prevention step the source node discards the path having presence of the malicious node and chooses another path to send the data to destination node.

#### V. RESULTS AND DISCUSSION

This scheme has been implemented on NS2.35 which is open source simulator. The performance of the system has been analysed on the basis of PDR, throughput and energy consumption. In order to simulate the network the input parameters has been described in the table 1 below:

Simulator	NS2.35
Channel	Wireless Channel
Propagation Model	Two Ray Ground
Queue	Drop Tail
Antenna	Omni-Directional
Energy Model	Radio Energy Model
Simulation Area	1200*1200
No of nodes	50
Routing Protocol	AODV
Initial Energy	100 Joules

Table1: Simulation Parameters

Throughput means quantity of data that is received at the destination node. It is the primary factor that reflects the performance of the network. The value of this factor shows an increase till 23 sec of the simulation time. Then the throughput becomes constant indicating there is no data received at the destination during that period of time indicating attack. After detecting this anomaly the throughput again achieves its normal upward trend. The value of throughput achieved is approx 55 Kbps.

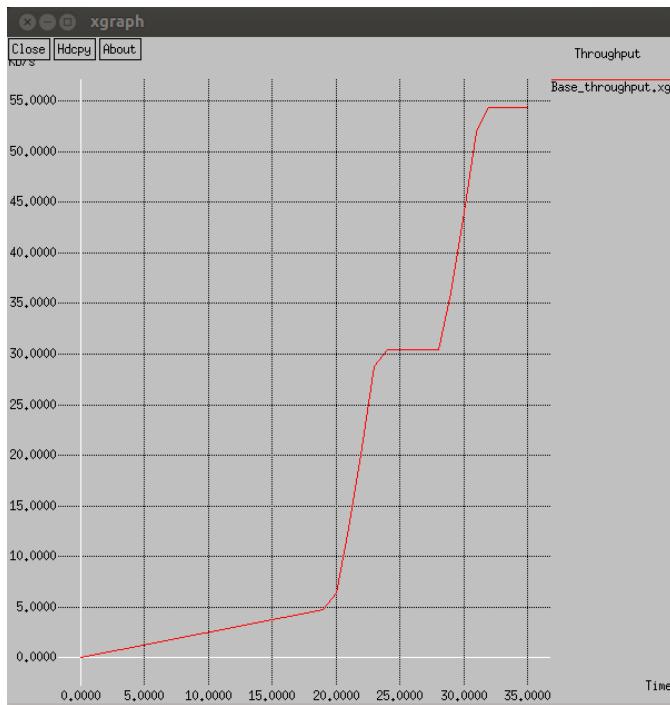


Fig. 1. Throughput

Packet delivery ratio is another important factor which is important to the performance of the network. It means the ratio among numbers of packets received to the amount of packets sent. Initially the graph shows a small drop that may result out of packets being dropped in the network during congestion occurring in the route request phase. Then larger drop in its value indicates attack where a large amount of packets are being dropped by malicious node. Then the parameter regains its normal value when the black hole has been detected and prevented successfully.

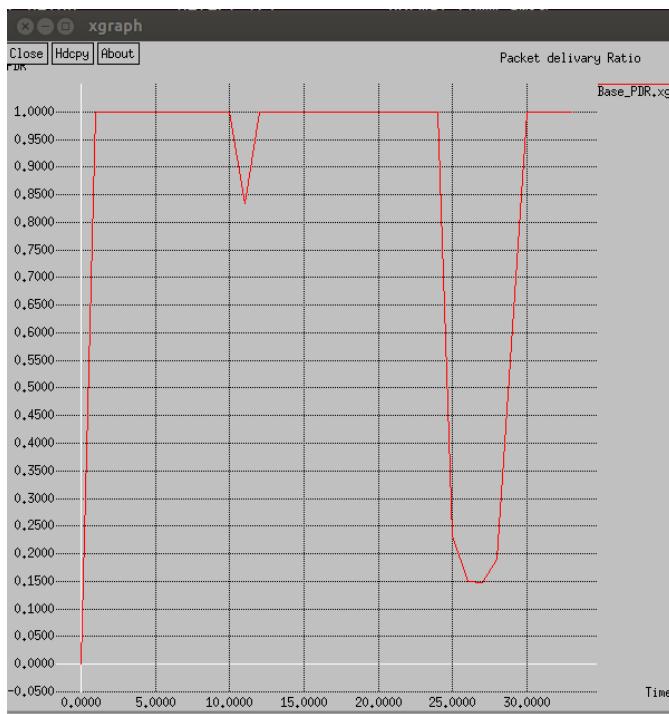


Fig.2 Packet Delivery Ratio

Energy consumption reflects the lifetime of the network. Smaller the energy consumed superior is the lifetime of the network. At first 100 joules of energy was given to the nodes, and remaining energy is 72 Joules.

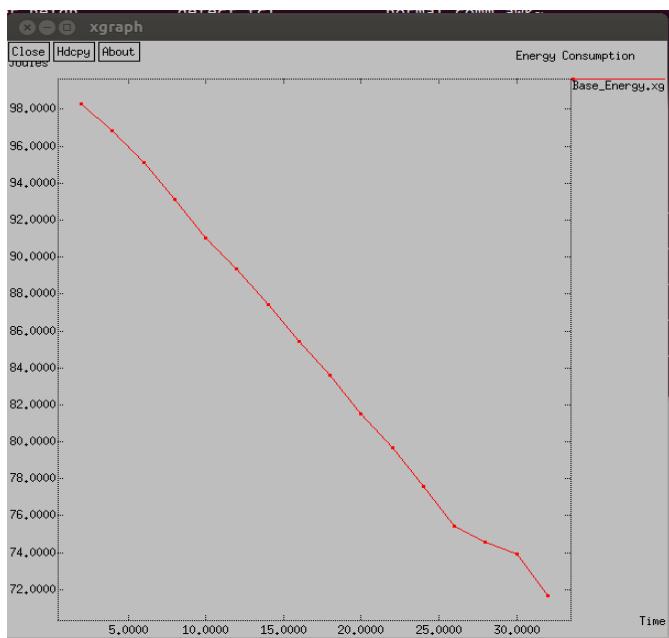


Fig 3. Energy Consumption

## VI. CONCLUSION & FUTURE WORK

In mobile ad hoc networks Black hole attack present a serious challenge as these attacks lead to dropping of data packets. The information which needs to be sent to destination is completely lost if network is prone to black hole nodes. In this paper, the malicious black hole nodes have been detected and prevented using IDS based on the anomalies occurring in the network. The scheme has been implemented in NS2.35. The performance of the system has been analyzed on account of throughput and PDR. This scheme has been successfully able to achieve its objective of detection and avoidance of such malicious nodes from the network. Also the energy consumption has been analyzed for the network. Since energy consumption is another issue in MANETs, in future this scheme can be combined with energy efficient algorithms so that network can have more lifetimes along with security.

## ACKNOWLEDGMENT

The paper has been written with the kind guidance, assistance and active support of computer engineering department who have always helped me in all required places. I would like to thank all ones whose encouragement and support has made the completion of this work possible.

## REFERENCES

- [1] Swati Jain, Naveen Hemrajani, "Detection and Mitigation Techniques of Black Hole Attack in MANET: An Overview", International Journal of Science and Research, 2(5), 70 - 73. (2013)
- [2] Kinagalapavani, Dr. DamodaramAvula, "Injection of attacks in MANET", IOSR Journal of Computer Engineering (IOSRJCE), Volume 4, Issue 3, Sep-Oct. 2012.
- [3] Sowmya K.S, Rakesh T. and Deepthi P Hudedagaddi, "Detection and Prevention of Black hole Attack in MANET Using ACO", International Journal of Computer Science and Network Security, Volume 12 No.5, May 2012
- [4] Rajib Das, Dr. BipulSyamPurkayastha, Dr. Prodipto Das, "Security Measures for Black Hole Attack in MANET: An Approach", International Journal of Engineering Science and Technology (IJEST), vol.3 No.4 Apr 2011.
- [5] Kulbhushan, Jagpreet Singh, "Fuzzy Logic Based Intrusion Detection System against Black hole Attack on AODV in MANET", International Journal of Computer Applications (IJCA), NSC No. 2 Dec 2011.
- [6] NeelamKhemariya, Ajay Khuntetha, "An Efficient Algorithm for Detection of Black hole Attack in AODV based MANETs", International Journal of Computer Applications (IJCA), Volume 66, No. 18, March 2013
- [7] Abhilasha Sharma, Rajdeep Singh, Ghanshyam Pandey, "Detection and Prevention from Black Hole Attack in AODV Protocol for MANET", International Journal of Computer Applications (IJCA), Volume 50, No. 5, July 2012
- [8] D.Geetha, B. Revathi, "AOMDV Routing based Enhanced Security for Black Hole Attack in MANETs", International Journal of Computer Applications (IJCA), ICRTCT, No. 2, Feb 2013
- [9] Debdutta Barman Roy, RituparnaChaki, "Baids: Detection of Black hole Attack in MANETs by Specialized Mobile Agent", International Journal of Computer Applications (IJCA), Volume 40, No. 13, Feb 2012

- [10] Parita Jain, Puneet Kumar Aggarwal, "Preventing MANETs from Black hole Attack using Black hole Node Detection Monitoring System", International Journal of Computer Applications (IJCA), Volume 44, No. 15, April 2012
- [11] Ms. Nidhi Sharma, "Black Hole Node Attack in Manet", 2012 second international conference on Advanced Computing & Communication Technologies, pp. 546-550.[12] Aarti, "Study of MANET: Characteristics, Challenges, Application and Security Attacks", International Journal of Advanced Research in Computer Science and Software Engineering, volume-3, Issue 5, May 2013, pp.252-257.
- [13] Wenjia Li and Anupam Joshi, "Security Issues in Mobile Ad Hoc Network- A Survey", Department of Computer Science and Electrical Engineering ,University of Maryland, Baltimore County Available [http://www.csee.umbc.edu/~wenjia1/699\\_report.pdf](http://www.csee.umbc.edu/~wenjia1/699_report.pdf)
- [14] Amin Mohebi, Ehsan Kamal, Simon Scot, "Simulation and Analysis of AODV and DSR Routing Protocol under Black Hole Attack", I.J. Modern Education and Computer Science", Vol.10, 2013.
- [15] Jaspal Kumar, Kulkarni, Daya Gupta, "Effect of Black Hole Attack on MANET Routing Protocols", International Journal of Computer Network and Information Security", Volume 5, pp. 64 - 72, 2013.

# Review of Performance Analysis of Routing Protocols for MANET

Harsimran Kaur<sup>#1</sup>, Jasvir Singh<sup>\*2</sup>

#harsimran.pamar@gmail.com

Post Graduation Student,

Department of Computer Engineering, Punjabi University, Patiala, Punjab, India

\*jassicct@gmail.com

Assistant Professor,

Department of Computer Engineering, Punjabi University, Patiala, Punjab, India

**Abstract-** MANET network is a collection of wireless mobile nodes that forms a network without any centralized control or access point. MANET stands for mobile ad hoc network. It is robust infrastructure less wireless network. MANET can be created by both fixed and any mobile nodes. In this paper we review the performance analysis of routing protocols for MANET like proactive, reactive and hybrid protocol.

**Keywords:** routing protocols, proactive, reactive and hybrid protocol, MANET.

## I. INTRODUCTION

Mobile ad hoc network is a smallest wireless network that creates a network beyond one access point. MANET stands for mobile ad hoc network. They composed of both router and hosts. MANET is used for different purpose for example, the military, for transmitting any data or information like audio video or any information from one node to another in the network. Routing is serving as a transferring of any knowledge from source node to a destination node. The main objective of ad hoc routing protocols is defining how to packets distribute between nodes without any access point. Routing protocols use different metrics to judge the perfect shortcut for routing the information to its final destination point. These metrics are a popular judgment that could be number of hops, routing algorithm used it to find out the way for the data to its final destination. The method of find out the way is that, routing algorithms keep out routing tables, that's include the whole knowledge for the packet. This way knowledge varies from one route to another in a network. Routing is further defined into three major categories:

- Proactive or Table driven protocols
- Reactive or on demand routing protocols
- Hybrid routing protocols.

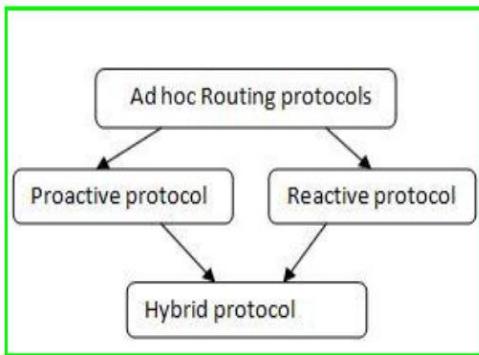


Figure 1

## II. LITERATURE SURVEY

This paper provides an exhaustive survey on routing protocols and also protocols performance. MANET characteristically is an independent system of nodes connected through wireless like without any access point and network may experience rapid and unpredictable topology update. The routing protocols are necessary in order to achieve communication among the whole network [3].

We have reviewed the different routing protocols with its advantage and disadvantages. Metrics have been analyzed that are used in analysis the network performance. MANET network increasing day by day. Routing protocols is wide area to review [4].

In this paper, we have surveyed the classification of routing protocols in MANET and also comparisons are given between routing protocols. The protocols are separated into classification like reactive, proactive and hybrid protocols [1].

### A. PROACTIVE PROTOCOLS:

Proactive Protocols are also called Table Driven Protocols. These protocols keep in proper stable update the topology of network. In this network whole node get idea around the network. The routing information saved into several tables. If any updating exists in network, then the records of tables are also changed according to the modification. The nodes swap network knowledge with one another. It also has route knowledge of whole nodes of network every moment whenever they wanted. Destination Sequenced Vector (DSDV), Optimized Link State Routing (OLSR), Wireless Routing Protocols (WRP) Protocols are examples of Proactive Protocols.

*Destination-Sequenced-Distance-Vector (DSDV):* [1] Perkins and Bhagwant advised the destination sequenced distance vector. It is a smaller amount robust as compared to link state routing protocols. The correspondent node resolves the way from supply node to end node. Routing table regulate each node that defines all node of entries. It

is also defines the list of the entire destinations through a series of number. The series of no. is defined with the destination node of the network topology.

*Pros and Cons:* It is the loop freeways. That network protocol is calculated to infinity trouble is reduced. No any modification still if diffusion of frequency due to useless promotion of routing data.

- *Optimized Link State Routing Protocols (OLSR):* Optimized Link State Protocol advised by Clausen and Jacquet. It is position to position proactive protocol that employs well-organized link state packet and multipoint relaying because it is forwarding method [1]. OLSR process primarily changing and maintaining data in a several tables. The route computation is also determined by the tables.

*Pros and Cons:* decrease the amount of rebroadcast of data packets in broadcast situation. A huge amount of transmission capacity and CPU control is necessary to calculate the way.

- *Wireless Routing Protocols (WRP):* it is advised by Murthy and Garica-Luna-Aceves [1]. Its purpose to maintaining routing information between each and every nodes of topology about the minimum distance to all target nodes.

*Pros and Cons:* in this topology of protocol is a loop routing protocols.

#### B. REACTIVE ROUTING PROTOCOLS:

This protocol also called On Demand Routing Protocols. Routes are defined among the nodes at any time when they are wanted to route the information packets on the network. Source nodes created the route to its destination node that does not have route knowledge. It starts to create the method that's goes through whole node of network before coming to the target destination. Ad Hoc on Demand Distance Vector (AODV), Temporally Ordered Routing Algorithm (TORA), Dynamic Source Routing (DSR) are example of Reactive Routing Protocol.

- *Ad Hoc on Demand Distance Vector (AODV):* This is a mixture of both Dynamic Source Routing and Destination Sequenced Distance Vector Routing Protocols. AODV follows the necessary method for Route detection and Route Maintenance, series of numbers. At any time node requests to send information packet to target node, it initially telecasts the route request (RREQ) packet [2]. The neighbor node telecasts the information packet to their neighbor node and this method continues before coming to the target destination. Throughout forwarding packet demand to neighbor node list the address of neighbor node and this list record is saved in routing tables, which is useful for creation of reverse path.

*Pros and Cons:* Ad-hoc network can handle dynamic behavior of Vehicle. In this algorithm requires for the nodes in telecast can identify others telecasts.

- *Dynamic Source Routing (DSR):* Dynamic Source Routing protocol is depended on the link fixed (static) algorithm. In this routing protocol sender node estimate the route from starting node to its target node. It is

also add the address of neighbor nodes to its list of record in the packet. Dynamic Source Routing was created for multi hop network and also for small Diameters areas [2].

*Pros and Cons:* No necessary to change any nodes time by time. In high mobility is does not perform well.

- *Temporally Ordered Routing Algorithm (TORA):* This protocol is depended on link reversal concept [4]. TORA specially created to localize algorithmic function to topology changes by advance several routes to the end node. Longer routes are mostly used to decrease the overhead of discovery latest routes and shortest hop path gives the extra importance. TORA is considering under the category of stability.

*Pros and Cons:* reestablish the route of high routing overhead. This protocol is defining the changeable invalid route.

### C. HYBRID PROTOCOL:

Hybrid Routing Protocols is mixture of mutually Proactive Protocols and Reactive Protocols. Hybrid protocols is split information of whole network with its neighbor nodes, in this all node have awareness of its next neighbor node. All node have its personal routing zone, zone radius specified the size of route, that's specified by a metric like a number of hops. Zone Routing Protocol (ZRP), Zone Based Hierarchical Link State Routing Protocol (ZHLS) are example of Hybrid Routing Protocols.

- *Zone Routing Protocol (ZRP):* This Routing Protocol is effectively mixture of superlative features of Proactive and Reactive Routing Protocols [3]. Each node specifies the zone radius and the zone around itself is specifies in the number of hops to the perimeter of the zone.

*Pros and Cons:* ZRP decrease power of traffic organized by periodic flooding of routing information packets. Routing zones is overlapping [4].

- *Zone Based Hierarchical Link State Routing Protocol (ZHLS):* Zone Based Hierarchical Link State Routing Protocol is categories into non overlapping zones [3]. Node has knowledge of the connectivity of nodes inside its personal zone. It has knowledge of whole network connectivity zone.

*Pros and Cons:* eliminate the problem of overlapping. Extra traffic defined by the creation and maintenance of the zone level based topology.

### III. CONCLUSION

This paper is describing the overview of Routing protocol for MANET with its types. The main purpose of this paper is to provide some basic knowledge about routing protocol and its purpose. In this paper, different techniques of routing protocol also reviewed and described briefly with its pros and cons. In future, we will study about algorithms associated with these techniques and some more concepts of routing protocol. This is vast area of protocol, so we will try to cover as many as topic which are necessary for our work area.

REFERENCE

1. K.Gupta Anuj, Sadawarti Harsh, and K. Verma Anil, “ Review of Various Routing Protocols for MANETs”, “International Journal of Information and Electronics Engineering”, Vol. 1, No. 3, November 2011.
2. Dr.S.S.Dhenakaran and Parvathavarthini.A, “An Overview of Routing Protocols in Mobile Ad-Hoc Network”, “International Journal of Advanced Research in Computer Science and Software Engineering”, Volume 3, Issue 2, February 2013 pp. 251-259.
3. Paul Hrituparna and Dr. Prodipto Das, “Performance Evaluation of MANET Routing Protocols”, “IJCSI International Journal of Computer Science Issues”, Vol. 9, Issue 4, No 2, July 2012.
4. Parvinder Kaur, Dr. Dalveer Kaur & Dr. Rajiv Mahajan, “The Literature Survey on Manet, Routing Protocols and Metrics”, “Global Journal of Computer Science and Technology: E Network, Web & Security ”, Volume 15 Issue 1 Version 1.0 Year 2015.
5. Muralishankar V.G. and Dr. E. George Dharma Prakash Raj, “Routing Protocols for MANET A Literature Survey”, “V.G.Muralishankar et al, International Journal of Computer Science and Mobile Applications”, Vol.2 Issue. 3, March- 2014, pg. 18-24.

# Data-hiding in 2D Barcode using Steganography

Priya Sidhu<sup>#1</sup>, Er.Gaurav Deep<sup>\*2</sup>

#Department of Computer Engineering, Punjabi University, Patiala  
Patiala, Punjab, India

\*Assistant Professor, Department of Computer Engineering, Punjabi University, Patiala  
Patiala, Punjab, India

**Abstract:** Steganography is a procedure of concealing key messages in a cover media while transmission happens between sender and receiver. Security of secret information has by and large been a vital matter from days passed by occasions to the present time. It is certainly the important subject for scientists to create secure methods to convey information without presenting it to anyone other contrasted with receiver. Every once in a while, scientists have built up many systems to meet secure move of data and steganography is one of them [10]. An extremely powerful process for concealing data behind photographs or whatever other advanced media and to make them better from the interlopers is proposed [9].

**Keywords:** Steganography, Methodology, Barcode, Steganography Techniques

## I. INTRODUCTION

The word Steganography means "included writing" from Greek. In Steganography, it offers secrecy of text or images to avoid them from attackers. Steganography gives secret conversation which supposed hacker unable to get the presence of information or information in it. The idea is that it includes a cover thing that is applied to cover up the first information image, a bunch that's the key image which is to be transmitted [2]. A stego-key that is connected to conceal the data image into cover image, and the steganography algorithm to transport out the vital object. The output which we get is stego-picture that contains the hidden message. At that point the stego image is sent to the device which recovers the data picture by making utilization of the de-steganography [12].

Steganography represents the main position in secret message communication. Different message covering techniques have been developed and implemented in the past applying audio/video files, electronic images, and other medias. An electronic image is a collection of information/data about the pixels in it. It is truly a large number of information. This is large compared to the information we want to hide. Thus we generally choose an electronic image for covert communication. We can make use of their purity for covering information [9].

### A. Process

- Secret Message: The secret message is information which helps to be hidden into some feasible digital media.
- Cover Message: It is the carrier of message such as image, audio, video or other digital media.
- Stego Key: It is utilized to embed message considering upon the encrypting algorithm. Embedding algorithm is the technique for installing the mystery message into the cover image [17].

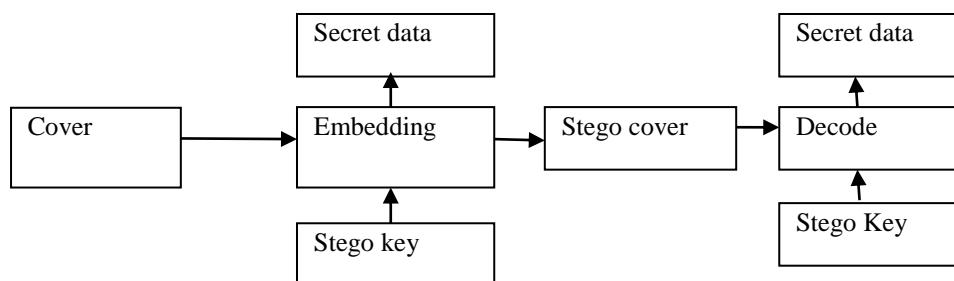


Fig.1 Process of steganography

The general style of Steganography keeps running on the Cover Image (any photo that can be utilized to put on key information inside), the encrypted message (the individual information that will be sent covertly), a stego key that is connected to recorder the mystery message all together that acknowledgement turns out to be hard and a steganography algorithm/technique (the procedure to conceal key message) [9]. The consequences of the procedure could be the stego image that is the hidden message covered up inside. That stego image is accommodated the telephone where telephone are sure to get the hidden information out of the stego image by applying deciphering algorithm [6].

#### B. Steganography techniques:

Classification of stenography techniques on the basis of the cover modifications used in the embedding process is as follows:

- Least significant bit (LSB) process: This technique is extremely basic. In this methodology minimum critical bits of a few or a large portion of the bytes inside a picture is changed with bits of the key message [3].
- Transform domain techniques: This technique implants key information in the recurrence area of the sign. Change space strategies shroud messages in critical zones of the spread picture improving them made to issues, for example, for instance: pressure, editing, and some picture taking care of, in contrast with LSB approach.
- Mathematical methods: This technique implants key information in the recurrence area of the sign. Change space strategies shroud messages in critical zones of the spread picture improving them made to issues, for example, for instance: pressure, editing, and some picture taking care of, in contrast with LSB approach [14].  
i.e. if "1" is transported then cover is changed usually it is remaining as such.
- Distortion techniques: In this method, the learning of novel spread in the interpreting procedure is key at the beneficiary side. Beneficiary methodology the distinctions with the principal spread keeping in mind the end goal to remake the gathering of alteration utilized by sender [16].

#### C. Aim of Steganography

A good means of image Steganography aims at following aspects.

1. Imperceptibility to a Human Visual Process: The difference between stego image and input image ought to be little in a way that the unapproved individual can't recognize Key data. On the off chance that more points of interest is hidden in the organization picture, it will bring about weakening of stego image [10].

2. Effective: Stego image ought not be changed regardless of the possibility that it experiences change, sharpening, separating, running, obscuring, trimming and other adjustment and so on. It ought to make the input image after converse of procedures. That ensures the idea to be covered stays safe likewise when it gets contaminated and controlled.
3. Simplicity of recognition and extraction: It is for an individual who offers the individual vital to recuperate the idea, it must to be easy to get it. For other users who do not have a key, it should be extremely mind boggling to reveal its substance. That guarantees that the watermarking is perfect and might be simply deciphered by receiver [4]..
4. Large data capacity: The watermarked picture must figure out how to take huge data without loading the course or the underlying picture. That property recognizes the amount of data must be implanted for right carriage.
5. Invisibility: The information is undetectable to people in general and no one can get to the mystery information without authorization of sender. On the off chance that anybody has the code to unscramble just that individual can split the secret message [8].

#### *D. Types of Steganography*

Steganography is the art of hidden communication. This communication takes place by hiding data inside data. The various techniques used in steganography are as follows:

1. Text Steganography
2. Image Steganography
3. Audio Steganography
4. Video Steganography

#### *E. Image Steganography*

Images are the main technique for concealing key information. An image is a gathering of quantities of various light intensities. Diverse sorts of images are utilized as a part of picture steganography. In photographs you can discover different sort of record configurations are way out. All of record organization has its particular points of interest. Diverse steganographic techniques exist, for these distinctive kinds of picture record designs. Spread picture is the primary source in which we can conceal the shrouded subject; it can be anything like content, photographs, sound, motion picture and so on. The span of cover image dependably is bigger than the disguised item. The resultant picture is called stego-picture. With the assistance of PSNR we can gauge the bore of stego image. PSNR is most effectively characterized by means of the mean squared error (MSE) [13].

The PSNR is defined as:

$$\text{PSNR} = 10 \log_{10} \left( \frac{R^2}{\text{MSE}} \right)$$

Bigger PSNR demonstrates great nature of the image or in different terms lower distortion. Bigger the PSNR esteem the smaller the possibility of visible attack by human eye.

## II. DATA HIDING IN 2D BARCODE

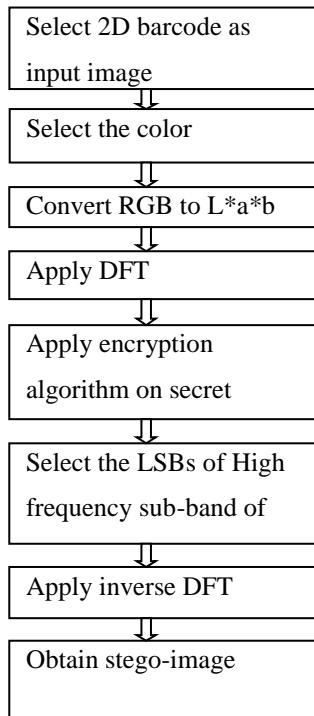
Since last couple of years, two-dimensional (2D) barcode have picked up the consideration of the general population from the mechanical foundations and bit by bit supplanted one dimensional standardized tags in numerous projects on account of their higher data stockpiling limit. QR code is subsequently as a data compartment which can be gotten and decoded by brilliant phones straightforwardly. Network scanner tag containing a ton more amount of data than their 1D form an ordinary QR code is only a sq highly contrasting pixilated box. Encoded data may incorporate simple content, work of art, or direct clients to a site or greeting page for extra data. Moreover, Barcode innovation is predominant in rate, stead quickness, data volume, all inclusiveness, and expense. The information in the code could be ensured, which needs extraordinary programming to unravel and interpret, which guarantee more prominent security [7].

Quick Response barcode is a two-dimensional scanner tag with high information thickness, mistake remedy capacity and simple security instrument. Amid, the beginning of the QR codes, the reason was to make utilization of the snappy association with the particular website page with the URL data changed over to the QR code design. However, nowadays they remain as information compartments that give more security when encoded following from the perspective of information covering up examines, QR code should then be viewed as the undeniable watermarks. The aforementioned is the motivation behind why we utilize QR code as an information holder and scramble it in our arranged application [7].

## III. METHODOLOGY

The 2D barcode image is taken as cover image to hide the information. The first code is fragmented, to conceal the information, into smaller parts, where smaller part is chosen shade of the standardized identification design that can be made by applying DFT. The information in each part is the bit of encoded into standard 2D barcode relating to that bit of information [6].

*At sender side:*



The main steps that are required to hide data by using this technique are as follows:

*Step 1:* Select the 2D data matrix Barcode and filtering step to cover image.

*Step 2:* Select the color in which data is to be hidden.

*Step 3:* Convert Image from RGB Color Space to L\*a\*b\* Color Space.

*Step 4:* Apply DFT on L\*a\*b\* color space.

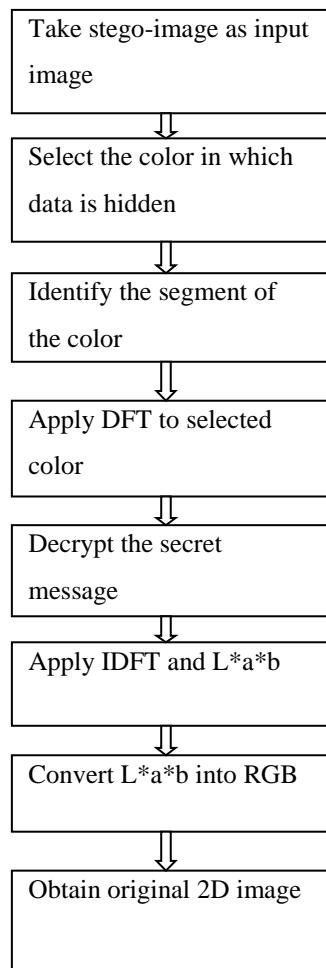
*Step 5:* Secret Text is converted into encoded text using either public or private key encryption algorithm.

*Step 6:* The DFT coefficients of Least Significant Bit (LSB) at high frequency sub-bands are selected to hide the data.

*Step 7:* The encoded secret message is hidden in the DFT Coefficients of LSB at high frequency sub bands of DFT.

*Step 8:* Finally apply the inverse DFT and combine all the segments to obtain the stego images [11].

*At Receiver side*



At the receiver side procedure is applied in inverse form which is as follows:

*Step 1:* In the first step of extraction phase, take the stego-image as input.

*Step 2:* Select the color in which data is hidden.

*Step 3:* Identify the Segment in which information is hidden.

*Step 4:* Apply DFT to the selected Segment. Select the LSBs Plane at High Frequency Sub Band of DFT.

*Step 5:* Then Decrypt the secret message using the same key.

*Step 6:* Then Apply IDFT and  $L^*a^*b^*$

*Step 7:* Convert RGB image in order to obtain the original cover image [11].

#### IV. IMPLEMENTATION

On sender side:

The initial step is that that 2D barcode image as cover image in which we need to scramble the information. At that point, to remove noise from the image applies the filtration and remove the haziness. Now, make the segments of the color in which information is to be hide using clustering using k-means. It will make different clusters of various colors which are present in the image. An exhaustive search is conducted to pair up the same color pixels within a cluster. Each pixel which is similar in color within threshold value is included in a cluster. Apply Discrete Fourier Transform to the cluster which contains the largest number of pixels. It will convert image to grayscale level as it is easier to encode data into it. Now, select the segment and then, select the LSB plane encode the data. After the message is encoded into the segment, the stego-image is created [16].

On receiver side:

Take stego-image as input image to decrypt the data. Then, select the color and identify the segment where the data is hidden. Now, apply DFT to the selected segment and select the LSBs plane of it, then decrypt the secret information from the 2d barcode image. Also, apply IDFT and  $L^*a^*b$  to combine the colors and obtain the original 2D barcode image [10].

#### V. CONCLUSION

The experimental benefits reveal that the proposed method is a better way to add concealed information reporting without distortion and with the use of conversion it becomes really difficult for the unauthorized users to recognize the changes in stego picture and it offers a way to secure the information from illegal user. Our proposed method provides better PSNR value where larger PSNR indicates better quality of the picture or in other terms decreases distortion.

This function is to enhance the steganography system using QR- code design image and different techniques. Skillful practices of code design style, unit block segmentation, design and therefore on, have now been proposed for information data embedding and extraction. Therefore, the proposed program method has been doing the information moving successfully in protected manner centered on Least Significance Bit method and QR code design image. Eventually, the occurred function is to improve the device efficiency stage examine to existing system.

## REFERENCES

- [1] Pitas, I., "A Method for Signature Casting on Digital Images," in International Conference on Image Processing, IEEE Press, 1996, pp. 215-218.
- [2] Maxemchuk, N.F., "Electronic Document Distribution", AT&T Technical Journal, September/October 1994, pp. 73-80.
- [3] Pritam Kumari, Chetna Kumar, Preeyanshi, Jaya Bhushan, "Data Security Using Image Steganography And Weighing Its Techniques", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 2, ISSUE 11, NOVEMBER 2013,ISSN 2277-8616
- [4] Gagandeep Kaur, Er.Gaurav Deep, "HSI Color Space Conversion Steganography using Elliptic Curve", International Journal of Innovations & Advancement in Computer Science, IJIACS, ISSN 2347 – 8616, Volume 4, Issue 6, June 2015
- [5] Praveen.T, Muthaiah.RM, Krishnamoorthy.N, "TRANSMITTING BULK AMOUNT OF DATA IN THE FORM OF QR CODE WITH CBFSC AND CHUNKING TECHNIQUES- FIGHTING AGAINST CRYPTANALYTIC ATTACKS", International Journal of Computer Engineering and Technology (IJCET), ISSN 0976 – 6375, Volume 4, Issue 4, July-A
- [6] Harpreet Kaur, Gaurav Deep, "Level's 4 Security in Image Steganography", International Journal of Computer Applications (0975 – 8887) Volume 121 – No.23, July 2015
- [7] Basant Sah, Vijay Kumar Jha, "A New Approach to Data hiding using Replacement of LSB and MSB", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013, ISSN: 2277 128X
- [8] Ravi K Sheth, Rashmi M. Tank, "Image Steganography Techniques", International Journal of Computer Engineering and Sciences(IJCES) Volume-1 Issue-2, 2015
- [9] International Standard. ISO/IEC 18004. Information technology --Automatic identification and data capture techniques -- QR Code 2005 bar code symbology specification. Second Edition. 2006-09-01.
- [10] Dr.Husainy.M, "Message Segmentation to Enhance the Security of LSB Image Steganography", International Journal of Advanced Computer Science and Applications, Vol. 3, No. 3, 2012.
- [11] D. Antony Praveen Kumar ,M. Baskaran, J. Jocin.and Mr. G. Diju Daniel,"Data Hiding Using LSB with QR Code Data Pattern Image", International Journal of Science Technology & Engineering, Volume 2 ,Issue 10 ,April 2016.
- [12] C. K. Chan, L. M. Cheng, "Hiding data in image by simple LSB substitution", pattern recognition, Vol. 37, No. 3, 2004, pp. 469-474
- [13] Arvind Kumar, Km. Pooja, "Steganography- A Data Hiding Technique" International Journal of Computer Applications ISSN 0975 – 8887, Volume 9– No.7, November 2010 enclosure," IEICE Trans. Commun., vol. E85-B, no. 7, pp.1360-1367, July 2002.
- [14] Chamkor Singh, and Gaurav Deep," Cluster Based Image Steganography Using Pattern Matching", International Journal of Emerging Trends & Technology in Computer Science, Volume 2, Issue 4, July – August 2013.
- [15] S. Uma Maheswari, D. Jude Hemanth." Frequency domain QR code based image steganography using Fresnel transform" International Journal of Electronics and Communications, 69 (2015) 539–544
- [16] Ankita Patel, Rahul Joshi, "SECURING MEDICAL DIAGNOSIS REPORT USING IMAGE STEGANOGRAPHY", International Journal of Advance Engineering and Research Development (IJAERD) Volume 1, Issue 5, May 2014
- [17] Pritam Kumari, Chetna Kumar, Preeyanshi, Jaya Bhushan, "Data Security Using Image Steganography And Weighing Its Techniques", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 2, ISSUE 11, NOVEMBER 2013,ISSN 2277-8616

# Digital Image Tempering Detection Techniques

Navneet Kaur<sup>#1</sup>, Navdeep Kanwal<sup>\*2</sup>

<sup>#</sup>Department of Computer Engineering, Punjabi University, Patiala

Patiala, Punjab, India

<sup>\*</sup>Assistant Professor, Department of Computer Engineering, Punjabi University, Patiala

Patiala, Punjab, India

## Abstract

Digital visual media have turned into the principle data bearers in the computerized time. Recently the quality of advanced visual data has been questioned because of straightforwardness in duplicating both its source and content. Digital image forensics is a shiny new research field which goes for approving the genuineness of pictures by recouping data about their history. Two fundamental issues tended to are: the recognizable proof of the imaging gadget that caught the image, and the detection of hints of forgeries. In the fields for example crime scene investigation, medical imaging, e-commerce and mechanical photography validness and integrity of advanced pictures is the key. Different strategies and different issues including the tempering detection and picture verification have been examined and reasonable suggestions for security situation have been exhibited.

## [1] Introduction

Images and videos have turned into the fundamental data transporters in the advanced period. The least difficult video in TV news is regularly acknowledged as a confirmation of the honesty of the reported news. Correspondingly, video observation, recordings can constitute principal trial material in an official courtroom. Together with undoubted advantages, the availability of advanced visual media brings a major drawback. Image processing specialists can undoubtedly get to and alter picture content in this way its significance without leaving outwardly noticeable follows is lost. Besides, with the easy availability of editing tools, the craft of altering and forging visual substance is not any more limited to specialists. As a result, the alteration of image for malicious purposes have become common. Digital forensics is the way toward revealing and translating electronic information. The objective of procedure is to detect any proof in its most beginning structure while carrying out an organized examination by gathering, distinguishing and approving the computerized data for reason for reproducing past confirmations.

Photography lost its significance numerous years back. Just couple of decades after niepce made the principal photo in 1814 that was at that point manipulated .with the appearance of high determination computerized cameras, intense PCs and modern photograph altering programming, the control of photographs is turning out to be more regular. Here we briefly given the case of photograph treating all through history ,in figure1 there is the print implies to be of general Ulysses S.Grant before his troops at city point ,Virginia ,amid the American common war .Some extremely pleasant analysts at the library of congress uncovered that this print is a composite of three separate prints :

- (1) The head in this photograph is taken from the picture of Grant

- (2) The steed and body are those of real broad alexander M.McCook ;and
- (3) The foundation is of confederate detainees caught at the clash of Fisher's Hill,VA.

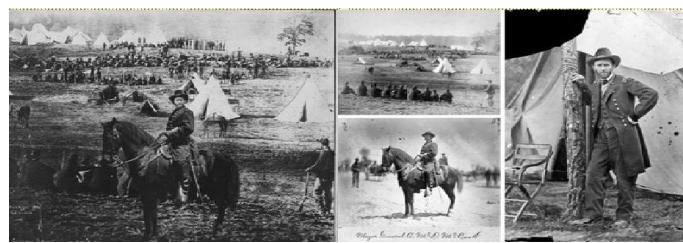


Figure 1

An image can experience diverse types of assaults amid the tempering process. Among these assaults, the least complex one is the copy move, in which the part of the picture is copied inside the same picture. The picture parts can likewise be replicated from different pictures (image splicing) and the picture itself or the altered locales can other distinctive sorts of changes to make the tempering undetectable. likewise selected to assemble the distinctive visually impaired image forgery detection approaches under five noteworthy categories i.e., pixel-based, compression based , camera-based, physics based, and geometric based methods[1].

A. Pixel based method:

This method depends on distinguishing the measurable inconsistencies happened in picture pixels amid the tempering process. These systems additionally break down relationships among pixels acquainted due with the particular type of altering in a spatial area or a changed space. These methods pixel based visually impaired forgery detection location systems have been the most generally utilized methods particularly when we realize that the easiest and most generally utilized ways to deal with falsification are additionally pixel based. Such strategies depend on the investigation of between pixel relationships that emerge from altering, either straightforwardly or indirectly. As said in the Introduction, the most regular pixel-based falsification discovery methodologies are copy move, Image splicing, Re sampling, lastly retouching detection.

B. Compression based method:

The change of a forged image with the end goal of compression and different applications can make forgery detection an exceptionally difficult errand. JPEG picture pressure, for instance, is appeared to make falsification recognition extremely troublesome. In any case, in crime scene investigation examination, a few properties of JPEG compression are abused to recognize the follows left by altering. These procedures can themselves be assembled into JPEG quantization based [2], double JPEG compression based [2-4].

C. Camera based method:

The image acquisition process in an advanced camera framework includes distinctive handling stages. In the first place, the light enters the camera lens then goes to the sensors through Color Filter Array (CFA). The sensor contains a cluster of photo detectors that catch occurrence light and change over it into voltages took after by the Analog-to-Digital (A/D) change stage. Today advanced cameras depend for the most part on Integral Metal-Oxide Semiconductor (CMOS) innovation with couple of producers as yet utilizing the conventional Charged Coupled Device (CCD) innovation. To catch color images from these sensors, CFA is utilized. The sensors catch one only color and the rest of the hues are assessed utilizing additions (demosaicing). The connections presented in the addition step can be utilized as a part of altering discovery. Before the final stockpiling, the picture quality is enhanced utilizing different upgrade procedures like Gamma revision and white parity. The curios presented in the distinctive phases of the picture creation procedure are abused to recognize hints of altering. Chromatic abnormality [5], camera source recognizable as clue, shading channel cluster, demosaicing relics [6], furthermore, sensor noise imperfections can help in estimation of various camera artifacts.

D. Physics based method:

Neutral photos are typically taken under various lighting conditions. In this way, the lighting of a produced locale may not coordinate the first in joining operations (where two or more pictures are utilized to make a fashioned picture). In material science based procedures, the irregularities in light source between particular items in the scene are utilized to uncover the clues related to tampering [7–9]. Johnson et al. [7] proposed an altering discovery strategy that uses the course of occurrence light and figured a low-dimensional descriptor of the lighting environment in the picture plane. The algorithm estimates the brightening heading from the force circulation along physically clarified object limits of homogeneous shading. Kee et al. [9] extended this way to deal with misusing known 3-D surface geometry.

## [2] Image tempering detection techniques

Image tampering is a computerized craftsmanship which needs comprehension of image properties and great visual inventiveness. One tempers pictures for different reasons either to appreciate fun of advanced works making unimaginable photographs or to deliver false proof. regardless of whatever the reason for act may be, the counterfeiter ought to utilize a solitary or a blend arrangement of picture preparing operations .To distinguish the manufactured locales there are treating recognition systems ,we break down the best in class of forgery exposal strategies.

### 2.1 Copy move forgery detection techniques:

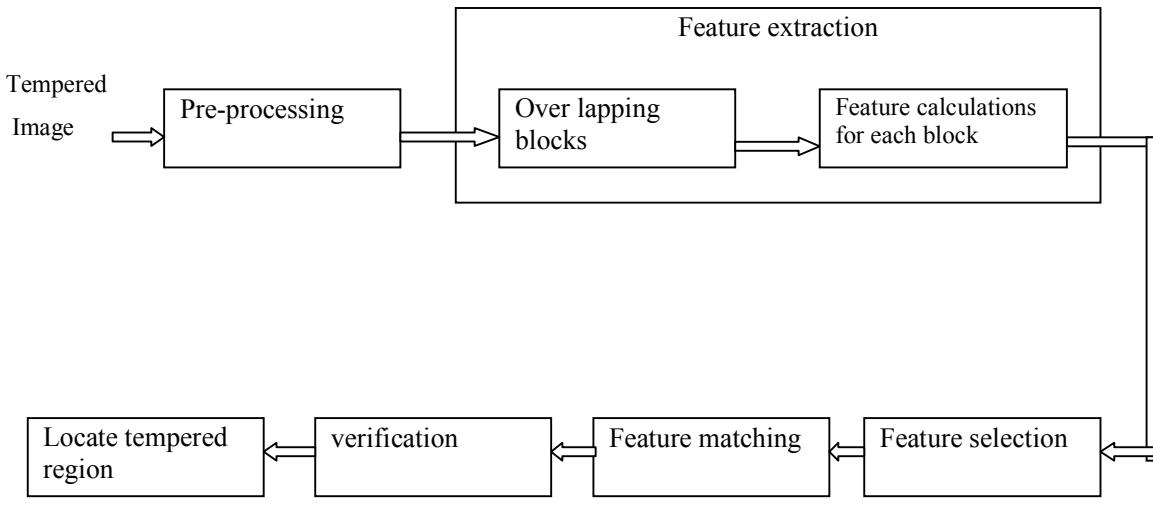
The diverse parts are copied and moved to the same image in duplicate move fraud as shown in figure 2, thus solid connection exists between these which can be utilized as a proof for forgery identification. Be that as it may, the primary test is to find proficient elements and matching algorithms for finding related portions. In these techniques, to start with, trademark elements are computed either by isolating the picture into covering pieces or computing neighborhood key focuses for the entire picture.

The positions of every piece (or key point) are additionally put away in the element vector. At that point, the element coordinating is performed to discover comparable components inside the same picture. The imitation confinement is finished by showing the coordinated pieces (or key focuses) in colors comparing.



**Figure 2- example of copy move tempering**

One of the most copy move forgery detection algorithms is proposed by Fridrich et al. [10], in light of isolating the picture into settled size covering squares and put away as 1-D feature vector. An imperative was made in connection to the decision of square size, as it ought to be not exactly the span of greatest copy move block section. Next, a movement vector methodology was utilized for feature vector and hinders with the same movement were announced as tempered regions. This general system is regular among most copy move temper detection strategies and the primary steps are portrayed as takes after:



**Block diagram for copy move tempering detection algorithm**

All the work in [10] vary just as far as the features utilized for the tempering detection process. Here, we propose to arrange these calculations into three general classes: Spatial, Transform based and feature matching methods[1].

#### 2.1.1 Spatial domain methods:

In spatial space, the pixel area specifically portrays the content in a picture. The energy in spatial space is for the most part consistently appropriated and neighboring pixels are related. This makes the coordinating procedure very computational. The duplicate move forgery detection methods in spatial domain can either be founded on moments, intensity, key focuses or textured based components.

#### 2.1.2 Transform based methods:

In transform domain, the coefficients show more often than not lesser relationship and just couple of coefficients convey the vast majority of the vitality, which implies that exclusive couple of coefficients can be utilized as feature for every overlapping block. Transform domain strategies can be partitioned into either frequency based, composition based, and reduced dimension based strategies.

#### 2.1.3 Feature matching methods:

When components have been extricated for every image block(or key points) from the entire image, a correlation is performed to match features from various (or key points) with each other. The principle issue is the means by which to qualify similitudes among elements. The most direct approach is a comprehensive research where each feature vector is contrasted and the others. This methodology is computationally with a time complexity of  $O(MN)$  for a picture of size  $M N \times$ . Other than this, it does not work for recognizing blocks that have experienced a few adjustments

### 2.2 Image splicing detection techniques:

Image splicing is a glue up created by staying together photographic images as shown in figure 3. While the term photomontage was initially used to allude a work of art or demonstration of making composite photo can be followed back to the season of camera innovation[11]. The errand of image splicing detection among bona fide and tempered image is a case of binary classification .It begins with the preprocessing stage which is generally color to grey-scale transformation took after by the element extraction stage. Diverse sorts of elements are separated from credible and altered pictures for a given dataset. The component extraction stage is basic and the characterization execution relies on upon the choice of best components for the issue under investigation. The removed components are used to prepare a classifier and the prepared model is utilized to characterize the genuine and altered pictures. At last, in the post-preparing stage, the altered areas are limited.

Amid the image splicing process, for the instance of out of focus blurred images, the altered regions are falsely blurred to coordinate the blur of the actual image. The irregularity is left in the blur and can be utilized as a proof of altering. Kakar et al. [12] proposed an image splicing detection technique in view of irregularity in the movement blur. Among various pixel based splicing detection algorithms talked about here, the probabilistic-based Markov highlights have been appeared to be exceptionally compelling in distinguishing splicing based altering which likewise accomplish the most elevated recognition exactness contrasted with different methodologies. The previously stated methodologies were utilized for splicing detection, nonetheless, the majority of these can't limit the splicing regions in the altered picture. Among various joining confinement systems, Amerini et al.[13] proposed a image splicing localization algorithm based on first digit highlights extricated from the DCT coefficients and an SVM classifier. The altered areas were limited by characterizing picture obstructs into single and double JPEG compacted blocks. The primary favorable position of the calculation was its adequacy as for various estimations of the compression quality element and additionally the size.



Figure 3-An example of image splicing tempering technique

### 2.3 Image retouching technique:

Image retouching is a typical strategy utilized as a part of the media industry. It is seen as an adequate and at times an attractive strategy for changing images. It doesn't result in any critical change in a picture rather it accentuates (or decreases) some attractive (or undesirable) components of the picture (see Fig. 4). It is a well known procedure utilized with magazine photographs and as a part of motion pictures. The picture is upgraded to make it more appealing and here and there a few areas are changed, (for example, evacuating wrinkles) to acquire the last photograph, while such sort of control is not seen as producing, we incorporate it here as it includes altering the first image[14]



Figure 4-Example of retouching tempering technique

Since the most recent decade, electronic and print media as very much utilized picture correcting devices like Adobe Photoshop, and many other editing tools, to make the photographs more regular and alluring. Retouched photographs are utilized to make a honorable representation of genuine excellence. While the writing has diverse methodologies in identifying the inventiveness of a picture what's more, are sorted by location of upgrade operation.

### 2.3.1. Image inpainting detection:

Image inpainting is utilized to regain and/or evacuate a few sections of the image with no perceptual loss. It depends on copy and move of most comparative parts of the picture with some all around characterized standard. It is not quite the same as copy move forgery one might say that diverse patches are originating from various areas of the picture rather than same persistent range of the picture. Its applications are to expel content or stamps from pictures and to evacuate wrinkles in picture modifying. So copy move tampering algorithms are not specifically pertinent for image inpainting recognition.

As of late Liang et al. [13] proposed an effective algorithm for image inpainting location. It began with looking for comparative blocks to recognize doubted areas what's more, blocks having a place with the uniform ranges were filtered utilizing vector likeness. The tempering localization was enhanced utilizing Multi-Region Relation (MRR), to expel the dubious blocks having a place with the uniform regions. The computational velocity was enhanced utilizing weight change based element coordinating. The test has shown the adequacy of the algorithm on various inpainted pictures and in addition for copy- move tempered image.

Image temper technique	Image operations or Tool used	Temper detection techniques
Copy-move (exhaustive search ,block matching)	Copy , move (using DCT or PCA)	Spatial domain method ,transform based methods and feature matching methods
Image splicing	Copy ,resize ,move ,analysis, feature extraction, feature mapping and verification	Pixel based splicing algorithms
Image retouching	Image correcting ,blurring image editing	Image inpainting detection algorithms

Table 1: Image tempering techniques and detection techniques

### 3. Conclusion

Due to significant change in processing and system advances ,and the accessibility of better transmission capacities ,the previous couple of years have seen an impressive ascent .we illustrated that advanced picture imitation identification approaches have customarily been classified into dynamic and inactive strategies. Under the dynamic structure, hearty strategies have been created for phony location, including watermarking also, signature era. These, nonetheless, have constrained applications as they require some pre processing at the picture creation stage. All the more critically, the vast majority of the web pictures are not inserted with computerized signature on the other hand watermark. Under such a situation, computerized pictures can't be validated utilizing dynamic strategies. Among the distinctive approaches utilized under this visually impaired situation, we appeared that pixel-based strategies keep on being the simples we illustrated that advanced image forgery identification approaches have customarily been classified into dynamic and inactive strategies. In this paper the various forgery detection techniques are discussed and their detection methods and algorithms are also discussed. Under the dynamic structure, hearty strategies have been created for forgery detection. These, nonetheless, have constrained applications as they require some pre processing at the image creation stage. Under such a situation, computerized pictures can't be validated utilizing dynamic strategies. Among the distinctive approaches utilized under this visually impaired situation, we appeared that pixel-based strategies keep on being the simples. Our extended overview of image forgery detection techniques shows that this area of research is still in its flourishing stage, and holds a huge potential for future R&D applications.

### References

- [1] M.A. Qureshi, M. Deriche / Signal Processing: Image Communication 39 (2015) 46–74.
- [2] M.K. Johnson, H. Farid, Exposing digital forgeries through chromatic aberration, in: 8th Workshop on Multimedia and Security, ACM, New York, NY, USA, 2006, pp. 48–55.
- [3] T.Bianchi,A.Piva, Detection of nonaligned double JPEG compression based on integer periodicity maps,IEEE Trans.inf.Foerensics secur.7(2)(2012) 842-848.
- [4] A.E. Dirik, N.D. Memon, Image tamper detection based on demosaicing artifacts., in: International Conference on Image Processing (ICIP), IEEE, San Diego, CA, USA, 2009, pp. 1497–1500.
- [5] M.K. Johnson, H. Farid, Exposing digital forgeries by detecting inconsistencies in lighting, in: 7th Workshop on Multimedia and Security, ACM, Geneva, Switzerland, 2005, pp. 1–10.
- [6] M.K Johnson ,H.Farid, Exposing digital forgeries in complex lightening envioronment ,IEEE Trans.Inf .Forensics Secur. 2(3) (2007)450-461.
- [7] Kee, H. Farid, Exposing digital forgeries from 3-D lighting environments, in: International Workshop on Information Forensics and Security (WIFS), IEEE, Seattle, WA, USA, 2010, pp. 1–6.
- [8] A.J. Fridrich, B.D. Soukal, A.J. Lukáš, Detection of copy-move forgeryin digital images, in: Digital Forensic Research Workshop (DFRWS),Citeseer, 2003.
- [9] International journal of Advanced Research in Computer and Communication Engineering Vol.2,Issue 10.October 2013.
- [10] P.Kakar ,N.Sudha ,W.Ser, Exposing digital image forgeries by detecting discrepancies in motion blur, IEEE Trans.Mutimed .13 (3) (2011) 443.452.
- [11] Amerini, R. Becarelli, R. Caldelli, A.D. Mastio, Splicing forgerieslocalization through the use of first digit features, in: International Workshop on Information Forensics and Security (WIFS), IEEE, Atlanta, GA, USA, 2014, pp. 143–148.
- [12] K.Eismann ,Photoshop Restoration and Retouching,Pearchpit Press,2005.

- [13] B. Li, Y.Q. Shi, J. Huang, Detecting doubly compressed JPEG images by using mode based first digit features, in: 10th Workshop on multimedia Signal Processing, IEEE, Cairns, QLD, Australia, 2008, pp. 730–735
- [14] Fu, Y.Q. Shi, W. Su, A generalized Benford's law for JPEG coefficients and its applications in image forensics, in: Electronic Imaging, vol 6505, International Society for Optics and Photonics, San Jose, CA, USA, 2007, p. 65051L.

# Evaluating TCP Variants Performance According to Scenarios

Mandeep Kaur<sup>1,a</sup>, Abhinav Bhandari<sup>2</sup>

<sup>1</sup>*M.Tech Scholar, Department of Computer Engineering, Punjabi University Patiala, Punjab, India.*

<sup>2</sup>*Assistant Professor, Department of Computer Engineering, Punjabi University Patiala, Punjab, India.*

**Abstract-**In today's modern era of high-speed multimedia transmissions congestion on the network is eminent. By the need of time a lot of variants have been proposed by the researchers to cope with the problem of congestion. Every variant being proposed have their pros and cons for one or the other scenario. In our study we have considered the three scenarios that are Access-link, Dial-up and Geo satellite Scenario. We have taken up some parameters of TCP Variants to be compared for these scenarios in order to provide a comparison for the use of a particular variant in a specified scenario. The purpose of this paper is to analyze and compare the performance parameters of TCP variants for three different scenarios.

## I. INTRODUCTION

TCP is a connection oriented reliable protocol for end-to-end data transmission over the communication network. TCP was originally modeled to control congestion over wired networks. The wired scenarios were very less susceptible to delay and corruption of data packets and the only cause of packet loss was congestion on the network [1]. TCP Tahoe and Reno was basically made to control congestion at that time. By the time due to change in scenario from wired to wireless the original TCP was not able to handle the congestion. Modification to original TCP was required to be done. Wireless scenarios were more prone to network congestion due to the problem of variable and high delay with high bit error rate [1]. The other cause of congestion is the immense increase in number of network users which leads to the sharing of resources like bandwidth, buffers and queues. When more than one packet reaches the receiver end node at the same time only one packet is accessed at a time and the others are being queued or dropped [2].

The performance of TCP variants we are going to evaluate are as follows with the brief introduction of mechanism they use to control congestion:

1. RENO: TCP Reno implements four phases for congestion control that are: Slow Start, congestion avoidance, fast retransmit and fast recovery. Congestion indications are time out and reception of three duplicate acknowledgements. After a congestion episode cwnd is decremented to its half [3].

2. SACK: SACK is basically an option added to TCP header in order to modify fast recovery phase of TCP Reno. It helps to recover from non-contiguous packet losses from same window by reducing the number of packets to be transmitted again and again by providing information about the received packets [3].

3. BIC: Binary increase congestion control is a variant proposed for high speed networks with high latency (long fat networks) [4]. BIC implements three phases in its mechanism. The first one is Binary Search increase phase which searches for an appropriate window size with no packet losses i.e. target window size. Second phase is additive increase phase which increases the window size additively unless a packet loss occurs. And the third Max probing phase, if no packet loss occurs at updated window it probes for a new maximum window size [5].

4. CUBIC: Cubic is a high-speed TCP variant implemented by modification to TCP BIC congestion control mechanism in order to achieve intra-protocol fairness among competing flows [6]. For window size updating TCP Cubic following function:

$$\text{cwnd} = C(t - K)^3 + \text{MWS}$$

Where, C = Scaling factor

t = elapsed time from last window reduction.

$K = K = 3\sqrt{(\text{MWS} \cdot \beta)/C}$ , where  $\beta$  is multiplicative decrease factor [7].

5. HSTCP: High speed TCP was deployed in networks with large bandwidth delay products [8]. HSTCP is the modification to standard TCP. It works same as standard TCP when window size is low and increases its window size aggressively when a loss is detected. It uses two functions for congestion control. In normal mode it increases its window size by  $a(w)/w$  where, w is congestion window size and after congestion detection window size is decreased by  $(1-b(w))w$  that multiplicative decrease function.

6. HTCP: HTCP is another protocol deployed for high speed and long-distance networks. The feature which makes it different from other variants is its compatibility with conventional networks. HTCP rapidly respond to changes in the bandwidth and utilize it in an efficient manner. It has been observed by experimentation that  $\alpha_i$  (additive increase) should be small for conventional networks and larger for high-speed and long-distance networks [9].

7. STCP: Scalable TCP is a variant deployed for high-speed wide area networks. The algorithm of STCP is the modification at sender end only for congestion avoidance phase [10]. The congestion window is updated by sender in the following manner:

- on every received acknowledgment:

$$\text{cwnd} = \text{cwnd} + 0.01$$

-when a loss is being detected:

$$\text{cwnd} = 0.875(\text{cwnd})$$

This paper aims to provide a comparative performance evaluation of TCP variant in three different scenarios. The key contributions of this paper are:

- 1) To provide a brief introduction to TCP Congestion control variants.
- 2) To evaluate the performance of TCP Congestion Control variants and providing their comparison w.r.t different scenarios.

Rest of the paper is organized as follows:

In the next Section we provide related works done before. Section 3 defines the simulation set that is common to all the three scenarios. Section 4 provides a brief introduction to the scenarios which are simulated for the comparison of TCP variants. Section 5 includes the performance evaluation of TCP Variants in the form of tables, meant for comparison. Section 6 provides the conclusion to the work being done.

## II. RELATED WORKS

Congestion is problem which may never have a permanent solution to end it up. A lot of work is reported in the existing literature. For example The performance Behavior of TCP Variants is also being analyzed in terms of

throughput, fairness and friendliness [4]. A Survey of Congestion Control Mechanisms have been done in which provides a brief of variants for wired, wireless and satellite networks [8].

### III. SIMULATION SETUP

To perform our test evaluation we are using network simulator 2 and TCP evaluation suite. The following parameters are taken in our simulation common to all the three scenario configurations:

Simulation time: - 100seconds

Topology: - Dumbbell

Traffic Setting

Number of forward ftp flows: - 3/s

Number of reverse ftp flows: - 3/s

HTTP connection generation rate: - 2/s

Number of forward video streaming flows: - 2

Number of forward video streaming flows: - 2

Rate of each streaming flow: - 640kb

Packet size: - 840B

Number of Tmix flows: - 3

### IV. SIMULATED SCENARIOS

- i. ACCESS-LINK SCENARIO: Scenario which contains a number of links to access data from other places or databases is access-link scenarios. In our test bed we have taken 10mbps bottleneck bandwidth, core delay of 2ms and buffer length 10ms for access-link scenario.
- ii. DIAL-UP SCENARIO: Dial-up is a way to connect to internet through telephone line. These types of internet connections are very slow. Our dial up scenario prefers bottleneck bandwidth of 0.064mbps, core delay of 5ms and buffer length 7000ms.
- iii. GEO-STATIONARY SATELLITE SCENARIO: These types of scenarios include communication through satellites. Our test bed prefers bottleneck bandwidth of 1mbps, core delay of 300ms and buffer length of 100ms for geo-stationary satellite scenario.

## V. PERFORMANCE EVALUATIONS

ACCESS-LINK SCENARIO:

TABLE I  
 PERFORMANCE EVALUATION OF TCP VARIANTS IN ACCESS-LINK SCENARIO

	RENO	SACK	BIC	CUBIC	HSTCP	HTCP	STCP
Average Utilization	0.408	0.408	0.406	0.393	0.41	0.409	0.411
Average Queuing Size(No. of packets)	0.5	0.5	0.6	0.2	0.5	0.5	0.5
Average Queuing Delay(seconds)	0.00029	0.000295	0.000306	0.000153	0.000299	0.000327	0.000289
Packet Drops (no. of Packets)	1515	1494	1574	535	1450	1974	1476
Average throughput(bps) initiator sending	52096426.67	52049733.33	53660773.33	31483120	49918800	10594176	52715066.67
Average throughput(bps) initiator receiving	57014.27	57484.87	57457.67	59529.13	58076.13	60697.4	58007.07
Average throughput(bps) initiator sending	1235442.93	1230315.47	1248474.93	925049.6	123228.13	627407.47	1241152.8
Average throughput(bps) acceptor receiving	1281762.93	1277975.73	1295213.87	970082.93	1283266.4	811872.8	1291352.8

Table I: provides results of simulation for access-link Scenario. As per our comparative study in case of access-link scenario clearly, the TCP Variant –CUBIC outperforms of all the TCP variants in terms of queuing delay, queue size, packet drop rate and throughput whereas STCP utilizes the bottleneck bandwidth in most efficient manner of all the variants . And the other variants have slighter differentiation in their performance where TCP BIC and HTCP shown worst results with little differences in throughput.

DIAL-UP SCENARIO:

TABLE II  
 PERFORMANCE EVALUATION OF TCP VARIANTS IN DIALUP SCENARIO

	RENO	SACK	BIC	CUBIC	HSTCP	HTCP	STCP
Average Utilization	0.974	0.914	0.984	0.966	0.983	0.95	0.933
Average Queueing Size(No. of packets)	28.2	24.4	27.6	25.9	29	25.6	23.4
Average Queueing Delay(seconds)	0.536364	0.457165	0.527651	0.474027	0.549077	0.467889	0.4550663
Packet Drops (no. of Packets)	5593	5211	5480	5254	5779	5156	5395
Average throughput(bps) initiator sending	372242.93	321453.33	387700.53	399974.1	394091.2	339250.67	353773.07
Average throughput(bps) initiator receiving	11796	9973.33	11641	11278.67	12128.33	10395.33	10943.33
Average throughput(bps) acceptor sending	6623.9	5662.12	6885.5	6675.81	6883.02	6252.39	6412.33
Average throughput(bps) acceptor receiving	13027	11835	13948	13185.33	13370.67	13363	13146.33

Table II: Provides comparative simulation results for dial-up scenario that depicts very less differentiation in their performance, in case average utilization BIC is utilizing the bandwidth in more efficient manner than the others. STCP has minimum average queueing size and delay. HTCP has a minimum packet drop rate and BIC outperforms in case of throughput. On the negative view of evaluation SACK is performing worst in case of dial-up scenario.

GEO-STATIONARY SATELLITE SCENARIO:

TABLE III  
 PERFORMANCE EVALUATION OF TCP VARIANTS IN GEO-STATIONARY SCENARIO

	RENO	SACK	BIC	CUBIC	HSTCP	HTCP	STCP
Average Utilization	0.316	0.321	0.33	0.297	0.323	0.31	0.315
Average Queueing Size(No. of packets)	0.7	0.7	0.8	0.6	0.8	0.8	0.7
Average Queueing Delay(seconds)	0.002395	0.02428	0.002449	0.002045	0.002462	0.002695	0.002344
Packet Drops (no. of Packets)	1269	1414	1461	1286	1370	2069	1319
Average throughput(bps) initiator sending	7451960	7447821.33	7754301.33	7452472	7605522.7	7387808	7336682.7
Average throughput(bps) initiator receiving	26074.67	25812.33	25544.67	24530.67	28644.67	28715.67	26998
Average throughput(bps) initiator sending	141386.93	141648	145301.07	138657.87	143781.87	136375.73	138556.27
Average throughput(bps) acceptor receiving	178649.33	176088.8	181472.27	172073.07	182664	213566.67	175549.33

Table III: In case of Geo-stationary scenario TCP BIC utilizes the bottleneck bandwidth in better way than other variants and also have higher throughput. TCP CUBIC has least average queuing delay and lesser number of packet drops whereas HTCP is the variant performing worst in terms of queuing size, delay, packet drop rate and throughput.

## VI. CONCLUSION

As per our parameter consideration in our comparative study we hereby conclude that in case of access-link scenario TCP CUBIC is a variant that performs better than others whereas HTCP has provided worst results. In case of Dial-up scenario there is a little bit difference in results of all the variants being considered but SACK is an exception which has provided worst results in terms of throughput. The other one scenario is Geo-satellite that has verified BIC and CUBIC as the best and HTCP as the worst performing variant.

#### REFERENCES

- [1] B. Singh, "A Comparative Study of Different TCP Variants in Networks," International Journal of Computer Trends and Technology (IJCTT), vol. 4, pp. 2962-2966, August 2013.
- [2] N. V. \. D. A. M. G. Mehta Ishani, "Comparative Study on Various Congestion Control Protocols: TCP, XCP and RCP," The SIJ Transactions on Computer Networks \& Communication Engineering (CNCE),, vol. 2, no. 5, pp. 56-60, September 2014.
- [3] B. Sikdar, S. Kalyanaraman and K. Vastola, "Analytic models and comparative study of the latency and steady-state throughput of TCP Tahoe, Reno and SACK," in Global Telecommunications Conference, 2001. GLOBECOM '01. IEEE, 2001.
- [4] C. Callegari, S. Giordano, M. Pagano and T. Pepe, "Behavior analysis of TCP Linux variants," in Performance Evaluation of Computer and Telecommunication Systems (SPECTS), 2010.
- [5] L. Xu, K. Harfoush and I. Rhee, "Binary increase congestion control (BIC) for fast long-distance networks," in INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies, 2004.
- [6] I. Abdeljaouad, H. Rachidi, S. Fernandes and A. Karmouch, "Performance analysis of modern TCP variants: A comparison of Cubic, Compound and New Reno," in Communications (QBSC), 2010.
- [7] I. R. L. X. Sangtae Ha, "CUBIC: a new TCP-friendly high-speed TCP variant," ACM SIGOPS Operating System Revie - Reseach and development in Linux kernal, vol. 42, no. 5, pp. 64-74, july 2008.
- [8] S. G. M. P. a. P. T. Christian Callegari, "A Survey of Congestion Control Mechanisms in Linux TCP," Springer International Publishing Switzerland 2014, p. 28-42, 2014.
- [9] R. Morris, "Scalable TCP congestion control," in INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, 2000.
- [10] L. S. M. A. M. Jawhar Ben Abed, "COMPARISON OF HIGH SPEED CONGESTION CONTROL PROTOCOLS," International Journal of Network Security \& Its Applications (IJNSA), vol. 4, no. 5, pp. 15-24, 2012.
- [11] R. S. D.Leith, "H-TCP: TCP for high-speed and long-distance networks," Research Gate, 2004.
- [12] An Efficient Congestion Control Protocol for TCP. Hanaa Torkey, Gamal ATTIIYA, Ahmed Abdel Nabi. 1, s.l. : IJECCE, 2014, International Journal of Electronics Communication and Computer Engineering, Vol. 5.
- [13] [17] A comparison of mechanisms for improving TCP performance over wireless links. Katz, H. Balakrishnan and V. N. Padmanabhan and S. Seshan and R. H. 6, s.l. : IEEE, 2009, IEEE/ACM Transactions on Networking, Vol. 5, pp. 756-769.
- [14] [18] A New Survey on Improving TCP Performances over Geostationary Satellite Link. Pirovano, Alain and Garcia, Fabien. 1, 2013, Network and Communication Technologies, Vol. 2.
- [15] [19] A Survey on TCP Congestion Control Schemes in Guided Media and Unguided Media Communication. Jain, Monal, Tomar, Deepak Singh and Tomar, Shiv Kumar Singh. 3, 2015, International Journal of Computer Applications, Vol. 118, pp. 20-29.
- [16] [20] Sharma, S. Poojary and V. Approximate theoretical models for TCP connections using different high speed congestion control algorithms in a multihop network. Communication, Control, and Computing (Allerton). s.l. : IEEE, 2013, pp. 559-566.
- [17] [21] Ding, Nan and Rui-Qing Wu, Hong Jie. Enhanced TCP Congestion Control Based on Bandwidth Estimation and RTT Jitter for Heterogeneous Networks. s.l. : Springer, 2015, Vol. 322, pp. 623-632.

# A Survey on Software Defined Network

Kuldeep Kaur<sup>1</sup>, Asst. Prof. Anantdeep Kaur<sup>2</sup>

Department of Computer Engineering, Punjabi University Patiala

**Abstract-** Traditional networks are based on static configuration are improve to the Modern high speed network of dynamic configuration. Now day's the computer networks are modern, more difficult and complex, based on dynamic configuration [2] and traffic management challenges [3]. These challenges can be require the large number of connections and various switches, routers and firewalls that are connect with each other. To accomplish these tasks and various traffic challenges, SDN can be easily and confidently used. When SDN is applying to NFV (Network Function Virtualization) can help in addressing the challenges of dynamic resources allocation, resource utilization and resource management [6]. The architecture of SDN is focused on some additional security requirements and threats issues because of newly deployed infrastructural entities [7]. SDN can also be requiring the various improvements [2] in the Network Management tasks. In the network management, SDN focus on changes to network condition and state. SDN provides the support of high level of programming language for network configuration. This paper is based on the improvement in networking, traffic control management, congestion management, configuration management.

## I. INTRODUCTION

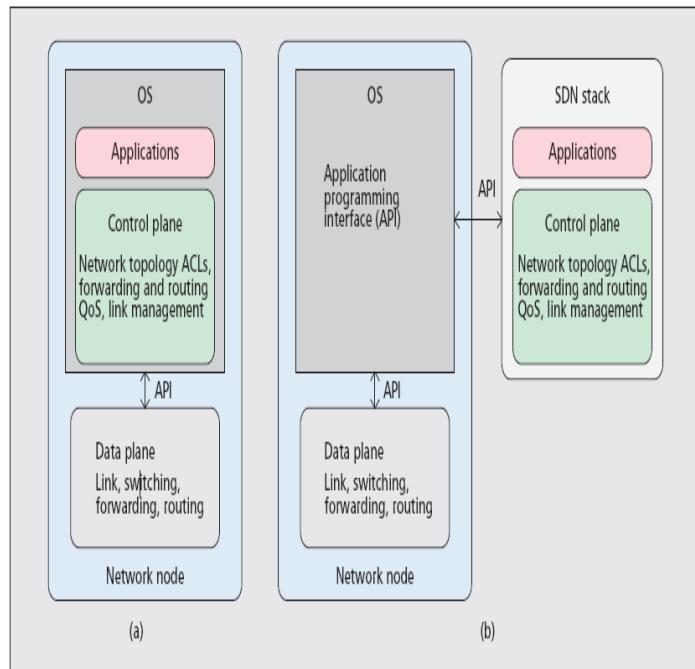
A new paradigm is comes in networking, SDN that separate the control plane and data forwarding plane and maintaining the services within this framework. Software Defined Networking provides the relationship between the network devices and the software that control them. SDN is an architecture that is dynamic, manageable cost effective and adaptive that supports the higher bandwidth of today's applications. This architecture decoupled the network control and forwarding function to enable the directly programmable and abstract the underlying infrastructure for application and network services. Software Defined Networking allow the network operation to express the required services and functionality using high level of abstraction at the control plane into the low level functionality at the data plane. SDN is an interface to network switches enable more flexible and predictable network control, so that make easier to extend network function.

SDN focus on the some key points [3]:

- Separate the control plane and data plane.
- Provide relationship between the network devices and software.
- A centralized controller and view of the network.
- Programmability of the network by external application.
- Manageable cost effective and adaptive architecture.

#### A. Why We Use SDN?

In traditional network, switch determines where packets of data and stream of data are forwarded. In SDN, network administrator is able to shape network traffic [4], using some traffic rules. They can also change the data traffic rules on the fly if they need to. SDN is open-source product and as such is open interface to the various network devices. In SDN the task of connecting up to different application, network devices and clouds are very simple. SDN allows network administrators to use software for much of work they used to do manually. The network administrator is able to control the network traffic through the SDN controller interface that SDN provides.



**Figure 1. Traditional Network v/s Software Defined Network [4]**

#### II. SDN CONTROLLER INTERFACE

In the network, Software Defined Network controller [5] supports the various tenants. Each tenants in the SDN, has its own topology and own control logic. In SDN, centralize controller are enable to manage the various switches control through the SDN controller interface. The controller interface provides the table for indexing the various packets that contain the flow rate and list of the rules. It is the responsibility of the SDN controller to define the forwarding packet processing detail that included number of packets, specific device port and lost packet detail.

### III. ACTUALIZATION OF SDN

In the network management system complexity was arise due to the abstraction, decomposition and modularization. The principle of the SDN [4] is used to limit network control complexity. SDN control services and functionality is divided into the multiple abstract layers and modularize the control tasks. The architecture of SDN uses the three distinct layers: data layer, control layer and Management layer. The data layer at the data plane, the lowest abstraction layer of the architecture. SDN enables the programmability at the control plane. SDN provides separation of network devices, aiming to transform each device into a forwarding engine.

### IV. VARIOUS CHALLENGES OF SDN

- A. *Scalability*: SDN can be focus on controller scalability [3] in which three specific challenges is identified. First, how the information is exchanging between multiple nodes and a single controller. Second, how SDN controller can communicate to another controller. Third, size of controller back-end database is small.
- B. *Flexibility*: One fundamental challenge of SDN, how to handle high speed, high security and high performance of packaging processing in very efficient and easiest way. Flexibility [3] enables the system adaptable feature that support the new feature e.g. application, protocol and security.
- C. *Security*: SDN controller provides protection against the unauthorized access and exploitation. SDN allow the security policy [3] and reduced the frequency of misconfiguration.
- D. *Interoperability*: SDN provides the solution can be integrated [3] the SDN into the existing network system. SDN support various business application, new infrastructure of data center.

### V. RECONFIGURATION IN SDN

Software Defined Networking is depend on header processing during the reconfiguration [4] in netork swiching. Each packet's header contain the source and destination address and other network details.

#### A. *Header Processing*

Header Processing [4] is needed for confidentialy transfer the packets to its destination. Header processing deciding how many number of the packets are send and replicated. Recongiuration that require switching as when packet is enter in a device, its header is added, checks various rules and protocol that needed after any operation apply on this. During the analysis some specific format is checked using some special bits, it detect any wrong format. Highly Reconfiguration system define the which protocol is suitable or not suitable, number of header being used in packets and which operation is performed when multiple protocol are used. Various reconfiguration mechanism supported by the hardware in network system. These mechanism provide a table for header processing, register for configuration.

### *B. Traffic Control Management*

It is the mechanism that are with two plane i.e. data plane and control plane. SDN focus on header processing at the data plane and traffic management [4] at the control plane. Header processing is deal with the where the packet are send but traffic management at control plane deals with handling of stream of packets at the destination. In SDN, manage the configuration management that more complex task of the network. Because configuration management focus on the changes in the traffic management i.e. shaping parameter, flow rate of traffic.

## VI. CONCLUSION AND FUTURE SCOPE

In this work we studied the overview of the SDN architecture. The traditional network architecture that is based on static configuration can not handle the user-requirements efficiently. Modern network architecture is based on dynamic configuration [3] that allows the new features and some advanced technical user-requirements very efficiently. Software Defined Network is a promising architecture that support the modern network architecture and there functionality. SDN also deals with the performance issues, reconfiguration issues [4] and scalability issues [2]. The business and industry needs time for proper synchronized the devices, services & functionality according to the architecture of SDN. Future work may be based on multi-dimentional packet classification rules for cacheflow [1]. Earlier time we also studied on a full network virtualization [5] solution for SDN. SDN, in future can be required more improvements related to technical and traffic engineering in the networking. The question comes “How the architecture of SDN can be easily support these technical advancements?”. In the future our concentration on the solution of the problem that arises due to the technical advancements and improvements [2] in network configuration and management.

## REFERENCES

- [1] Green, K. (2009). Emerging Technology. (biotech, Ed.) *TR10 Software Defined Network* , 20.
- [2] Kim, H., & Feamster, N. (2015, feb). Improving Network Management with SDN. *IEEE Communications Magazine* , 6.
- [3] Sezer, S., Scott-Hayward, S., & Chouhan, P. K. (2013, july). Implementation Challenges for SDN. *IEEE Communications Magazine* , 8.
- [4] Zilberman, N., & M.Watts, P. (2015). Reconfigurable Network Systems and Software-Defined Networking. *Proceedings of the IEEE* , 103, 23.

[5] Drutskoy, D., Keller, E., & Rexford, J. (2013). Scalable Network Virtualization in Software-Defined Network. *IEEE*, 6.

[6] LI, Y., & CHIN, MIN. (2015, DEC 9). SDN Function Vir  
tualization. *IEEE*, 12.

[7] A. A., Anuar, N. B., & Gahi, A. (2016). Secure and Dependable Software defined network. *Elsevier Ltd*, 6.

# Classification, Clustering and Regression in Agricultural Data Mining

Gurpinder Singh<sup>1</sup>, Kanwal Preet Singh Atwal<sup>2</sup>

*Department of Computer Engineering, Punjabi University, Patiala, India<sup>1</sup>*

[Gaggibenipal07@gmail.com<sup>1</sup>](mailto:Gaggibenipal07@gmail.com)

*Assistant Professor of Computer Engineering, Punjabi University, Patiala<sup>2</sup>*

[Kanwalp78@yahoo.com<sup>2</sup>](mailto:Kanwalp78@yahoo.com)

**Abstract**— Agricultural Data Mining is relatively novel field for research as compared to that of Educational data mining, Medical data mining, Business data mining etc. However, it is attracting many researchers for contributing to this area. This Paper provides an overview on the Data Mining Techniques which are frequently used in Agricultural Data Mining and some of their applications. These techniques will include Classification, Clustering and Regression to be naming a few. K-nearest, K-means and Multiple Linear Regression Algorithms are discussed with their use in applications like predicting crop yield, olive production and others. Focus of this paper is to introduce the readers with a wide range of possibilities for research in the Agriculture.

**Keywords--** Data Mining, Classification, Clustering, Regression, K-Means, K-nearest, Multiple Linear Regression.

## I. INTRODUCTION

Data Mining is the process of discovering previously unknown and potentially interesting patterns in large amount of datasets [1]. Data Mining gives us the ability to predict the output of some certain inputs by applying certain rules to the datasets. Simply stated, data mining refers to extracting or “mining” knowledge from large amounts of data [2]. The word data mining actually is considered to be a misnomer. It should have been called “information mining” in relevance to other real world mining. But to emphasize on the word “mining” it is called Data Mining. Data Mining is part of a larger process called Knowledge Discovery in Databases (KDD).

Agricultural data mining is a novel field and researchers are conducting more and more experiments every day. A variety of data mining techniques are used in Agriculture like- classification, clustering, association, regression, Machine learning, Neural Networks etc. Different Algorithms are applied to the training data and then compared.

The algorithm which produces the best result is kept for further predictions. Today, many tools are equipped with the basic data mining algorithms by default. WEKA (Waikato Environment for Knowledge Analysis) is used frequently for the purpose of data analysis and predictive modeling. Other tools may include R-tool, rapid miner, Orange and Knime.

Agriculture can be seen as a business with many risks (weather, time of irrigation, disease and pests etc.). In order to minimize these risks proper prediction of the factors which affect the production of the crop must take place. Besides farmers, agriculture provides a means of business to huge industries that depend on the agricultural products. It is a known fact that India's majority of the population is dependent on Agriculture. Agriculture plays a big role in India's Gross Domestic product i.e. 25% of the total GDP [3].

This Paper explains various techniques used for prediction and description rules. These techniques are not specially designed for the agricultural data mining but these are some of the basic methods used in any research field of data mining. Classification (Sec. 2), Clustering (Sec. 3), Regression (Sec. 4) are some of the widely used techniques of prediction and description rules used in agricultural data mining. Each of these techniques empowers many algorithms which are designed specifically for special tasks. This paper includes 3 algorithms with their applications in the real world. These algorithms are K-means (Clustering), K-nearest (Classification) and Multiple Linear Regression (Regression).

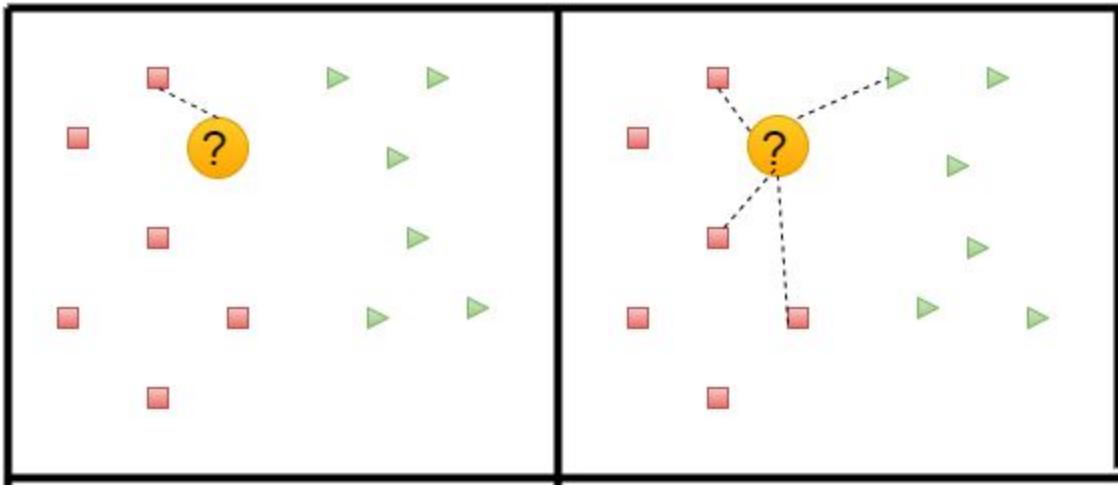
## II. CLASSIFICATION

Classification, as the name suggests helps in finding or predicting the class of any object whose class label is unknown. The classes are predetermined but algorithm predicts a certain class (from the predetermined classes) for the object whose class is unknown. In simple words, sometimes we are interested in knowing the class of some objects which we have no idea about. In that situation the algorithm would simply predict the class to which the object belongs.

The model or algorithm or rule which predicts the class of the unknown object is basically derived after the analysis of the training data. Training data is the data whose class labels are known. Training data is used to train the classification technique how to perform classification [4]. Classification predicts categorical (qualitative) class labels. For instance classification can predict the quality of a seed as Good, bad or medium by analyzing its various attributes. Here Good, Bad and medium are predetermined classes to which the future unknown object labels will be labeled. Ramesh Vamanan and K.Rmar (2011) surveyed various classification techniques for the soil database to generate meaningful relationships [5]. K-nearest Neighbor, Support Vector Machine, ZeroR, OneR, Neural Networks, Naïve Bayes Classifier, Decision trees and Fisher's linear discriminant are some of the algorithms used for classification. Classification has its applications in many fields where most of the time it works as a data mining procedure. These areas could be Medical Imaging, video tracking, Geostatistics, Speech recognition, Handwriting recognition, Biometrics, Pattern recognition and many more.

For **K-nearest Neighbor Algorithm**, a training set is available which is used to classify the unknown objects. Basic assumption in k-NN is that the similar objects should have similar classification [6]. K is the number of similar samples (known) which are used to assign class labels to the unknown objects.

K-NN is a simple method for classification. It computes the distance of unknown samples to known samples. The computational cost could be high for this method which is why it is quite expensive to practice. A. Mucherino [6] presented a figurative way to express the basic idea behind classifying which is shown in Figure 1. K-NN uses training set every time it needs to classify an unknown object or sample.



**Figure 1:** The point marked by? is unknown label which is classified according to its distance from its nearest neighbor. **(a)** K=1, **(b)** K=4.

### III. CLUSTERING

Unlike classification, Clustering doesn't have any training data set. Clustering analyses unknown object without the knowledge of known samples (classes). For e.g. let us suppose that we have pictures of some animals. Then if we are applying classification, we will know the classes in advance (say mammals, birds). And on the basis of these classes we will predict the class of another image whose class label is not known. On the other hand Clustering is quite opposite of that, consider the case that we even don't know the classes (mammals, birds) in advance. So in this situation the algorithm which is used is called a clustering algorithm. A cluster is a group of objects where the objects in a cluster are more similar to each other and dissimilar to the objects of some other cluster. The principle which is used here is maximizing the intra-class similarity and minimizing the interclass similarity (Jiawei Han et al.). Basically clustering algorithms are divided into two categories unsupervised linear clustering algorithm (which includes Gaussian clustering algorithm, Hierarchical clustering algorithms, k-means algorithm, fuzzy c-means clustering algorithm and Quality threshold clustering algorithm) and unsupervised non-linear clustering algorithm

(which includes MST based clustering algorithms, kernel k-means clustering algorithm and density based clustering algorithm).

This paper discusses about the k-means algorithm which comes in handy for agricultural data mining. K-means has been used in a research about agricultural yield data [7]. Another major application of clustering was encountered in the prediction of olive production in Thassos [4]. Focus of k-mean is to partition a dataset in which the data in a group is more similar to each other. K in k-means describes the number of clusters that should be made. Centers are marked for all the clusters in a way that they are as far from each other as possible because they can produce results if kept close.

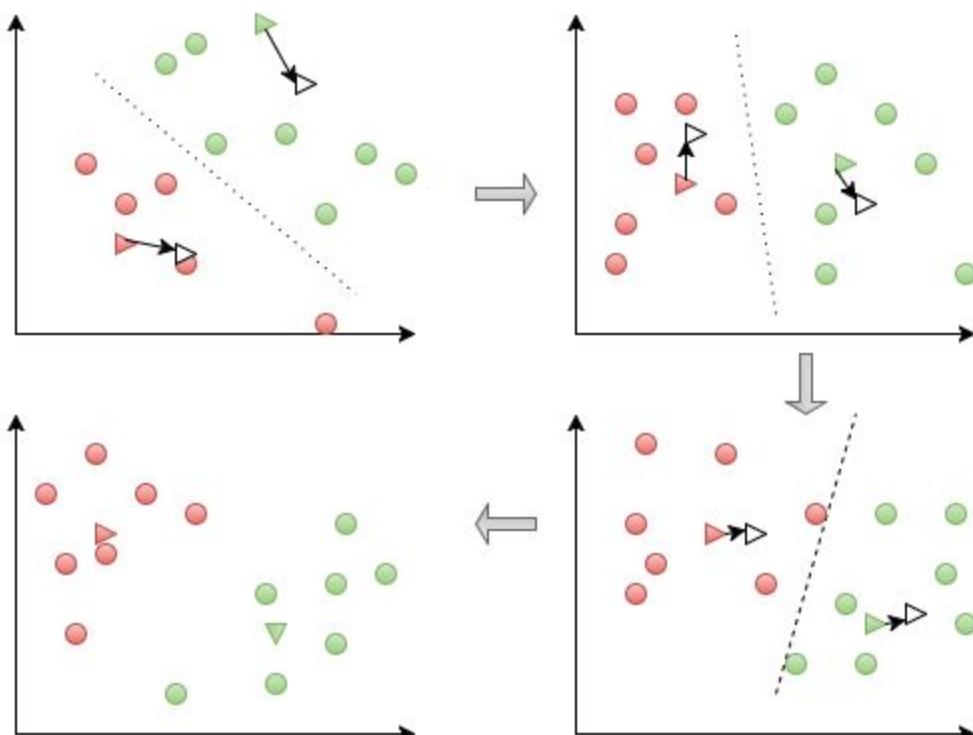


Figure 2: K means clustering example

Euclidean distance can be used for partitioning and then the objects which are near to a certain centroid will be considered a part of that cluster. After all the points are calculated, the first stage is done. After the first stage, re-calculate new k centroids and repeat the calculations of all the points. This method is performed until there is no ambiguity in the clusters and they are clearly away from each other. Figure 2 represents k-means where Euclidean distance is used for partitioning the clusters and then the distance of certain points is calculated.

#### IV. REGRESSION

Regression is a statistical approach towards predicting a value based on the best fit line for the current sample. When there is only one attribute in hand and we want to predict the future values, in that situation the best fit line will always be the mean of sample and that would be the predicted value. In case where two attributes or variables (dependent and independent) are present, then two models are compared. The first model will always be generated by considering only the dependent variable where its mean will be the best-fit line with slope of zero. Then its residuals or error will be computed. In the second model, both dependent and independent attributes are considered and similar process is followed. If the residuals are much less than that of the first model then the second model will be opted. If residuals or error is similar, then second model will be discarded. Regression or Simple linear regression is used interchangeably as they are the same thing. Different algorithms are used for data mining purposes like- Ordinary Least squares Regression, Linear Regression, Logistic Regression and Multivariate Adaptive Regression Splines (MAR). The equation used in regression is similar to the equation of a line.

#### V. APPLICATIONS

Data mining techniques are used in the development or betterment of many agricultural properties and also reduce the agricultural risks. Besides Agriculture these techniques have been introduced in medical, business, educational and governmental applications. Classification has been used to classify the mushrooms into different quality groups as the human inspectors do [8]. A dataset of 282 mushrooms was taken with objective and subjective attributes. The class labels were given as 1<sup>st</sup> grade, 2<sup>nd</sup> grade and 3<sup>rd</sup> grade. J4.8 algorithm was applied to the data in WEKA.

In 2006 P.Tittonell, K.D Shepherd, B. Vanlauwe, K.E Giller studied the effects of crop and soil management on maize productivity by applying Classification and Regression Tree (CART) analysis [9].Support vector Machine is used to classify crops [10] and scenarios relating to changing weather are analyzed using SVM is well [11]. D Ramesh and B Vishnu Vardhan used Multiple Linear regression and K-means to predict the yield data [7]. Multiple Linear Regression showed 98% of the accuracy where as K-means showed 96%, so MLR turned out to be more beneficial for the prediction of yield based on the parameters; Year, Rainfall, Area of Sowing and production.

ZeroR, K-means and Association rules were applied for classification, clustering and association [4] of the dataset. This research was conducted for the prediction of olive production in Thassos Island. Neural Networks are used to differentiate between bad and good apples, x-rays images of the apples were used to monitor the presence of water cores in Apples [12]. SVM incorporated sensors are used for smelling milk [13].

## VI. CONCLUSION

Agricultural data mining is still in its initial phase. There are various applications where data mining techniques can be used. Most of India's Population depends upon agriculture for their living, so a good prediction model for any crop can improve the quality and quantity of the crop where as it will reduce the risks in agriculture. Data Mining techniques which are used in agricultural field are the basic ones and easy to implement. Complex researches could be taken out and more complex or sophisticated data mining techniques could be introduced for particular applications. Bioclustering techniques can be applied to more complex agricultural data. Also data mining techniques can be applied in a parallel environment which is still unexplored. Mathematicians and computer scientists have to come together to bring out the best results in agricultural data mining with the help of agronomists.

## REFERENCES

- [1] W.J. Frawley, G. P.-S. (1991). Knowledge Discovery in Databases: An Overview, *Knowledge Discovery in Databases*, 57-70.
- [2] Kamber, J. H. (2000). *Data Mining Concepts and Techniques*. Urbana- Champaign: Morgan Kaufmann.
- [3] G, Y. N. (2012). Applying Data Mining Techniques In the Field o0f Agriculture and Allied Sciences. *International Journal of Business Intelligents* , 72-76.
- [4] Theodosios Theodosiou, S. V. (n.d.). Application of Data Mining Techniques to Olea europaea var. media oblonga production from Thassos. 487-497.
- [5] Ramar, V. R. (2011). Classification Agricultural Land Soils: A Data Mining Approach. *Agricultural Journal* , 82-86.
- [6] A. Mucherino, P. P. (2009). A survey of Data Mining Techniques applied to Agriculture. *Springer-Verlag* .
- [7] D Ramesh, B. V. (2013). Data Mining Techniques and Applications to Agricultural Yield Data. *International Journal of Advanced Research in Computer and Communication Engineering* , 3477-3480.
- [8] Holmes, S. J. (1999). Developing Innovative Applications in Agriculture Using Data Mining.
- [9] P. Tittonell, K. S. (2006). Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya—An application of classification and regression tree analysis. *Agriculture, Ecosystems & Environment* , 137-150.
- [10] Camps-Valls G, G.-C. L.-M.-O.-G. (2003). Support Vector Machine for crop classification using hyperspectral data. *Lect Notes Comp Sci* , 134-141.
- [11] Tripathi S, S. V. (2006). Support Vector Machine Approach to Downscale Precipitation in Climate Change Scenarios. *J Hydrol* , 621-640.
- [12] Shahin MA, T. E. (2001). Artificial Intelligence classifiers for sorting apples based on watercore. *J Agric Eng* , 265-274.
- [13] Brudzewski K, O. S. (2004). Classification of Milk by means of an electronic nose and SVM Neural Networks. *Sens Actuators* , 291-298.

# Various Tools and Techniques to Assess Information from Big Data

Gagandeep Kaur<sup>#</sup> gaganmalhotra1791@gmail.com, Er. Harpreet Kaur<sup>\*</sup> khasria.harpreet@gmail.com

<sup>#</sup> Department of Computer Engineering, Punjabi University, Patiala, Punjab, India.

<sup>\*</sup>Assistant Professor, Department of Computer Engineering, Punjabi University, Punjab, India.

**Abstract -** Big data refers to sets of data with high computational complexity and that is larger than the capacity of traditional software tools to seize, accumulate and investigate. It relates to structured and unstructured data. Big data actually revolves around 3 V's-velocity i.e. speed, volume i.e. quantity and variety i.e. types of data. Big Data is data generated from social media (Facebook, Twitter etc.) , the data generated by networks, for example IOT (Internet of Things). This research paper sheds light on various issues related to tools available, languages used to explore big data and also mining techniques needed to fetch and analyze big data. The Methodology used is the Beautiful Soup which is a python library that can perform parsing of html page and web scraping .Web scraping helps to transform unstructured data to structured form. From the study, it has been observed that in today's era, python is the most powerful language to fetch and analyze big data, because it can handle Zeta Bytes (ZBs) amount of data. Java and other languages cannot handle data more than Giga Bytes (GBs). Hadoop is the most useful and powerful tool for distributed storage and processing of large datasets, by the use of various plug-ins, it becomes easy to analyze big data.

**Keywords -** Big Data, Web Mining, Web Scraping, Beautiful Soup python library.

## I. INTRODUCTION

Big data refers to the data that is large in size. Big data is employed in explaining a huge volume of organized matter. Big Data relates to huge quantity, typical growing data sets with numerous sources, data accumulation and data collection ability. These data are rapidly expanding in all science and engineering stream, including physical, biological and medical sciences [10].

Data is generated from various sources such as devices and people. In this software world, big data is at boon. The data is flourishing at a rapid rate from the last twenty years; some of the findings about data are 2 million searching queries on Google, 277,000 tweets, 100 million emails, and 350 GB data processing on facebook every minute [9].Recent technological growth has led to plenty of data from different spheres (e.g. medical care and technological sensors, user-generated information and web data) since two decades. The term big data was coined to capture the meaning of this emerging trend [5].

#### A. Characteristics of Big Data

- 1) Volume: Volume means amount of data. The available technologies cannot handle enormous amount of data; this is a big problem for enterprises. Data can be transaction based data, streaming data and data generated by sensors. It may involve terabytes to petabytes of data and above it.
- 2) Velocity: Velocity defines the speed of data. Data is streaming at unprecedented speed and must be dealt according to the availability of time which is a very important factor. It relates to the speed at which data is processed.
- 3) Variety: Data can be organized, unorganized, half organized or a combination of three. It relates to the various types of data. It can take many forms like log files, tweets, pictures, videos, audio, text, PDF files, click streams etc [8].

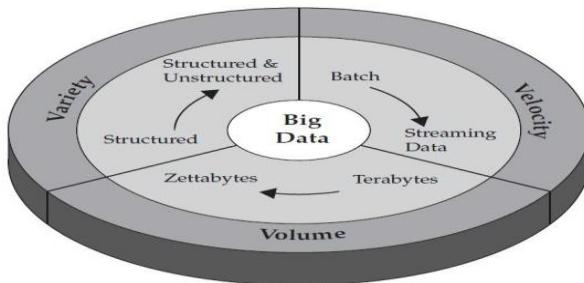


Figure 1: Characteristics of Big Data [10]

#### B. Categorization of Big Data

- 1) Structured data: This data relates to highly systematic and standard defined format that can be found in organized body of related information such as databases, data warehouses. Data is identifiable as it is organized into appropriate structure.
- 2) Un-structured data: It is a kind of raw data that has not been vulcanised into useful, meaningful and standard formats examples include log files, audio files, graphics and multimedia.
- 3) Semi-structured data: This data is the combination of the above two data which means it is neither typed data and nor raw data in a conventional database system. Geographical data, instant messages are the examples of this type of data.

### C. Web mining

Web mining relates to the implication of data mining methods and tools to capture the hidden pattern, facts and relationships from a complex network of interrelated elements. Web is a huge repository of data that is very complex and cannot be handled with simple tools. Web is huge collection of documents and hyperlink structures.

With the rapid increase of web usage, the demand of data mining also enhanced. Web mining is a vital application of big data and helps to discover usage systematic patterns from huge Web data stores. It is employed in understanding customer behavior patterns, examine the efficiency of a particular website and help to measure the triumph of a marketing campaign [2].

On the basis of type of data Web mining can be divided into three categories:

1) Web content mining – it is the method by which we can extract crucial information and content from the web records/data. Content may include textual matter, pictures, audio, visual or organized records for example lists and tables. Web content mining is compared from two separate points of view: Information Retrieval View and Database View [6, 3].

2) Web structure mining – it is the process which use graph theory to check and investigate the node and connection structure of a website. It attempts to ascertain the inherent link structure of the web. It can be used to develop information on the isomorphy or the distinctness between different websites.

3) Web usage mining – it tries to discover important knowledge from the data secured from web user sessions. It attempts to find usage systematic patterns from the web data to interpret and suffice the needs of Web-based applications in a better way. Some of the applications of web usage mining are adaptive websites, web personalization and recommendation, business intelligence [3].

## II. RELATED WORK

(Eloisa Vargiu, 2012) Web scraping is a technique to extract the embedded data and information from the internet sites. It also links to web indexing. It closely associates with the transformation of unstructured data in the form of html format into structured data that can be gone through analytics after storage. Web Scraping can be used in the monitoring of weather data, price comparison of online products etc. Various tools and techniques are available for scraping the data on World Wide Web.

(Mani, Bari, Liao, & Berkovich, 2014) Despite Present day era's tools and techniques are providing a lot of solutions for handling huge amount of data, but due to vast and enormous accumulation of big data from the global system such as emails, videos, images and text as well as the discontinuous data in detectors, sensors and wireless devices are demanding effective organizational and formatted structure.

(Sabia & Kalra, 2014) Big data is a main buzz phrase and new curve for IT today. Big data is a kind of data with high velocity, volume, variety, veracity and value. It comes from different sources like mobile devices, internet, social media, sensors, geospatial devices and other machine generated data.

RDBMS and data warehousing which are traditional data analysis technologies can no longer satisfy the growing challenges of enormous amount of data due to the high velocity and volume of big data, the most effective option is to store the big data in cloud such as bluemix, because it has capability to store and process massive and huge amount of big data.

(Hasan & Sharma, 2014) This paper describes the concept of Big Data and its research areas have been explored. This paper presents what all changes big data can bring in day-to-day lives. There are some life changing projects related to it which can completely change the way people think and look at the data and the information.

### III.FINDINGS

#### A.Traditional Database System and Big Data Solutions

In the olden days, type of information available was very small and limited. There was a well-defined set of technology approaches to deal with that information. But in today's world of science and technology the amount of data in the world has been exploding at a fast pace. It has grown from terabytes and petabytes to zettabytes and yottabytes. Table 1 explores the traditional data and big data solutions.

TABLE 1: Traditional Model and Big Data Solutions

Feature	Relational Database System	Big Data Solutions
1.Data types and formats	Structured	Semi structured and Unstructured
2.Schema	Static	Dynamic
3.Storage Volume	Gigabytes to terabytes	Terabytes, petabytes, and beyond
4.Economics	Expensive hardware and software	Commodity hardware and open source software

#### B. Analyzing Structured and Unstructured data

Analyzing structured big data (which is highly systematic as well as organized form of data) and unstructured big data (which is raw data that is transformed into useful form) can be done by various techniques as discussed in Table 2.

TABLE 2: Analyzing Structured and Unstructured Big data

Type	Techniques
1.Unstructured data	Data mining, Natural language processing and Text analytics.
2.Structured data	Machine learning and Data analysis technique.

### IV.TOOLS TO ANALYZE BIG DATA

Data can be extracted from various sources such as web pages, websites etc by using various tools as explored in Table 3. Through these tools we can pull the useful data from the web, investigate the data and draw patterns.

**TABLE 3: Tools related to Big Data and Web Mining**

Name of Tool	Importance
1.Hadoop	Framework for analyzing and storing massive amount of data.
2.Cloudera	A brand name for hadoop with extra services and security facilities.
3.R tool	Used for statistical analysis.
4.BeautifulsoupPackage using python	Used in Web Scraping.

## V. LANGUAGES USED TO HANDLE BIG DATA

Language is a way to communicate and control instructions of a machine, specifically a computer. The various programming language use to capture the data are explored in Table 4.

**TABLE 4:** Various Languages used to Capture and Analyze Big data

Language	Feature
1.Python	It is general purpose, high level programming language with greater readability.
2.R	It is an environment for statistical computing and graphics.
3.Java	It is a commonly used foundation for developing and delivering content on the Web.
4.Scala	It is designed to express common programming patterns.

## VI. WEB SCRAPING ANALYTICS USING BEAUTIFUL SOUP

Mining the data from the web is called web scraping or crawling. It is the process of gathering the data from World Wide Web for analysis purpose. Data that is available on the internet sites and web such as html tags, charts and tables can be extracted by various tools such as data tool, web scrapers. Beautiful Soup is a python library to extract and to pull out the embedded data and this library helps to convert the unstructured data into structured data using desirable python coding.

## A. Parsing HTML Page using Beautiful Soup Python

The 2.7.11 shell Python terminal is used in order to parse html content of Facebook Page. To do this we have to import the BeautifulSoup Library and Urllib that point to the specific url. Holding html\_content in a variable various commands are used to fetch the html content i.e. print soup.title. This title tag will fetch the title from parsed html page i.e. facebook.com. Parsing of page is shown in the screenshots as in Figure 2.

```
Python 2.7.11 Shell
File Edit Shell Debug Options Window Help
Python 2.7.11 (v2.7.11:6d1b6a68f775, Dec 5 2015, 20:32:19) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> import urllib2
>>> from bs4 import BeautifulSoup
>>> html_content=urllib2.urlopen("http://facebook.com")
>>> soup=BeautifulSoup(html_content)

>>> print soup.title
<title id="pageTitle">Facebook 'à"àØ‡ à"œà€ à"tà"‡à"tà", à""àØ,àØ° - à""àØà" - à"‡à"" à""*à"°àØ<
e>
>>> print soup.title.name
title
```

**Figure 2: Parsing HTML Page using Beautiful Soup Python**

### B. Fetching Structured data using Beautiful Soup Python

The following example shows how to pull out the data(quotations) from a website ([litemind.com](http://litemind.com)) and convert it into structured form from an unstructured form[11]. We can use various libraries such as `urlopen` to open a url and `Beautiful Soup`.

**Step 1**-Import `bs4` and `urllib` library and set the variable `html` to open a website such as [litemind.com](http://litemind.com)

```
>>> from bs4 import BeautifulSoup
```

```
>>> from urllib import urlopen
```

```
>>> html=urlopen('http://litemind.com/best-famous-quotes/').read()
```

```
>>> soup = BeautifulSoup(html)
```

**Step 2**-Find all the unstructured data wrapped in the html tags by using `findAll( )` function. By using inspect element we see each of the quotation is wrapped in a special div and class. We find a big list of unstructured data.

```
>>> print soup.findAll('div',{'class':'wp_quotepage'})
```

Output:

```
<div class="wp_quotepage"><div class="wp_quotepage_quote">1. You can do anything, but not everything.</div><div class="wp_quotepage_author">\u2014David Allen</div></div><div class="wp_quotepage"><div class="wp_quotepage_quote">2. Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away.</div><div class="wp_quotepage_author">\u2014Antoine de Saint-Exup\u00e9</div></div>, <div class="wp_quotepage"><div class="wp_quotepage_quote">3. The richest man is not he who has the most, but he who needs the least.</div><div class="wp_quotepage_author">\u2014Unknown Author</div></div>, <div class="wp_quotepage"><div class="wp_quotepage_quote">4. You miss 100 percent of the shots you never take.</div><div class="wp_quotepage_author">\u2014Wayne Gretzky</div></div>, <div class="wp_quotepage"><div class="wp_quotepage_quote">5. Courage is not the absence of fear, but rather the judgement that something else is more important than fear.</div><div class="wp_quotepage_author">\u2014Embrose Redmoon</div></div>, <div class="wp_quotepage"><div class="wp_quotepage_quote">6. You must be the change you wish to see in the world.</div><div class="wp_quotepage_author">\u2014Gandhi</div></div>, <div class="wp_quotepage"><div class="wp_quotepage_quote">7. When hungry, eat your rice; when tired, close your eyes. Fools may laugh at me, but wise men will know what I mean.</div><div class="wp_quotepage_author">\u2014Lin-Chic</div></div>, <div class="wp_quotepage"><div class="wp_quotepage_quote">8. The third-rate mind is only happy when it is thinking with the majority. The second-rate mind is only happy when it is thinking with the minority. The first-rate mind is only happy when it is thinking.</div><div class="wp_quotepage_author">\u2014A. Milne</div></div>, <div class="wp_quotepage"><div class="wp_quotepage_quote">9. To the man who only has a hammer, everything he encounters begins to look like a nail.<br><div class="wp_quotepage_author">\u2014Abraham Maslow</div></div>, <div class="wp_quotepage"><div class="wp_quotepage_quote">10. We are what we repeatedly do: excellence, then, is not an act but a habit.</div><div class="wp_quotepage_author">\u2014Aristotle</div></div>, <div class="wp_quotepage"><div class="wp_quotepage_quote">11. A wise man gets more use from his enemies than a fool from his friends.</div><div class="wp_quotepage_author">\u2014Baltasar Graci\u00e1n</div></div>, <div class="wp_quotepage"><div class="wp_quotepage_quote">12. Do not seek to follow in the footsteps of the men of old; seek what they sought.</div><div class="wp_quotepage_author">\u2014Basho</div></div>, <div class="wp_quotepage"><div class="wp_quotepage_quote">13. Watch your thoughts; they become words. \<br><div class="wp_quotepage_author">\u2014Watch your words; they become actions. \<br><div class="wp_quotepage_author">\u2014Watch your actions; they become habits. \<br><div class="wp_quotepage_author">\u2014Watch your habits; they become character. \<br><div class="wp_quotepage_author">\u2014Everyone is a genius at least once a year. The real geniuses simply have their bright ideas closer together.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">14. Everyone is a genius at least once a year. The real geniuses simply have their bright ideas closer together.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">15. What we think, or what we know, or what we believe is, in the end, of little consequence. The only consequence is what we do.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">16. The real voyage of discovery consists not in seeking new lands but seeing with new eyes.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">17. Work like you don\u2019t need money, love like you\u2019ve never been hurt, and dance like no one\u2019s watching.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">18. Try a thing you haven\u2019t done three times. Once, to get over the fear of doing it. Twice, to learn how to do it. And a third time, to figure out whether you like it or not.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">19. Even if you\u2019re on the right track, you\u2019ll get run over if you just sit there.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">20. People often say that motivation doesn\u2019t last. Well, neither does bathing \u2013 that\u2019s why we recommend it daily.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">21. Before I got married I had six theories about bringing up children; now I have six children and no theories.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">22. What the world needs is more geniuses with humility, there are so few of us left.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">23. Always forgive your enemies; nothing annoys them so much.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">24. I\u2019ve gone into hundreds of [fortune-teller] parlor[s], and have been told thousands of things, but nobody ever told me I was a policewoman getting ready to arrest her.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">25. When you go into court you are putting your fate into the hands of twelve people who weren\u2019t smart enough to get out of jury duty.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">26. Those who believe in telekinetics, raise my hand.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">27. Just the fact that some geniuses were laughed at does not imply that all who are laughed at are geniuses. They laughed at Columbus, they laughed at Fulton, they laughed at the Wright brothers. But they also laughed at Bozo the Clown.</div><div class="wp_quotepage"><div class="wp_quotepage_quote">28. My pessimism extends to the point of even suspecting the sincerity of the pessimists. </div><div class="wp_quotepage"><div class="wp_quotepage_quote">29. Sometimes I wonder about h
```

Figure 3: Fetching Unstructured data using Beautiful Soup Python

**Step 3**-In order to pull out the data from the html tags we have to choose `findChildren()` and `renderContents()` function.The data is converted into structured form by indexing it.

```
>>> for section in soup.findAll('div',{'class':'wp_quotepage'}):
```

```
quote = section.findChildren()[0]  
  
author = section.findChildren()[1]  
  
print quote,author  
  
break
```

Output:

```
1. You can do anything, but not everything.
```

```
>>> for section in soup.findAll('div',{"class":"wp_quotepage"}):  
  
    quote = section.findChildren()[0].renderContents()  
  
    author = section.findChildren()[1].renderContents()  
  
    print quote,author
```

Output:

```
1. You can do anything, but not everything. David Allen  
2. Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away. Antoine de Saint-Exupry  
3. The richest man is not he who has the most, but he who needs the least. Unknown Author  
4. You miss 100 percent of the shots you never take. Wayne Gretzky  
5. Courage is not the absence of fear, but rather the judgement that something else is more important than fear. Ambrose Redmoon  
6. You must be the change you wish to see in the world. Gandhi  
7. When hungry, eat your rice; when tired, close your eyes. Fools may laugh at me, but wise men will know what I mean. Lin-Chi  
8. The third-rate mind is only happy when it is thinking with the majority. The second-rate mind is only happy when it is thinking with the minority. The first-rate mind is only happy when it is thinking. A. A. Milne  
9. To the man who only has a hammer, everything he encounters begins to look like a nail. Abraham Maslow  
10. We are what we repeatedly do; excellence, then, is not an act but a habit. Aristotle  
11. A wise man gets more use from his enemies than a fool from his friends. Baltasar Gracian
```

Figure 4: Fetching structured data using Beautiful Soup Python

## VII. ANALYSIS METHODOLOGY OF BIG DATA MINING

Data mining is an important part of knowledge discovery, which is the process of converting raw data into processed information. It is used to uncover the hidden pattern from the various dimensions.

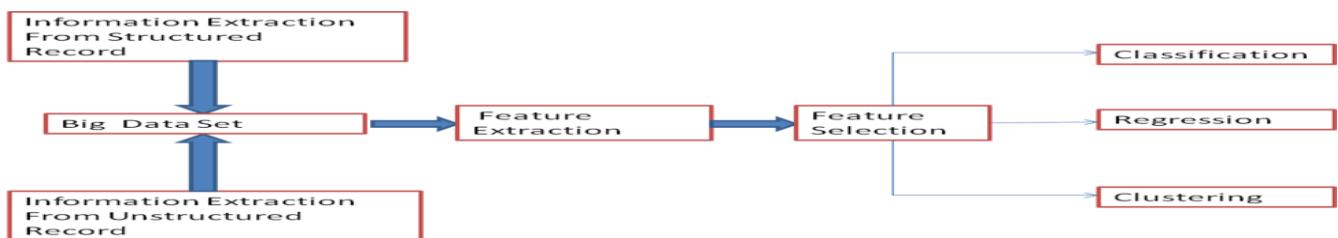


Figure 5: Methodology to analyze big data

Figure 5 presents analytical steps to extract useful information from big data. Big data analysis is basically done using tools like Hadoop, cloud-ERA etc. But data mining works as a catalyst to analyze big data. There are various data mining techniques such as clustering, classification, if-else-structured analysis. It has been observed from literature that clustering is best suitable technique to predict facts and to extract patterns. As structured-if-else analysis makes the execution slow on processing huge amount of data.

### VIII. CONCLUSION

Big data is basically new phenomenon in the computer world. The field of big data is growing from various perspectives. It reveals the importance of the latest technology which is a current trend in the software world or the IT world. Business organizations are emphasizing on the issue to organize and handle structured and unstructured Big Data. Various tools, languages and mining techniques are available to resolve the issues related to large data. It has been concluded from research that python language, Hadoop tool and clustering mining technique are best methods to fetch and to analyze Big Data. With the help of Web Scraping the disorganized form of data is converted into structured and organized form using Beautiful soup. The ability to analyze and store massive amount of data promises ongoing opportunities for various areas such as marketing, sports, business firms etc. In current scenarios large amount of data is generated regularly. In the realm of Future work of research will be to analyze human behavior on social media by apply various data mining techniques and to continue research in web scraping with more advanced software.

### REFERENCES

- [1] Eloisa Vargiu, M. U. (2012). *Exploiting web scraping in a collaborative filtering based*.
- [2] G.C, B., & T.V, S. (2009). Web Mining in Technology Management. *IJCTT*.
- [3] Gupta, R. (2014). Journey from Data Mining to Web Mining to Big Data. *IJCTT*, 18-20.
- [4] Hasan, F. D., & Sharma, A. K. (2014). Big Data: The Next-Gen Google. *International Journal of Computer Applications* , 19-21.
- [5] HU, H., WEN, Y., CHUA, T.-S., & LI, X. (2014). Toward Scalable System for Big Data Analytics:A Technology Tutorial. *IEEE* , 2, 652-687.
- [6] Jaideep Srivastava, P. Desikan, Vipin Kumar, [http://dmr.cs.umn.edu/Papers/P2004\\_4.pdf](http://dmr.cs.umn.edu/Papers/P2004_4.pdf)
- [7] Mani, G., Bari, N., Liao, D., & Berkovich, S. (2014). Organization of Knowledge ExTraction From Big Data Systems. *IEEE*, (pp. 63-69). Washington,DC.
- [8] Sabia, & Kalra, S. (2014). Applications of Big Data: Current Status and Future Scope. *ISSN* , 3 (5), 25-29.
- [9] Shilpa, & Kaur, M. (2013). Big Data & Methodology-A review. *International Journal of Advanced Reseach in Computer Science and Software engineering* , 991-995.
- [10] Tiwari, A. K., Chaudhary, H., & Yadav, S. (2015). A Review on Big Data and Its Security. *International Conference on Innovations in Information Embedded and Communication Systems*.
- [11] <https://www.youtube.com/watch?v=0mAGb6sCZWc>

# Comparison Review- Various X-Ray image enhancement methodologies

Sukh Sehaj Singh<sup>#1</sup>

#Department of Computer Engineering, Punjabi University,  
Patiala, Punjab, India  
sukhsehajsingh@yahoo.com

**Abstract-** Various image enhancement techniques are being used for the medico-diagnostic purposes. These techniques are necessary for the feature extraction or minutiae specification in the particular images. The need is to enhance the contrast and removal of the noise and blurring effect from an image. Low- contrast structures can be found in many real time images such as x-ray images. For highlighting the precise and trivial details in a image, we need to apply carefully certain algorithms and filters on it. In this paper I have compared various algorithms with respect to the x-ray system to find which is most performance efficient and gives perceptual image of damaged bones. The objective is to compare the contrast enhancement capabilities of various filters on a single x-ray image

**Keywords:** Contrast limited adaptive histogram equalization, Adaptive histogram equalization, discrete wavelet transform, Mean square error

## I. INTRODUCTION

With the invention of the conventional methods for medical diagnostic purposes it becomes very important to generate the high quality images to extract the spatial features of the image. Industrial x-ray processing requires the contrast enhancement algorithms in order to cope up with the low contrast image and to produce visually high quality images. These images such as an X-Ray image generally suffer from an unwanted fluctuation due to presence of noise. The performance of imaging sensors is affected by a variety of factors, such as environmental conditions during image acquisition .The factual capability of the medium and the limited capacity of the medium lead to the degraded image to some extent. So, improving the contrast of an X-Ray image is an important aspect in X-Ray image enhancement. Various approaches have been proposed in order to increase the contrast of the X-Ray image used for medico-diagnostic purposes. These approaches include histogram equalization, region based adaptive enhancement, discrete wavelet transform, CLAHE i.e. contrast limited adaptive histogram enhancement, morphological based image enhancement etc. All the approaches have different enhancement capabilities which can be obtained by calculating certain measures such as MSE i.e. mean square error and SNR i.e. signal to noise ratio.

## II. SEVERAL TECHNIQUES

**Adaptive histogram enhancement algorithm:** It covers the local contrast of the image and gives out the more detailed picture. This technique covers the number of histograms and each histogram corresponds to the different segments. But there is one problem with this technique i.e. it generally enhance the contrast of the homogeneous regions of the image. Li Jin[8] proposed that first step of AHE algorithm is to input image block  $A(x,y)$  and generate the local histogram of the block.

**Contrast limited adaptive histogram enhancement algorithm:** This technique uses the histogram equalization to the contextual regions. Value component is processed by CLAHE without affecting the hue and saturation. In this approach the value of each pixel is reduced to the maximum of user selectable. In first step, the captured image is converted from RGB color space to HSV color space. The image after been processed in HSV converted back to the RGB color space.

**Discrete wavelet transform:** There are two types of wavelength transforms – serial wavelength transforms and discrete wavelength transforms. The discrete wavelength transform is more useful than serial wavelength transform. Jiao Feng[2] proposed that For any square integral signal  $(x) \in L^2(R)$ , the wavelet transform is defined by the equation:

$$WT(b,a) = \langle f(x), \psi_{a,b}(x) \rangle = |a|^{\frac{1}{2}} \int_{-\infty}^{\infty} \overline{\psi(\frac{x-b}{a})} dx$$

Where  $\psi_{a,b} = |a|^{1/2} \psi((x-b)/a)$  is a wavelength function (wavelet transform use wavelet function to represent the signals approximately) , which is produced by another mother function  $\Psi(x)$ , a is a scale factor, b is a shift factor, “ “ is a conjugation function.

The 2 dimensional discrete wavelength function is represented by two dimensional filter functions using separable scaling and wavelet functions in horizontal (**n1**) as well as vertical (**n2**) direction as:

$$\phi(n_1, n_2) = \phi(n_1)\phi(n_2) \dots (1), \quad \psi^H(n_1, n_2) = \psi(n_1)\phi(n_2) \dots (2), \quad \psi^V(n_1, n_2) = \phi(n_1)\psi(n_2) \dots (3), \quad \psi^D(n_1, n_2) = \psi(n_1)\psi(n_2) \dots (4)$$

The above four equations represent approximate signals with horizontal details, vertical details and diagonal details etc.

**Mean filter:** This filter is sometimes called as the averaging filter or a low pass filter. The output response of this smoothing, linear spatial filter is the average of the pixels contained in the neighbourhood of filter mask or kernel. The idea is replacing every pixel in the image by the average of intensity levels in the neighbourhood defined by the filter mask or kernel. Because of random noise consist of sharp transitions, smoothing filters are widely used for noise reduction. Mean filter requires an image to be convolved with a  $(2K + 1 \times 2L + 1)$  kernel or mask, for example when  $L=K=1$ , we obtain a 3 by 3 smoothing filter as :

$$w(k, l) = \begin{Bmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{Bmatrix}$$

**Median Filter:** This is an order statistic filter (nonlinear) whose result is based on the ranking of the pixels in the area encompassed by the filter, replacing the centre pixel with the value determined by the ranking result. The median filter generally reduces the value of the pixel by the median of the intensity value in the neighbourhood of the pixel. For example in the neighbourhood of 3 by 3 pixels, the median is the 5<sup>th</sup> largest value.

**Combined approach :** This approach can be used efficiently for enhancing the contrast of the images specially for the medico-diagnostic purposes like X-Ray. This approach make implementation of different algorithms such as mean, median and histogram enhancement etc. on a single image with certain defined procedure one after another. This approach follows a systematic procedure which can be defined in steps as:

First step is to involve the sharpening and averaging matrix for further application. Now obtain histogram equalized view of the image to be enhanced followed by application of min filter. Next is to get the complement of image by obtaining the negative of the image. Now apply the sharpening filter on this complement image followed by histogram equalization of the resultant image. Final image is obtained by application of mean filter followed by corresponding max filter.

#### IV. RESULTS AND OUTPUTS

The above discussed algorithms have been implemented and the performance of these algorithms has been analyzed based on certain measures. The mean square error (MSE) and root mean square error (RMSE) has been used to measure the enhancement performance of these algorithms. The figure(s) shows the enhanced output images after the applications of filters over the input image.

Figure 1: Enhanced images corresponding to the original x-ray image

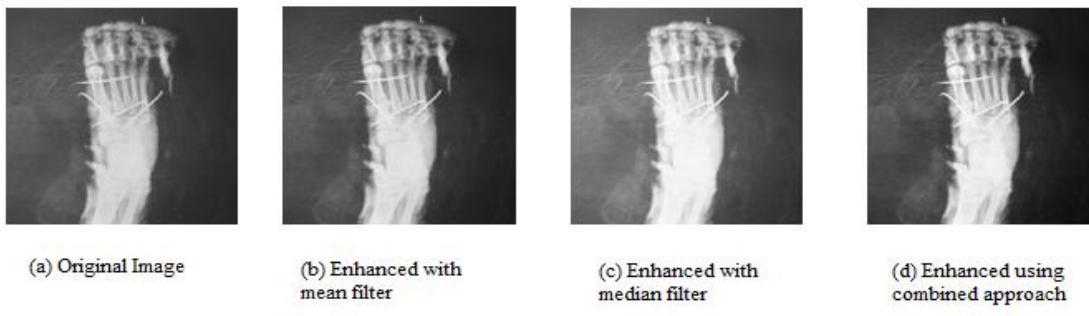


Table 1: Comparison of MSE Values

Algorithms →	Mean filter(3 by 3)	Median filter	Combined approach
Mean square error (MSE) value	0.02168	0.02228	0.02526

Figure 2: Enhanced images corresponding to original x-ray image

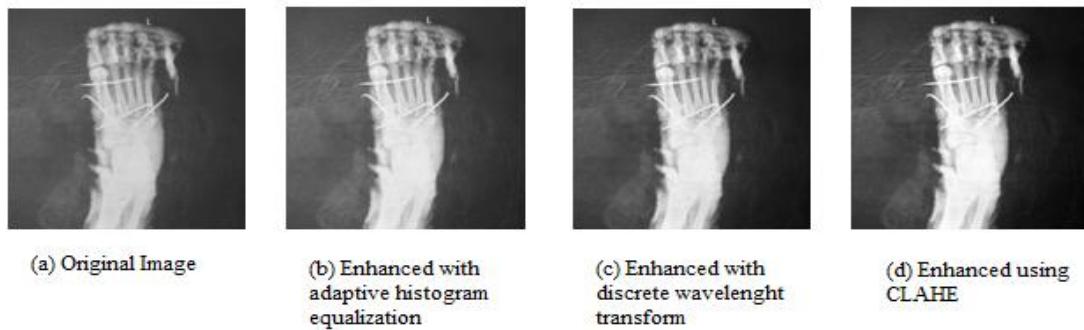


Table 2: Comparison of MSE Values

Algorithms→	AHE	DWT	CLAHE
Mean square error (MSE) values	0.0278	0.0296	0.0322

Where mean square error is given by formula:

$$MSE = \frac{1}{MN} \sum_{y=1}^M \sum_{x=1}^N [I(x,y) - I'(x,y)]^2$$

$I(x,y)$  ; input image &  $I'(x,y)$ ; enhanced or output image

M,N ; dimensions of the image to be enhanced

For the enhancement algorithm to be good, the value of mean square error (MSE) should be high.

## V. CONCLUSION

From the above we can conclude that the median filter has better performance and enhancement capability as compared to mean (3by3 or 5by5) or averaging filter. The combined approach (concurrent application of mean filter, averaging filter and histogram equalization) has better output and enhancement capability as compared to the independent implementation of each of these.

From the other Results we can conclude that the CLAHE (contrast limited adaptive histogram enhancement) algorithm is more efficient in comparison with the AHE (Adaptive histogram) and HE (histogram enhancement) algorithms.

## ACKNOWLEDGMENT

Sincere thanks for the support and the required facilities which has been provided by the ‘Department of Computer Engineering’, Punjabi University Patiala to carry out the related work efficiently.

## REFERENCES

- [1] Rafael C. Gonzalez University of Tennessee, Richard E. Woods. Digital image processing Third Edition 2008 .Pearson Education
- [2] Jiao Feng, Naixue Xiong, Bi Shuoben , ‘X-ray Image Enhancement Based on Wavelet Transform’ 2008 IEEE Asia-Pacific Services Computing Conference, pp. 1568-1573, Dec 2008
- [3] Ili Ayuni Mohd Ikhsan, Aini Hussain, Mohd Asyraf Zulkifley ,’ An Analysis of X-Ray Image Enhancement Methods for Vertebral Bone Segmentation’ , 2014 IEEE 10th International Colloquium on Signal Processing & its Applications (CSPA2014)
- [4] Navdeep Kanwal, Akshay Girdhar, Savita Gupta, “Region Based Adaptive Contrast Enhancement of Medical X-Ray Images”, Bioinformatics and Biomedical Engineering, (iCBBE) 2011 5th International Conference, pp. 1-5, May 2011
- [5] Zheng Wei, Yang Hua, Sun Hui-Sheng, “X-Ray image enhancement based on multiscale morphology”, 2007 1st International Conference on Bioinformatics and Biomedical Engineering pp. 702-705, July 2007

- [6] Juliastuti, E., and L. Epsilawati (2012), “Image contrast enhancement for film-based dental panoramic radiography”, System Engineering and Technology (ICSET), 2012 International Conference IEEE 2012, pp. 1-5.
- [7] Khan, MohdFarhan, Ekram Khan, and Z. A. Abbasi (2012), “Multi segment histogram equalization for brightness preserving contrast enhancement”, In Advances in Computer Science, Engineering & Applications, Springer Berlin Heidelberg, pp. 193-202.
- [8] Li Jin,Wang Yan Lei, Wang Lei, “Industrial X-Ray image enhancement algorithm based on adaptive histogram and wavelet”, 2011 The 6<sup>th</sup> international form on Strategic technology pp. 836-839,Aug 2011

# A Review of Detection of Model Smells in UML Model

Jaspinder Kaur

M.Tech(CE),Punjabi University,Patiala

[jaspinder257@gmail.com](mailto:jaspinder257@gmail.com)

+919417622847

Brahmaleen Kaur Sidhu

Assistant Professor, Punjabi University, Patiala

[brahmahaleen sidhu@yahoo.co.in](mailto:brahmahaleen_sidhu@yahoo.co.in)

+919501008858

**Abstract-**An increasing trend of MDE and need of high quality software has made the quality of model of great importance to the software development industry. Model smells i.e. bad smells in models are the structures in models that indicate the poor quality of design and an opportunity of improvement. Model Smells are the results of poor design decisions and make the design fragile, difficult and costly to maintain, evolve and reuse. To develop high quality software in MDE paradigm, detection of model smells is a crucial activity for developers. This paper explains the model smells, approaches and tools available for their detection.

**Keywords:** Model Smells, Detection, Quality, UML

## 1. INTRODUCTION

Model Driven Engineering (MDE)is a new trend in software development that advices the use of model as key artifact throughout all the stages of software development.AS model has become decisive software development, quality and quality assurance of models has become an important issue too. Refactoring is used for improving the quality of the software by making the internal structure of the system better without making any changes in its functionality. Bad Smell Detection is the idea of identifying the opportunities for refactoring-called bad smells. Bad smells are the structures in system that are the results of poor design decisions making the system design fragile, difficulty and costly to maintain, evolve and reuse. Bad smell detection is very crucial to produce high quality software and maintain that quality when software undergoes changes during its life. Aside from bad smell detection in model being important due to increasing importance of model s in MDE ,it also offers other advantages like earlier smell detection and detection technique being independent of any implementation language. Manual inspection of models for finding bad smells will be very time consuming so tool assistance is needed.Three of the important bad smells of class diagram are discussed here.

**Data Clump:** It is a commonly occurring model smell that occurs due to recurrence of same set of more than attributes at a lot of places like as attributes in classes or parameters in operation signatures. [1] They really ought to be an object. Illustration of data clump in UML class diagram is below.

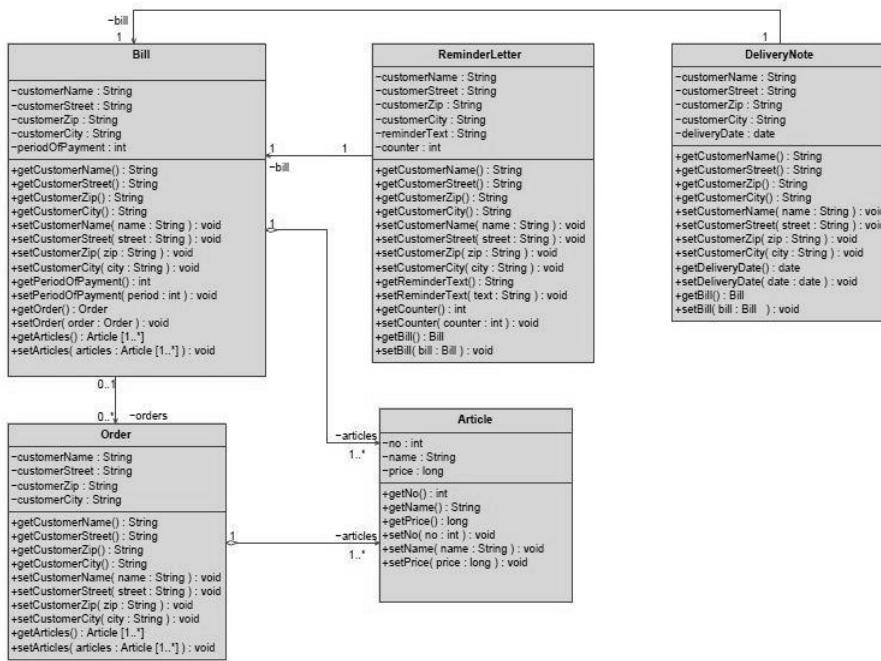


Figure 1 An example of Data Clump

**Swiss Army Knife:** It is a class that is exposing its maximum functionalities by implementing a large number of interfaces. The classes that are associated with the Swiss army knife class provide a lot of public interfaces.

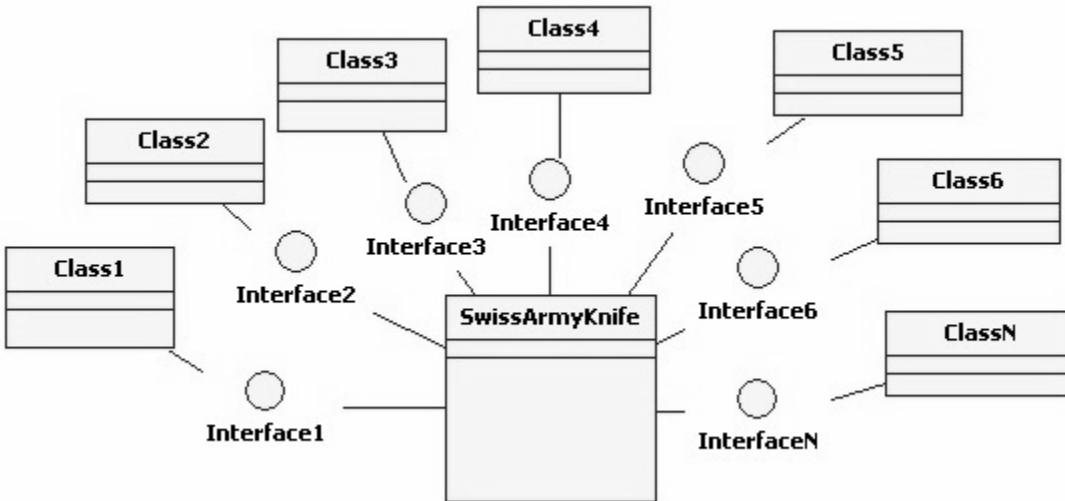


Figure 2 An example of Swiss Army Knife

**Blob:** It is a large class consisting of a large number of attributes and methods. This class is actually highly dependent upon its surrounding classes.

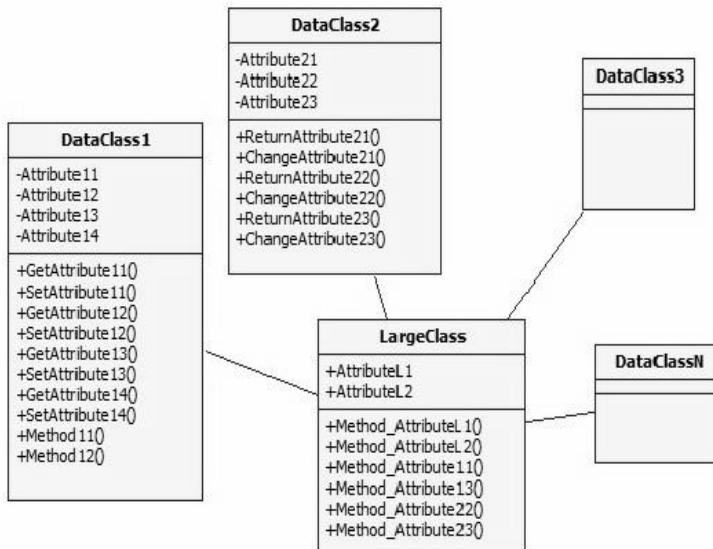


Figure 3 An example of Blob

A lot of works has been done in area of bad smell detection at code level but at model level it is still at infant stage. Very few tools exist for bad smells detection in models like SDMetrics, Magic Draw, EMF Refactor. Various tools that are mentioned in research papers of bad smell detection either does not exist on internet or are developed at academic level and are not industrially approved.

## 2. MOTIVATION

Main motivation behind detection of bad smells is to make the software maintenance and evolution easy and cost effective and for doing it in UML models is increasing importance of UML due to trend of MDA.

Software evolution and software maintenance are now the central activities of software development because most of the software development is done through changes in the form of iteration. A widely references study by has shown that 70% of the total expenditure spent on software is on these two activities. Changes are an inevitable part of software and a costly affair. If software is not of good quality that makes the change difficult and also with changes software quality goes down making the further software changes more difficult and expensive. Process that works towards quality assurance of models and counteract against is Refactoring. Refactoring is the process of improving the internal structure of the system without making any changes in its functionality. Refactoring of models is a two step process that works by first detecting the bad smells in models and then applying appropriated refactoring action. Thus bad smell detection is very crucial to the quality assurance of models. [2]

Model driven Engineering is technique that has become very famous in recent years. It makes the model as main artifact of software development taking the focus away from the code. Model Driven Architecture (MDA) is most famous approach of Model Driven Engineering. MDA is more restrictive than MDE and focus on UML as de facto standard of modeling languages. Thus making UML models the center of software development. With increasing importance of model quality of models has become an important issue too making the bad smell detection crucial activity to software development .Bad smells are the indicators of poor quality of design in artifact thus decided on where to apply refactoring. [3]

### **3. REVIEW OF AVAILABLE TECHNIQUES OF BAD SMELL DETECTION IN UML MODELS**

A lot of work has been done in the field of detection of bad smells using code whereas field of detection of bad smells in models is still in its infancy stage. Development in the automation of model transformation has fueled the interest in this area of research.

Out of the 22 code smells proposed by Fowler in 1999, only 10 are extended and discussed with regards to model refactoring. Although some of these smells cannot be applied to UML models exclusively (such as Long Method, Switch Statements, Comments etc.), quite a few can be addressed by considering other UML views or richer models such as ones augmented with formal specifications. For instance, inter-class smells such as Divergent Change, Shotgun Surgery, Feature Envy and Inappropriate Intimacy can be identified by using message exchange details from interaction diagrams such as sequence diagrams and functional composition information from use case diagrams. [2]

Mohammed Mishaudding, Mohammad Alshayeb (2013), has mentioned that based on the study different research papers there are three main approaches being followed to specify bad smells in models Metric Based, Pattern Based, Rule Based. Metric based approach has advantage that it could be integrated to already existing modeling tools. Its con is specification of appropriate threshold value for metrics as it has decisive influence on detection accuracy. Pattern based approach is based on the concept of first search for design problems within the model and then suggesting corrections for them in the form of design patterns. Rule-Based smell detection approach identifies both model smells and ant patterns using a declarative rule definition. These rules are manually defined to identify the symptoms that characterize the smell. [4]

Ananda Rao, K Narendar Reddy (2008), has mentioned that design smells are used to detect design defects in object oriented software design. Detection of bad smells allows us to imply Refactoring thus improving the quality of the software and making it more maintainable. In existing systems of bad detection there is a lack of quantitative approaches. Quantitative methods are free from any human bias and could be automated. This paper proposes the idea of use of DCPP matrix to detect two important bad smells Shotgun Surgery and Divergent change. These smells are also referred as maintenance smells as their effect on maintenance efforts is noticeable one. A matrix is created on the basis of how a change made in one artifact is propagated to other artifacts. Using these matrix artifacts of design suffering from above mentioned smells are identified. [5]

Anti-patterns are the bad solutions to the design problems that occur repeatedly. Anti-patterns provide the solution to the current problem but their effect on the software quality and maintenance cost is adverse. This approach is combination of pattern and metric based approaches towards detection of bad smell in models. Most of the anti-patterns outlined in literature have been defined in terms of code quality metrics. Detection of anti-patterns at design level offers advantages like improved quality of code, language independent detection approaches and reduced cost of detection of defects during later stages over their detection at code level. [6]

Rahma Fourati, Nadia Bouassida, Hanene Ben Abdallah has mentioned a metric based approach for Anti-pattern detection in UML diagrams in their papers. This approach examines the class diagram and behavioral diagrams of the software system for detection of Anti-patterns at design level. Use of existing metrics and few newly defined has been made for detection of five types of Anti-patterns Blob, Lava Flow, Functional Decomposition, Poltergeists and Swiss Army Knife. [7]

Thesis report by Petra Beczi was the first approach toward the anti-pattern detection in UML class diagrams. He has introduced approach called RADAR, which is a solution to detect *inter alia*, Complex Class, Large Class, Lazy Class, and ManyFieldAttributesButNotComplex (MFABNC) in UML class diagrams and returns warnings of the results. Essentially, RADAR uses a combination of some existing software design (SD) metrics and rules of the anti-patterns, moreover provides a

flexible algorithmic procedure. This approach's accuracy of RADAR in the detection of Large Class is between the averages of 38%-75%, and 55%- 80% for MFABNC. Regarding to the average accuracy of the other two anti-patterns, the detection of Complex Class is determined as between 68%-70%, and 80%-99% for Lazy Class on the basis of survey. [8]

Erlangung and Doktorgrades presented an integrated approach towards assessment and improvement of quality of UML models. They have developed a prototype tool to verify their approach using the UML model studied in the practical course. Study showed that their approach was practically feasible and applicable. [9]

#### 4. Tools

Refactoring and smell detection in UML lacks a standard and industry accepted tool. Most of the tools that are developed for this purpose are academic based, poorly maintained or currently not available. Two tools considered for review are

- EMF Refactor
- SDMetrics

##### 4.1 EMF REFACTOR

EMF Refactor is an open source eclipse tool environment that provides standard quality assurance process for models based on eclipse modeling framework. Tool environment consists of mainly two modules application module and specification module. Application module of emf-refactor provides the facility of calculation of existing metrics, smell and applying refactoring. Specification module allows the generation of new metrics and smells. This module provides us with the facility of generating new metrics and smells using specification technologies like Java, OCL, Henshin [9]. EMF Refactor also facilitates the generation of metric and smell detection reports for further analysis.

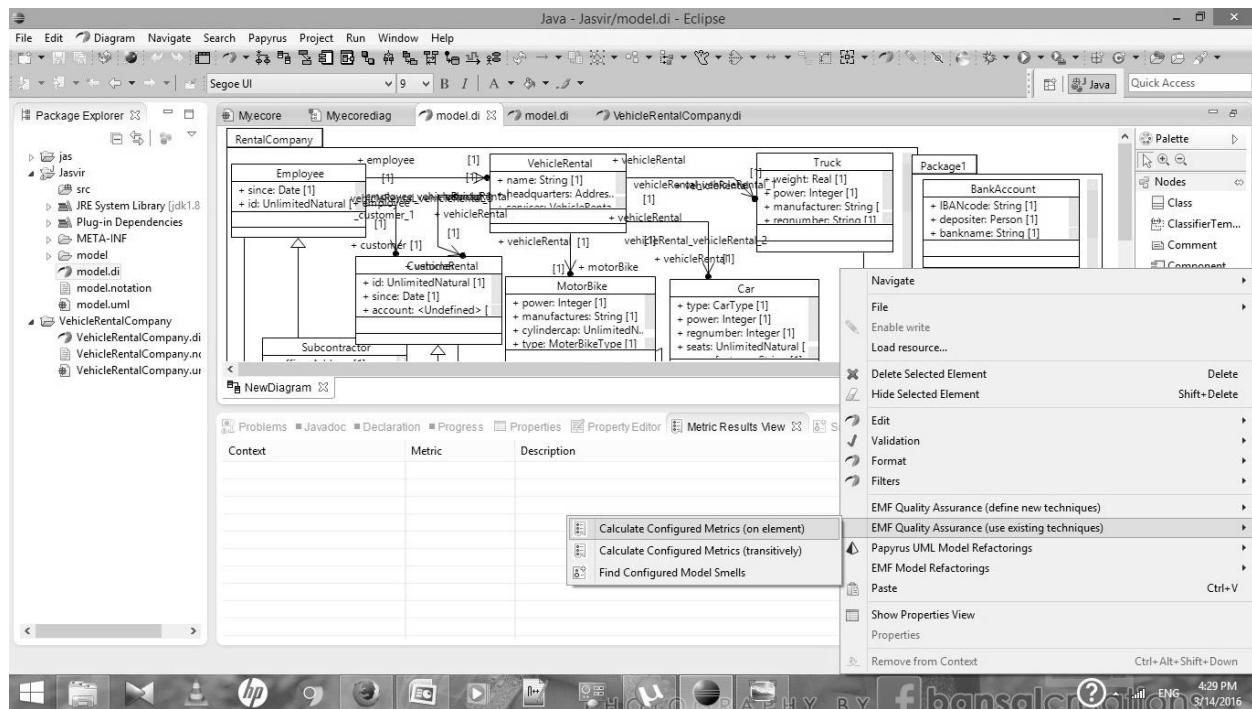


Figure 4 A sample of working with EMF Refactor

#### 4.2 SDMETRICS

SDMetrics is another tool available for smell detection in UML models. In SDMetrics there is no facility of drawing UML diagrams instead xmi file of UML model generated from another tool like ArgoUML etc is imported. SDMetrics tool provides the measure of all design attributes size, coupling, complexity etc. at all level of design like on model, package, subsystem etc. It facilitated the automatic check for design rules. SDMetrics refers to the process of bad smell detection as process of “checking design rules”. SDMetric checks adherence to the UML design rules like DupOps, God class etc. New metrics and design rules can be defined .We are not restricted to existing metrics and design rules. SDMetrics has a very large list of metrics and design rules against which the quality of model is measured. [10]

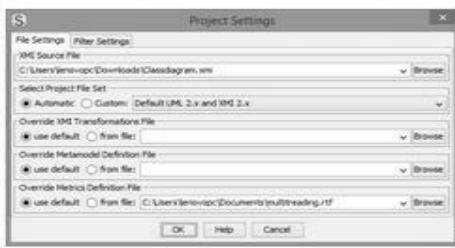


Figure 5 Uploading the XMI file in SDMetric

#### 5. CONCLUSION

The novel contribution of this paper two fold. First is providing the basic understanding of model smells and the need of the detection of model smells. It defines the various issues like recent trend of model driven engineering, costly software maintenance and evolution that has made the detection of bad smells in model an intruding issue to the researchers and individuals in software industry.

Secondly, this paper explains the various approaches and tools that are available for model smell detection. Various tools mentioned in literature are academic prototype and not industrial level. It could be said that issue of model smell detection had not received the attention it is worthy of.

As future work, relationship between UML class diagrams and object oriented languages will be explored in order to apply the approaches of code smell detection for model smell detection.

#### REFERENCES

- [1] Martin Fowler, *Refactoring Improving TheDesign Of ExistingCode.*, 1999.
- [2] Tushar Sharma, Girish Suryanarayana, and Ganesh S G, "Towards a Principle-based classification of Structural Design Smells," *JOURNAL OF OBJECT TECHNOLOGY*, vol. 12, pp. 1-29, 2013.
- [3] Mohammed Misbhauddin and Mohammad Alshayeb, "Towards a Multi-view Approach to Model-driven Refactoring," in *Software Engineering and Applied Computing (ACSEAC), African Conference for Software Engineering and Applied Computing*, 2012, pp. 60-66.
- [4] Mohammed Misbhauddin and Mohammad Alshayeb, "UML model refactoring: a systematic literature review," *Springer Science*, pp. 21-25, 2013.
- [5] Ananda A. Rao and Narendar Reddy K, "Detecting Bad Smells in Object Oriented Design Using," in *International MultiConference of*

*Engineers and Computer Scientists*, vol. I, Hong Kong, 2008, pp. 1-7.

- [6] Thorsten Arendta, Matthias Burhennea, and Gabriele Taentzera, "Defining and Checking Model Smells: A Quality Assurance Task for Models[submitted]," Philipps-Universit'at Marburg, FB12 - Mathematics and Computer Science, Hans-Meerwein-Strasse, D-35032 Marburg, Germany, Marburg, 2010.
- [7] Rahma Fourati, Nadia Bouassids, and Ben Hanene Abdallah, "A Metric-Based Approach for Anti-Pattern Detection in UML Designs," *Springer*, pp. 17-33, 2011.
- [8] PETRA BÉCZI, "RADAR: An Approach for the Detection of Antipatterns," Göteborg, Sweden, 2015.
- [9] Thorsten Arendt. (2014, April) <https://wiki.eclipse.org>. [Online]. <https://wiki.eclipse.org/Refactor>
- [10] (2016, March) <http://www.sdmetrics.com/>. [Online]. <http://www.sdmetrics.com/>
- [11] Anthony Finklestien, "The Future of Software Engineering," *ACM*, pp. 1-4, June 2000.
- [12] (2005, July) [www.omg.org](http://www.omg.org/spec/uml). [Online]. [www.omg.org/spec/uml](http://www.omg.org/spec/uml)
- [13] Throsten Arendt. (2014, April) [wiki.eclipse.org](https://wiki.eclipse.org). [Online]. [https://wiki.eclipse.org/EMF\\_Refactor\\_Architecture](https://wiki.eclipse.org/EMF_Refactor_Architecture)
- [14] Erlangung and Doktorgrades , "Quality Assessment and Quality Improvement for UML models," Göttingen, PhD Thesis 2011.

# Parametric Evaluation and Classification of Water in Punjab, India

Ripudamanjit Kaur<sup>1</sup>, Kanwalpreet Singh Attwal<sup>2</sup>

Computer Engineering Department

Punjabi University, Patiala, Punjab, India

Email-id: [ripudaman3291@gmail.com](mailto:ripudaman3291@gmail.com)<sup>1</sup>, [kanwalp78@yahoo.com](mailto:kanwalp78@yahoo.com)<sup>2</sup>

**Abstract:** Nature's driving force "Water" is essential to humans and pure water is the foremost medicine that leads to healthy living. But this free of cost medicine is rarely available to people of Punjab state, India. Punjab is the land of fertile plains but to obtain better crop yield, intensive agricultural activities are performed that prompted the usage of pesticides and fertilizers which eventually add up in water whereas sewage and industrial disposal is another factor affecting water quality of Punjab. This paramount concern had encouraged to pay heed to evaluate water quality in Punjab state. In this paper, eight potential water parameters namely Temperature, PH, Dissolved oxygen, Conductivity, Biochemical oxygen demand, Nitrite, Total Coliform, Fecal Coliform are used to highlight the prevailing status of water quality in Punjab. The permissible limits of the contaminants are delineated. To exhibit surface water contamination crisis, sampled report having past thirteen year data of 37 spots across four major rivers (Beas, Ghaggar, Sutlej, Ravi) and a lake (Harike Lake) in Punjab has been used. Neuro-fuzzy Technique is used to classify input parametric data into five water quality classes. It has been concluded that Ghaggar river is most polluted and Ravi river is least polluted among four major rivers of Punjab. Latest status reports released by officials, government authorities are delineated in this paper.

**Keywords:** Water Quality, Water quality classifications, Neuro-fuzzy technique, ANN, Fecal coliform, Nitrite.

## I. INTRODUCTION

Life doesn't stand a chance without water hence life cycle and water cycle are one. "Water" is a liquid which makes living on earth possible, but now a day it is becoming a "silent killer" due to its low quality caused by human negligence.

Punjab, the home of five rivers is known as "fertile state" as soil in this area is very fertile and hence produces around two third of the food grains annually in India. Unfortunately today Punjab is facing a huge water scarcity and is walking towards the dark future. The ground water levels are depleting very fast as water level goes down by 10 feet every year and tubewells are bored at between 350 to 500 feet which was earlier at 150 feet. To mitigate this crucial problem farmers should adopt new technologies like drip irrigation, underground piping system and micro-sprinklers so that water table could be maintained.

Their lies another problem of low water quality in Punjab caused due to industrial and agriculture run off, sewage and plastic dumping, excess usage of pesticides and fertilizers, dumping of toxic waste etc in water which lead to diseases and severe health hazards in Punjab. Impure drinking water has become the major cause of deaths in Punjab.. Resent surveys revealed that hepatitis, greying of hair, joint pain; skin disease, mental retardation, asthma and cancer are some of the serious issues arising due to bad water quality.[3]

Usage of high quantity of pesticides, insecticides and fertilizers for better crop yield is deteriorating surface as well as ground water as they have the potential of seeping into the ground. Hence untreated water is not pure and safe anymore especially for drinking purposes. In rural areas mostly people drink untreated water, which leads to severe health issues. Human blood testing in rural areas exposed the presence of pesticide residue in human blood.

In this paper, we are going to highlight the prevailing status of water quality in Punjab and classify it into five classes named Safe, Permissible drinkable, non-drinkable, low quality and hazardous. For this purpose, we are using eight potential water contaminants namely Temperature, PH, Dissolved oxygen, Conductivity, Biochemical oxygen demand, Nitrite, Total Coliform, Fecal Coliform as parameters for water evaluation. The permissible parametric ranges of these parameters for drinking purposes are shown in Table 1.

TABLE 1  
 PARAMETERS WITH THEIR PERMISSIBLE STANDARD RANGES [21] [22] [23] [24]

Parameters	Permissible Limit for Drinking Purposes
Temperature( $^{\circ}\text{C}$ )	12 – 35
PH	6.5 – 8.5
DO (mg/L)	>7.0
BOD (mgO/L)	5.0
Conductivity (millimols)	300
Total Coliform (mg/ 100ml)	<5000
Fecal Coliform (mg/ 100ml)	<100
Nitrite (milligram per liter)	1

To exhibit surface water depletion crises, classifications using Neuro-fuzzy Technique has been done by using database having past thirteen year data and eight parametric values from water samples of 37 spots across four major rivers (Beas, Ghaggar, Sutlej, Ravi) and a lake (Harike Lake) in Punjab has been used.

Neuro-fuzzy is a data mining technique which is a combination of Fuzzy technique and artificial neural network technique. During the classification, it uses ANN for automatic rule generation by applying training and testing method. Whereas Fuzzification which transform crisp values into fuzzy sets between 0 and 1, Defuzzification which convert fuzzy sets into crisp values and decision making by using IF-THEN rules generated by neural network is done by fuzzy logic. Figure 1, represents the Neuro-fuzzy classifier technique used during the work.

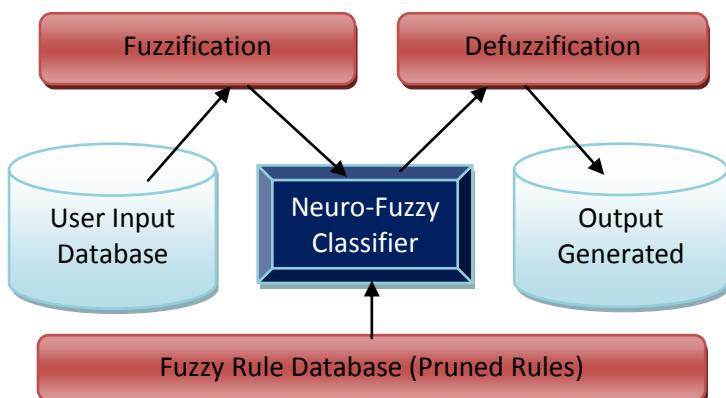


Figure 1: Neuro-Fuzzy Classifier

## II. STATUS REPORTS

Recently NASA sent a warning alarm to Punjab stating that there could be collapse of agriculture output in the state and big cities like New Delhi will have to face severe shortage of drinking water if some serious steps would not be taken to ensure sustainable usage of groundwater.[1] According to state government officials, out of 145 water blocks in Punjab 110 are “Dark Zones” (undrinkable water blocks) and around 40 percent blocks

are “Notified” (drinkables) and these blocks are restricted to be used only for drinking purposes and for agriculture or industrial usage

Jairam Ramesh(Union Rural Development Minister) said, high incidents of cancer and other water borne diseases are seen in Malwa belt of Punjab due to presence of high level of uranium(50% over WHO norms), arsenic, mercury and other heavy metals in water. Punjab is the only state having uranium in water with 1,140 samples out of 2,462 are tested uranium positive but sources are still unknown. [18]

Latest environment report ranked Punjab seventh among the states worst in checking water pollution. Punjab is among worst defaulters having atleast seven industrial units dumping toxic waste directly in rivers. [17] Punjab Pollution Control Board had identified more than 6,200 water polluting industries and decided to file cases against 16 industries in court.[20] Punjab ranks highest in water pollution caused by industries. In Kapurthala, untreated sewage waste is dumped into Kali Bein and Wadala Drain which ultimately fall into Beas River and hence deteriorate water quality in different parts of the State.[19]

The government had issued Rs 3,935 lakh which was advised by NITI Aayog to tackle portable water resource contamination problem by arsenic and fluoride material. Arsenic and fluoride are the contaminants polluting portable save water and depleting the health of individuals of that area.[4]. For water preservation some steps taken by Punjab Local Bodies are:

- Vehicle wash from direct supply line is restricted.[2]
- Watering of lawns will be allowed only after 5pm.[2]
- Average water consumption is 55 liters per person per day but Water Supply and Sanitation department is supplying 40 liters per person per day in the rural areas.[2]
- Penalty and fine have been induced on water wastage which will be Rs 1,000(1<sup>st</sup> time), Rs 2,000(2<sup>nd</sup> time), Rs 5,000(3<sup>rd</sup> time).[2]

In 2008, the government took initiative and passed an ordinance to curb the sowing nursery of paddy. Government is trying water saving techniques such as micro irrigation; permanently raised beds usage for planting; no tillage; use of tensiometers; and direct paddy seeding; use of laser leveller and rain water harvesting on rooftop.

### III. METHODOLOGY ADOPTED

First of all Matlab software has been downloaded and installed in which there is capability and Neuro-fuzzy tool which is used during this work. Flow chart of applied methodology is shown below in Figure 2.

*A). Creation of database for training:* A database had been created to be used to train the neuro-fuzzy network. This database was manually prepared using experts knowledge and WHO and BIS standard values available on permissible drinkable water parametric values. According to the standard values, we have created five water quality classes (1 to 5) Safe water, Drinkable water, non-drinkable water, low quality and hazardous water. Using this database, system recognises the ranges of parameters lying in each class as the database includes the parametric values as well as results in the form of class 1 to 5. Database is assumed surface water database which had Eight parametric values named Temperature, PH, Dissolved oxygen, Conductivity, Biochemical oxygen demand, Nitrite, Total Coliform, Fecal Coliform which corresponds to one class out of 1 to 5.

*B). Training neuro-fuzzy system:* This database is then loaded to train the system. Here neural network part of neuro-fuzzy system trains the system and automatically generates the provided input value based rules.

*C). Saving generated rules:* These automatically generated rules are then saved in .fiz file of Matlab. This rule database will be utilized further to get the results.

*D). Testing system and generated rules:* A database has been created having parametric values whose corresponding results are already known to us. This is done to check whether the results provided matches with the original results. If the results are not accurate the changes are made in the training system or some new entries are added to the system which leads to change in the rules generated to get the accurate results. The changes are made until we get needed outputs.

*E). Collect data and select parameters:* The water quality data has been collected and a database is created. This database includes the surface water parametric values having data collected from water samples of 37 spots across four major rivers i.e. Ravi, Ghaggar, Sutlej, Beas, lake in Punjab. The database used includes water quality data of past thirteen year from 2003 to 2015. Water quality parameters namely Temperature, PH, Dissolved oxygen, Conductivity, Biochemical oxygen demand, Nitrite, Total Coliform, Fecal Coliform are selected for surface water evaluation.

*F). Neuro-fuzzy system:* The system with accurate rule database after training and testing is then ready to be used. This neuro fuzzy system is then provided with databases created to obtain results using the already saved rule directory.

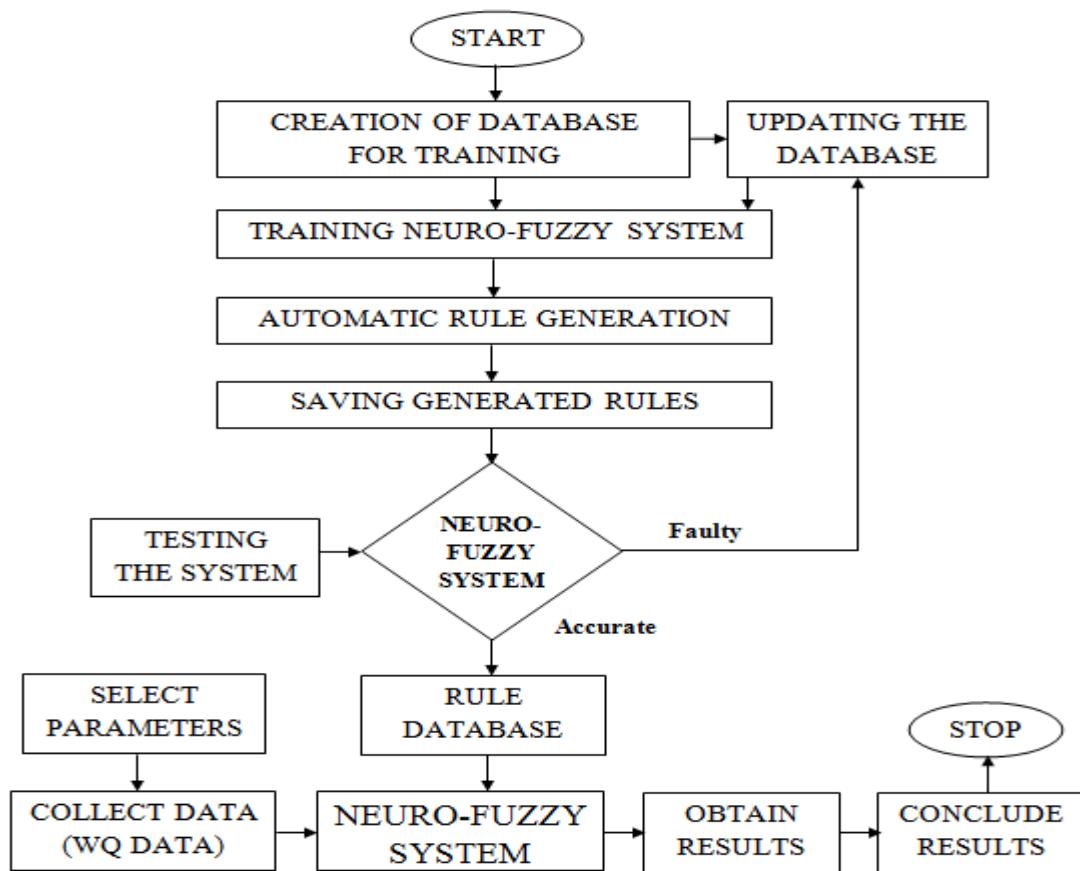


Figure 2: Flow chart of applied methodology

#### IV. APPLIED METHOD

Neuro-Fuzzy classifier converts inputs parameters into membership functions, and then rules generated or rule database obtained after training are used to obtain results. Schematic diagram of proposed method is shown in Figure 3. This method involves various steps to water quality classification using Neuro-fuzzy technique.

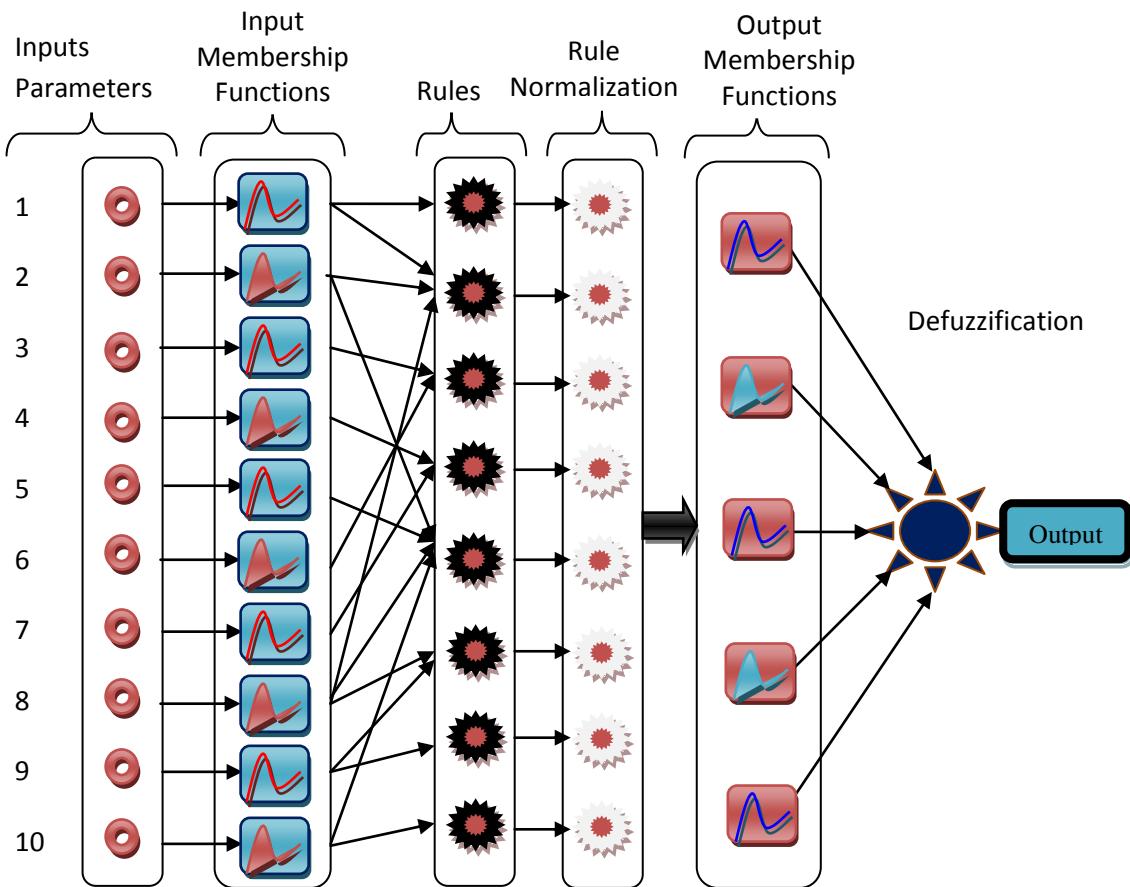


Figure 3: Schematic diagram of applied Neuro-Fuzzy Technique

Block diagram of proposed method is shown in Figure 4. Two water quality databases are used in this process, one for training the system and another for classification of water quality in Punjab. Steps involved are:

*A). System Initialization:* In the first step, system is initialized by loading the directory having ten input parameters and one output parameter. The input combination ranges correspond to five different water quality classes as outputs. These inputs and their respective outputs are manually prepared and are provided to train the system. Artificial neural network is trained and tested using these provided input and respective output values.

*B). Membership functions:* Fuzzy logic converts input and output variables into membership functions. Membership functions are curves which defines how input and output values are mapped between membership values of 1 and 0. Membership functions allow us to represent a fuzzy set graphically. It converts each input variable space into the range of 0 to 1.

*C). Training and Testing ANN:* The directory values are then provided to artificial neural network to train the network where interest keypoint features are extracted. Using two third of the inputs provided, ANN network train itself and make some logics and rules which are used for decision making when real life inputs are provided to the system. This trained system will provide results in next steps. The logics prepared by the ANN system are then tested for their exactitude. The trained system is hence tested for its accuracy and corrections needed are indulged. The ANN system is hence ready for the use.

*D). Automatic Rule Generation:* The ANN generated logics are then converted to the rules, these rules hence generated are in “IF-THEN” format statements. These generated rules represent the trained ANN system. The prepared rules could be further used to calculate results for provided inputs.

*E). Rule Normalization and Rule Database:* The rules generated so far are normalized for further usage over fuzzy set inputs. The normalization rules are then saved in a directory so that it could be further used directly when needed during classification.

*F). Fuzzification:* The provided inputs for evaluation are firstly fuzzified by fuzzy logic. The input value space in the form of crisp values is transformed into fuzzy sets i.e. in the range of 0 and 1. This is done to use normalized rules on input database.

*G). Inference Engine:* This step involves using of rules generated upon the inputs provided. Inference engine uses the rule directory to generate outputs for the provided inputs. The fuzzified inputs are used and rule database is applied over it to generate outputs in the form of classifications. All the decision makings are done in this step.

*H). Defuzzification:* The output generated is then defuzzified to obtain output values in the form of crisp value.

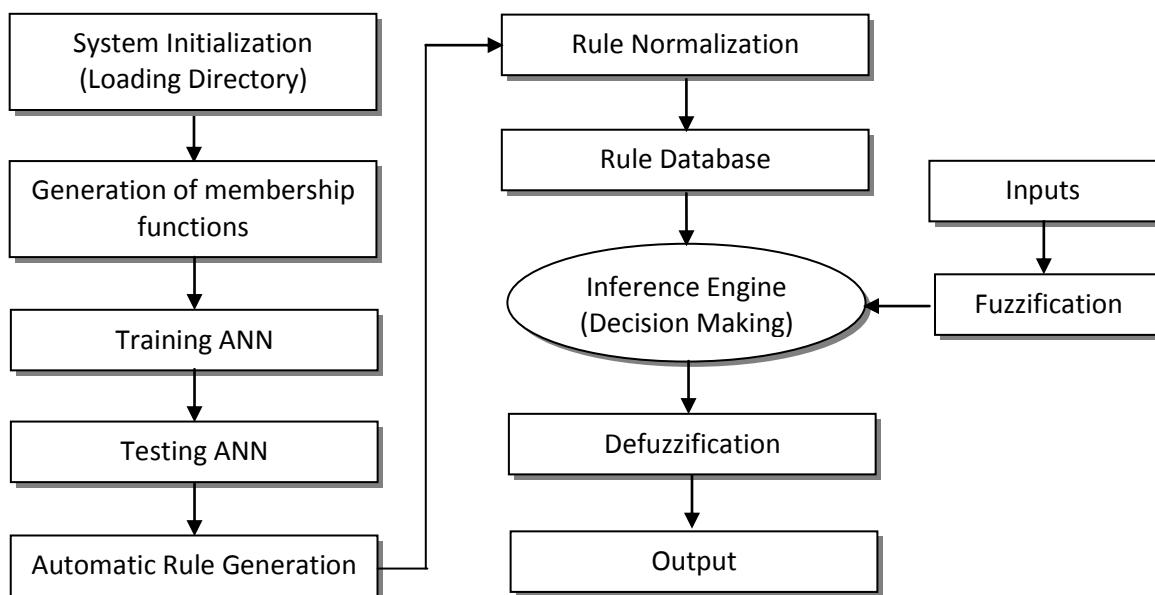


Figure 4: Flow Chart of applied Neuro-Fuzzy Technique

## V. RESULTS AND DISCUSSIONS

The proposed method is applied over database having water quality data of 37 spots across 4 major rivers i.e. Ravi, Sutlej, Beas, Ghaggar and Harike Lake in Punjab state, India. The database has past 13 year data on water quality in of these 37 spots of Punjab. Eight parameters are used as input parameters and water quality of each spot is evaluated.

Neuro-fuzzy classifier is used to classify the water quality in Punjab. It has been clear that water quality in Punjab is deteriorating as the year passes. Some of the conclusions made are:

- Ghaggar River is the most polluted one among the four rivers and it showed continuous increase in levels of pollutants in water.
- In Year 2013, water quality of the areas showed improvements as the pollutant dumping into rivers decreased.
- River Beas, Sutlej and Ghaggar have shown slighter increase of pollutants in water with respect to that in last year. Whereas River Ravi and Harike Lake has been less contaminated as compare to last year.

- All the rivers in year 2015, fall under Class 3 (non-drinkable water) i.e. the water of these rivers are not appropriate for drinking purposes, it needs to be treated first before consumption.
- It has been found that River Ghaggar has the highest level of Fecal and Total Coliform contamination in these years followed by River Sutlej. The presence of these contaminants indicates the sewage disposal and presence of human or animal waste in water.
- Water quality in Punjab has deteriorated from year 2005 to 2009, whereas it has improved from year 2010 to 2013 and again starts deteriorating from year 2014.
- It has been found that Harike Lake and river Ravi has low level of Fecal and Total coliform contamination as compared to others in past years upto 2015.
- River Ghaggar had highest level of Nitrite contamination in years 2010 to 2013, followed by River Sutlej and Harike Lake, whereas River Ravi has lowest nitrite contaminations all those years.

The water quality classification results of four rivers namely Sutlej, Ravi, Beas, Ghaggar and Lake Harike is shown in the graph i.e. in figure 5 bellow. It includes the past thirteen years i.e. from 2003 to 2015.

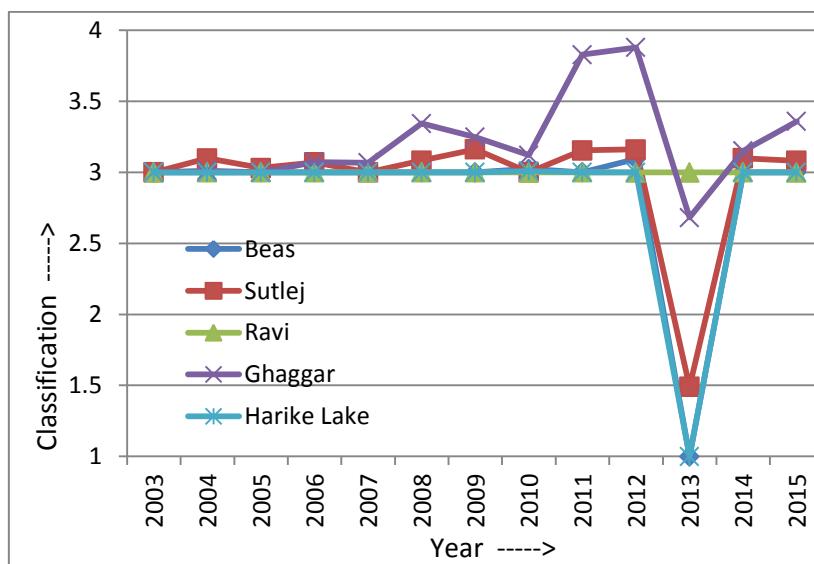


Figure 5: Graph Representing Water classification of rivers in corresponding year.

Table 2 below represents the locations across each river with their station code. Eight Parametric values obtained from water samples collected from these locations are used to obtain results. The database used include past thirteen year water parametric data from these locations but the table below represents the latest year 2015 data classification obtained during the work done.

TABLE 2:

WATER COLLECTION SPOTS AND THEIR AVERAGE CLASSIFICATION OBTAINED FOR YEAR 2015. [25]

Average Classification in 2015	Locations	Station Code
<b>BEAS (3.0000093)</b>	AT TALWARA H/W	1693
	AT U/S PATHANKOT	1694
	AT D/S PATHANKOT	1695
	AT MIRTHAL BRIDGE, GURDASPUR	1010
	AT 1KM.D/S OF EFFL. DIS. PT AT MUKERIAN	1294
	AT G.T.RD UNDER BDG. NR KAPURTHALA	1011
	AT U/S GOINDWAL	1696

	AT 100M D/S INDUST. DISCH. GOINDWAL	1012
	BEAS AT HARIKE,	1697
<b>SATLUJ (3.082192)</b>	AT 100M U/S OF HEADWORKS, NANGAL	1017
	AT 100M D/S ,NANGAL	1018
	AT 1 KM. D/S OF ZENITH	1293
	AT D/S KIRATPUR SAHIB	1814
	AT U/S HEAD WORKS ROPAR	1019
	AT D/S NFL, PUNJAB	1380
	AT U/S BUDHA NALLAH (UPPER)	1690
	AT 100M D/S BUDHA NALA CONFL.,LUDHIANA	1020
	AT BOAT BDG. DHARM- KOTNAKODAR RD, JALANDHAR	1021
	AT D/S EAST BEIN	1381
<b>RAVI (2.999981)</b>	AT U/S HUSSANIWALA - H/W FEROZEPUR	1691
	AT D/S HUSSANIWALA-H/W FEROZEPUR	1692
<b>GHAGGAR (3.3588263)</b>	AT BRIDGE HARIKE, AMRITSAR	1022
	RAVI AT U/S OF MADHOPUR HEADWORKS, GURDASPUR	1097
	AT MUBARAKPUR REST HOUSE (PATIALA)	1023
	AT 100M D/S CONF. WITH R. SARASWATI (PATIALA)	1024
	NEAR BANKARPUR,DERA BASSI	1295
	AT RATANHERI, D/S OF PATIALA NADI (AFTER CONFL.)	1473
	AT D/S CHHATBIR	1698
	AT U/S DHAKANSU NALLAH	1699
	AT D/S DHAKANSU NALLAH	1700
	AT D/S JHARMAL NADI	1701
	AT U/S JHARMAL NADI	1702
	AT MOONAK	1703
<b>HARIKE LAKE (3.0000105)</b>	AT D/S SARDULGARH	1704
	AT U/S SARDULGARH	1705
	D/S FROM CANAL	1382
	AT HARIKE VILLAGE, PUNJAB	1297

## VI. CONCLUSIONS

In this paper, we present report on presence of contaminations in water in Punjab state. The bearable limits of these contaminants in drinkable water are presented. Neuro-fuzzy technique has been used to classify input data having eight parameters into five water quality classes indicating whether water is drinkable or not. Each spot across rivers of Punjab are provided a classification and then its average results are calculated indicating the classification for the River. Some of the conclusions made are:

- Ghaggar is the most polluted and the River Ravi is least polluted among the four rivers of Punjab.
- Neuro-fuzzy technique used is simple, reliable and easy to understand.

This paper also presents the survey of latest government reports, official statements issued, newspaper reports regarding the water quality of Punjab. Some government initiatives and monetary fund issues, devoted to tackle the problem are presented in this paper. More accuracy and better results could be attained by using more parameters during evaluation and classification.

## VII. REFERENCES

- [1]. Sarbjit Dhalwal, ‘Punjab facing a veritable water crisis’, 2015. [Online]. Available: <http://www.tribuneindia.com/news/comment/punjab-facing-a-veritable-water-crisis/124027.html>.
- [2]. The tribune , ’ Now, pay fine for wasting water’, Apr 30, 2016. . [Online]. Available: <http://www.tribuneindia.com/news/punjab/now-pay-fine-for-wasting-water/229816.htm>.
- [3]. Bajinder Pal Singh, ‘In rural Punjab, drinking water is becoming a silent killer: study’, 24 May 2016, [Online]. Available: <http://www.livemint.com/Politics/1IUZGGDAgPDa6w2YLzDBMM/In-rural-Punjab-drinking-water-is-becoming-a-silent-killer.html>.
- [4]. Mukesh Ranjan, ‘Punjab, Haryana get central grant for potable water’, 4 April 2016, [Online]. Available: <http://www.tribuneindia.com/news/nation/punjab-haryana-get-central-grant-for-potable-water/217455.html>.
- [5]. Ranjit Singh Ghuman, ‘Why must Punjab save underground water?’, May 1, 2016, [Online]. Available: <http://www.tribuneindia.com/news/sunday-special/kaleidoscope/why-must-punjab-save-underground-water/230101.html>.
- [6]. Ruchika M Khanna, ‘Canal water for drinking, toxic groundwater for irrigation!’, May 18, 2016, [Online]. Available: <http://www.tribuneindia.com/news/punjab/canal-water-for-drinking-toxic-groundwater-for-irrigation/238301.html>.
- [7]. Central Ground Water Board, ‘Water quality issues and challenges in Punjab’ 01/03/2014, [Online]. Available: <http://www.indiaenvironmentportal.org.in/content/393645/water-quality-issues-and-challenges-in-punjab/>.
- [8]. Xiu Li, Jingdong Song, “A New ANN-Markov Chain Methodology for Water Quality Prediction” IEEE, 2015.
- [9]. Indira Khurana and Romit Sen, ‘Drinking water quality in rural India: Issues and approaches’, WaterAid department.
- [10]. Yan-Qing Zhang, A. Kandel, “Compensatory neurofuzzy systems with fast learning algorithms”, ISSN :1045-9227, IEEE 1998.
- [11]. Mohammad Malkawi and Omayya Murad, “Artificial neuro fuzzy logic system for detecting human emotions”, Springer 2013.
- [12]. Shina Dhingra, Palvinder Singh Mann , “An Adaptive Neuro Fuzzy Approach for Software Development Time Estimation”, ISSN: 2277 128X, IJARCSSE 2013.
- [13]. Zadah L: Fuzzy sets. U.S: National Science Foundation under Grant; 1965.
- [14]. Gao Qianqian; Zhang Ying, "A kind of classification method for evaluating water qualities", IEEE 2010.
- [15]. Asmaa Mourhir, Tajeeddine Rachidi and Mohammed Karim, “River water quality index for Morocco using a fuzzy inference system” Springer 2015
- [16]. Raakhi Jagga, ‘Save water, or Punjab won’t get any after 25 years: Parliamentary panel’ February 14, 2015. [Online]. Available: <http://indianexpress.com/article/cities/ludhiana/save-water-or-punjab-wont-get-any-after-25-years-parliamentary-panel/#sthash.aXLgHqq.dpuf>.
- [17]. Priya Yadav, ‘Punjab ranks high in water pollution by industries’, Jan 30, 2013, [Online]. Available: <http://timesofindia.indiatimes.com/city/chandigarh/Punjab-ranks-high-in-water-pollution-by-industries/articleshow/18249998.cms>.
- [18]. The Hindu, ‘Groundwater contaminated, Punjab battles uranium curse’, July 13, 2012, [Online]. Available: <http://www.thehindu.com/sci-tech/health/medicine-and-research/groundwater-contaminated-punjab-battles-uranium-curse/article3635131.ece>.
- [19]. Umesh Dewan, ‘Govt sleeps as toxic waste poisons water in Punjab’, [Online]. Available: <http://www.tribuneindia.com/2013/20131014/main6.htm>.
- [20]. Indian Environmental Portal, ‘Pollution in Punjab’ [Online]. Available: <http://www.indiaenvironmentportal.org.in/content/28277/pollution-in-punjab/>.
- [21]. Chemical fact sheets, [online]. Available: [http://www.who.int/water\\_sanitation\\_health/dwq/GDW12rev1and2.pdf?ua=1](http://www.who.int/water_sanitation_health/dwq/GDW12rev1and2.pdf?ua=1).
- [22]. US Environmental Protection Agency, ‘Ground Water and Drinking Water’ [Online]. Available: <https://www.epa.gov/ground-water-and-drinking-water/table-regulated-drinking-water-contaminants>.
- [23]. Agency for Toxic Substances and Disease Registry, [Online]. Available: <https://www.atsdr.cdc.gov/>.
- [24]. Water Research Cener, ‘National Secondary Drinking Water Standards’ [Online]. Available: <http://www.water-research.net/index.php/standards/secondary-standards>.
- [25]. ENVIS Centre on Control of Pollution Water, Air and Noise, ‘WATER QUALITY DATABASE’ [Online]. Available: [http://cpcbenvis.nic.in/water\\_quality\\_data.html#](http://cpcbenvis.nic.in/water_quality_data.html#).

# Cloud Based Software Testing

Ekta Rani<sup>#1</sup> e-mail: [ktsan313@gmail.com](mailto:ktsan313@gmail.com)

Er. Harpreet Kaur<sup>\*2</sup> E-mail: [khasria.harpreet@gmail.com](mailto:khasria.harpreet@gmail.com)

<sup>#</sup> Department of Computer Engineering, Punjabi University, Patiala, Punjab, India.

<sup>\*</sup>Assistant Professor, Department of Computer Engineering, Punjabi University, Punjab, India.

**Abstract-** Software testing is an essential step of software development life cycle. Cloud based testing is a type of software testing under cloud environment of web applications to simulate the movement of real world users by using different technologies. Cloud computing has improved the form of services provided to users when they demand it. In this paper, Software Testing steps are explained and focused on different testing techniques in cloud computing environment. Various steps and forms of cloud testing are included. Comparison of Conventional Testing and Cloud Based Software Testing is explored, Cloud based testing is scalable to all levels of testing Instead of conventional that is performed in fixed environment. Cloud based testing cost less because of "Pay as You Test" basis. SOASTA is very powerful tool to perform testing in Cloud by monitoring users virtually.

## I. INTRODUCTION

**Software testing** is an activity to check whether the software package functions meet the requirements defined by user and to ensure that the software system is defect free. It allows the developers to deliver the software that meets business and technical expectations, prevents unexpected results and improves long term maintenance of the software.

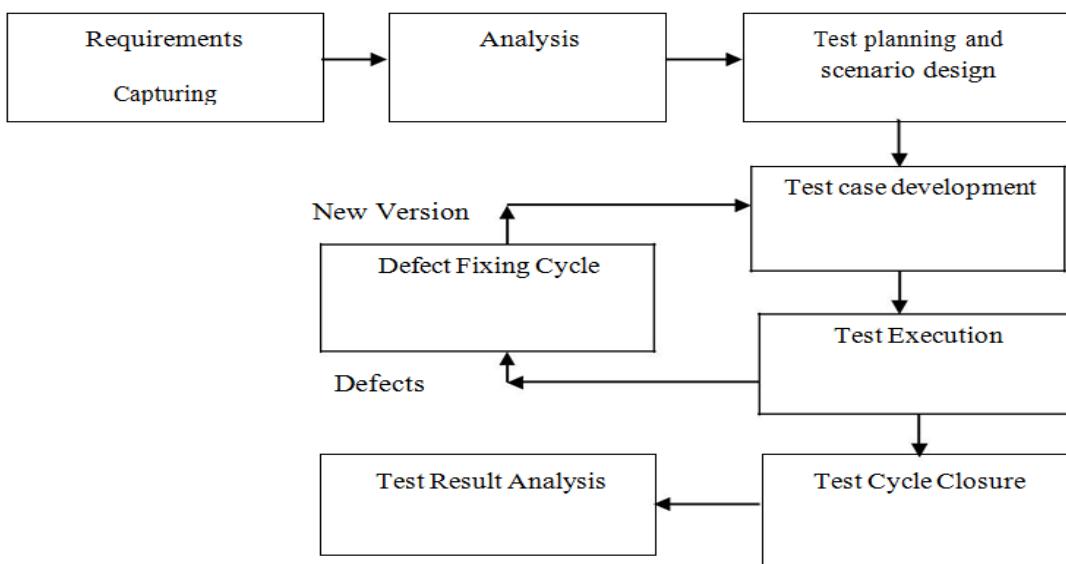
**Cloud computing**, also known as 'on-demand computing' depends on sharing of resources to achieve coherence over a network and provide services to the customer whenever and wherever needed. As a new computing standard, it provides and manages computer resources such as CPU, databases and storage systems in third-party data centers. With Cloud computing, we can store, get any data on internet rather than own physical computer. The cloud can be known as a metaphor for the Internet.[5][12]

**Cloud based software testing** is a form of software testing in which testing is done by using resources of cloud computing infrastructure. It uses cloud services to test the application by minimizing the cost and time with improved product quality. It uses infrastructure based, platform based and software based cloud services and assets for testing the software. Advantages of cloud based software testing include reduced costs from a shared resources and large-scale test environments [8]. Cloud testing provides an end- to-end solution that changes the way testing is carried out and helps an organization in supporting its intensity by reducing the cost of testing without impacting mission critical production applications. Scalability of cloud to test and on-demand service are the major advantages of cloud based software testing that provide powerful real time results. By using virtual resources and a shared cloud infrastructure it offers reduction in cost (hardware and software cost) and time.[5]

## II. SOFTWARE TESTING

Software testing is not a single activity, it is a series of planned tasks that are executed with the software

development activities to make sure that a product is delivered without any error/bug.



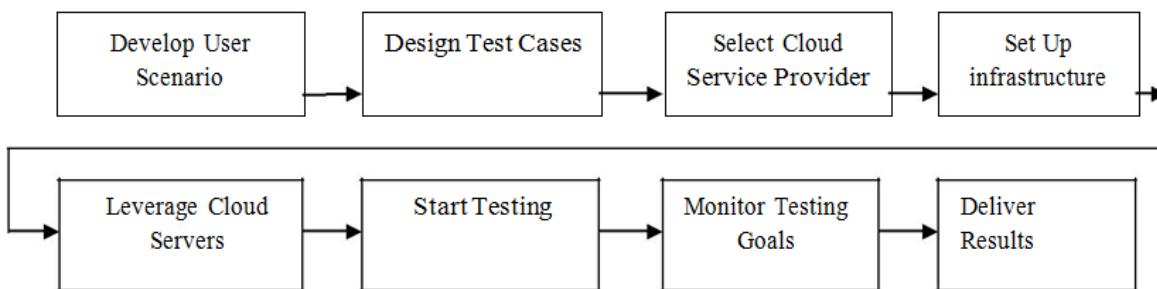
**Figure 1: Software Testing Life Cycle. [3]**

### III. CLOUD BASED SOFTWARE TESTING

It is testing and measurement activities on a cloud based environment and infrastructure by leveraging cloud technologies and solutions. It has four major objectives:

1. To ensure the quality of cloud based applications organized in cloud, with business processes, and system performance as well as scalability based on a set of applications-based system requirements.
2. To validate software as a service (SaaS) in a cloud environment, including software performance, scalability, security and measurements based on certain economic scales and pre-defined a.
3. To analyze the given functional services of cloud. Auto-provisioned function is an example of this type of testing.[3]

To test Cloud Compatibility and interoperation capability between applications in cloud infrastructure and SaaS. Example is APIs and their connection with others.



**Figure 2: Cloud Testing Life Cycle. [3]**

#### IV. FORMS OF CLOUD BASED SOFTWARE TESTING

1. **Testing a SaaS in a cloud:** It ensures the Quality of a SaaS in a cloud based on its functional and Non-functional service requirements.[1]
2. **Testing of a cloud:** SaaS vendors and end users are interested in carrying this testing. It validates the quality of a Cloud from an external view based on the provided cloud specified Capabilities and service features cloud.
3. **Testing inside a Cloud:** Only cloud vendors can perform this type of testing because they have accesses to internal structure. It checks the quality of a cloud from an internal view based on the internal infrastructure of a cloud and specified cloud capabilities
4. **Testing over Clouds:** This testing is performed by the cloud based application system providers. It tests cloud based service applications over clouds including public, private, and hybrid clouds according to requirements [2][3][5].

#### V. CLOUD COMPUTING TESTING TECHNIQUE TYPES

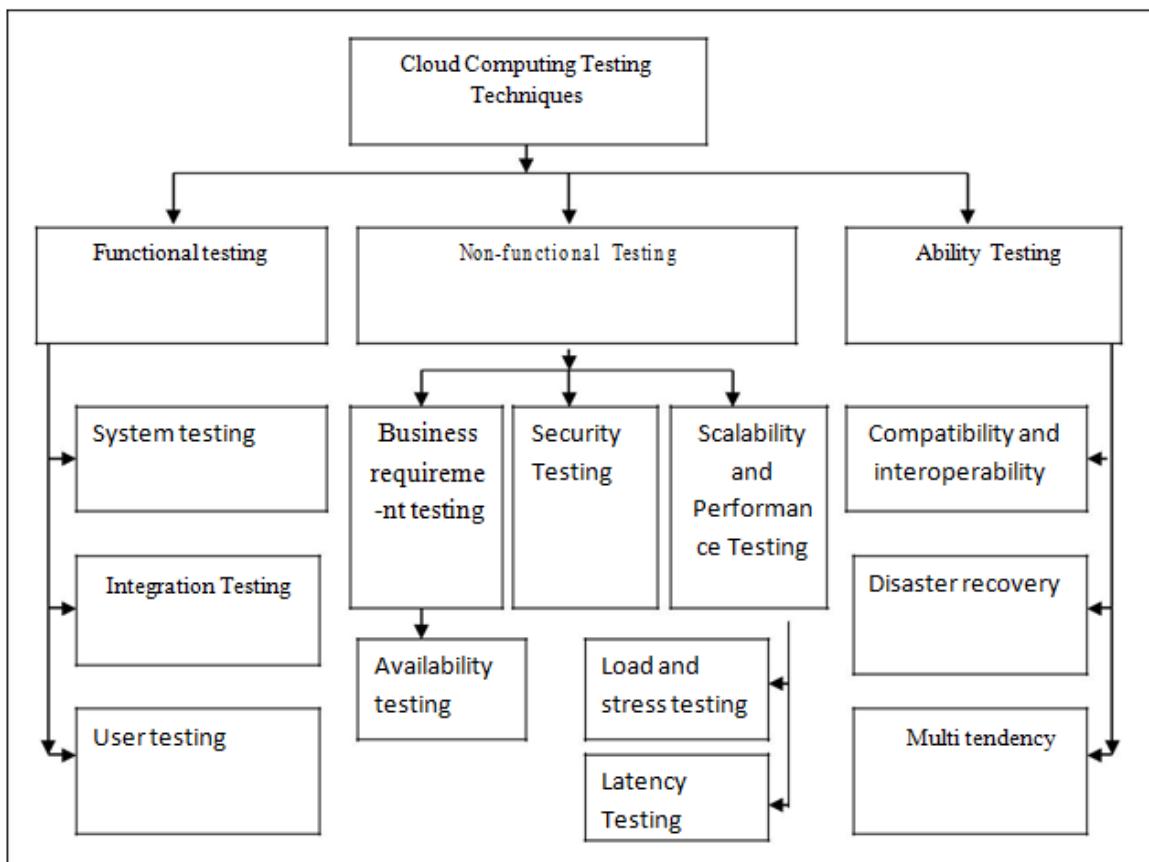


Figure 3: Cloud Computing Testing Techniques. [12]

##### A. Functional Techniques

It includes both hardware and Software features of software being tested. This is performs for both locally and virtually applications.

## **System Testing**

It is testing performed to check that systems functionality that it performs as it designed. When the input given to the system, the system parts work together and it delivers the expected output.

## **Integration Testing**

Integrated parts of system tested to verify that it works within the environment and at last it results the as needed in documents.

## **User Acceptance Testing**

It is a type of testing done to prove that user accepts the developed product cloud solution.

### *B. Non Functional Testing*

It is also known as performance testing because it is done to ensure the performance of application.

#### **i. Business Requirement Testing:**

Organizations migrating their business to a cloud then document their business requirements thus testing performed to verify that business functionalities are according to requirements specified.

#### **Cloud Availability Testing:**

This testing performs to check the availability of services all the time. There should be no down time to affect the business.

#### **ii. Cloud Security Testing:**

This testing performs to increase the security in business so that data remain secure no one outsider can access the business data.

#### **iii. Cloud Scalability and performance testing:**

Cloud scalability testing performed to check the capability of the cloud computing resources and cloud infrastructure. And cloud performance testing allow to measure performance of system.

#### **Cloud Load and Stress Testing:**

Load testing performed by creating large customer traffic and then measuring system's response to load of users. Stress testing done to test the ability of system to work efficiently and maintain the stability.

#### **Latency Testing:**

It is performed to check the latency time between the action of user and response of application in cloud computing.

### *C. Ability Testing Techniques*

This testing is done to make sure that services are given on demand to users.

#### **Compatibility and Interoperability Testing**

Compatibility testing is used to verify that application works and run across multiple environments. Thus it becomes easier to migrating application to cloud computing environment.

#### **Disaster Recovery Testing**

This testing performed to make sure that the service requested by user is not actually possible to available on demand and verification is done that service is back online with minimum time so that does not effect on business.

#### **Multi Tendency Testing**

Testing is performed that multiple users demanding services at a time and services should customized with security level for each client.

## VI. FINDINGS

	Conventional based Software Testing	Cloud based Software Testing
Objectives Of Testing	To ensure the functional quality of system and performance according to given requirements.	To ensure the performance and quality of SaaS, clouds and applications by leveraging The cloud environment.
Test Simulation	Simulate Online Traffic	Simulate virtually Online Traffic
Test Cost	Required hardware and Software Cost, Engineering Costs in testing Process	Pay as you Test Engineering cost in Cloud vendors
Testing Environment	A Pre-fixed Testing Environment in Lab	An open public Test Environment with Computing resources.
Test Execution Time	Offline execution of test Performed in a Lab	On demand test execution by third party
Scalability	Performed in a fixed test environment	Performed in a scalable test environment.

Table 1: Conventional Testing v/s Cloud Based Software Testing. [4][12]

## VII. CLOUD TESTING TOOLS

There are many cloud tools of IT development for various purposes. Cloud tools of different aspects of testing including test management, performance testing and using these tools provide advantages of reducing cost and flexibility. Cloud testing tools are SOASTA Cloud Test, Load Storm, Cloud Test Go, App Perfect, Jmeter etc. SOASTA tools many cloud based testing services and this is the first website testing product creates in industries. Thousands of users can simulate simultaneously by using Amazon service Elastic Compute (EC2). SOASTA CloudTest Lite is a single free test tool for web and mobile performance testing in cloud. This allows us to build and run tests with 100 virtual concurrent users. This tool can run without purchasing an additional license.

## VIII. CONCLUSION

Conventional Software testing results high cost such as expenditure on hardware, software and its maintenance to simulate user activity from different geographical locations. Now Cloud computing become need of benchmark to handle issues and risks with cloud computing environment. The contribution of this paper includes discussion about cloud based Software testing with requirements, benefits and also comparison with conventional testing. Cloud based testing is more advantageous than conventional testing in terms of cost, resources and due to online facility. In future it is getting more matured architecture of cloud computing and thus more testing benefits and testing challenges can explores.

## REFERENCES

- [1] Chaves da Silva, A., Correa, L. R., Vieira Dias, L. A., & cunhs, A. M. (2015). A Case Study Using Testing Technique for Software. *12th International Conference on Information Technology - New Generations* .
- [2] Neurla, E. T., & Sharma, E. G. (2014). Framework for analysing and testing Cloud Based Application. *International Journal of advance research in Computer Science and software Engineering* , 4 (6).
- [3] Malhotra, D. R., & Jain, P. (2013, july-august). Testing Techniques and Challenges in Cloud Computing Environment. *the SIJ transaction on Computer Science Engineering and its Applications(CSEA)* .
- [4] Tsai, W.-T., Bai, X., & Gao, j. (2013). Testing as a Service (TaaS) on Clouds. *IEEE Seventh International Symposium on Service-Oriented System Engineering* .
- [5] Chandane, S. H., & M. Bartere, P. M. (2013). New Computing Paradigm: Software Testing in Cloud, Issues, Challenges and Need of Cloud Testing in today's World. *International Journal of Emerging Research in Management &Technology* .
- [6] kumar, S., & Goudar, R. (2012). Cloud Computing:Resaerch Issue,Challenges,Platform and Applications: A Survey. *International Journal of Future Computer and Communication* , 1.
- [7] Iyer, G. N. Cloud Testing: An Overview. In S. Murugesan, & I. Bojanova (Eds.), *Encyclopedia of Cloud Computing, First Edition*.
- [8] katherine, M. A., & Alagarsamy, D. K. (2012, september). Conventional software testing Vs. Cloud Testing. *International Journal of Scientific and Engineering Research* .
- [9] Bai, X., Chen, B., Tsai, W.-T., & Gao, J. (2011). Cloud Testing Tools. *Software Engineering: an international Journal*.
- [10] Gao, J., Bai, X., & Tsai, W.-T. (2011). Cloud Testing-Issue,Challenges,Needs and Practice. *An International Journal* .
- [11] Malhotra, N. (2010). Cloud testing v/s testing a Cloud. *Intenational Software Testing Conference*.
- [12] (n.d.). Retrieved from the windows club: <http://www.thewindowsclub.com/types-of-clouds-and-cloud-computing>
- [13] Wikipedia. (n.d.). Retrieved from <https://en.wikipedia.org/wiki/SOASTA>

# Available Tools and Techniques For Text to Speech: A Review

Tejinder Kaur<sup>1</sup>, Asst Prof. Charanjiv Singh<sup>2</sup>

Department of Computer Engineering, Punjabi University, Patiala

tejinderkaur236@gmail.com, cjsinghpup@gmail.com

**Abstract-** Text-to-speech(TTS) system converts text into an audible form. The process of converting written text to equivalent message in spoken form. There are different types of languages spoken in India, but each language is the mother tongue of millions of people. In this paper, we are explaining various tools, websites and techniques are available for text-to-speech system.

Keywords: Text to Speech, NLP, Speech synthesis.

## 1. INTRODUCTION

Speech is most frequently used method of communication between humans, text-to-speech(TTS) system convert text into an audible form. It enables you to enter your own text and some of the languages and voices that we offer like Punjabi language. It gives a voice to your documents and making your office offline content more accessible with text to speech.

In actual, language express one's thoughts by means of a gestures and sounds written material is useless for uneducated , Deaf & Dumb peoples. Text to speech is very helpful for the deaf and dumb people now days. NLP researchers aim to gather knowledge on how human beings understand the language and use language so that appropriate techniques can be developed to make computer understand and manipulate natural languages to perform desired task.

## 2. TEXT TO SPEECH

The main function of TTS system is to translate an arbitrary text into a spoken waveform. In computer, the process of automatic generation of speech is called ‘speech synthesis’. The word ‘synthesis’ is defined by the ‘mixing’. Generally alphabets formed as words, and then words formed as sentences. TTS is used in different applications. The TTS system can be used to read text from web pages, news, blogs, talking books, games etc. Synthesized speech should be developed by link together, as through in chain pieces of record speech which is stored in a database.

## 3. LITERATURE REVIEW

Several systems has been developed which convert one language to another language and then that convert text to speech. NLP is directly associated to the human language in which it enables a user to translate a piece of text into a programmer friendly data structure. Synthesized speech should be created by concatenating the piece of recorded speech which is stored in database. Text to speech system is alternatively based on concatenative synthesis approach. In TTS system, we should translate a text into spoken form and for these type of

conversions it involves text processing and speech generation process. Font character mapping is necessary for font characters to of an Indian language to delineate vowel and consonant alphabet. Text analysis is the tone of equate or identify words in the text. Token identification identifies the number, symbols, token to words for which there should be well defined method of pronunciation. Text-pre-processing is generally a composite task and comprehend distinct languages dependent difficulties digits and numerals must be increased into full words. The second objective is to encounter or discover correct pronunciation for various contexts in the text. Some meaningful units of language, called homographs, cause the much hard circumstances in text to speech systems. To discover the correct tones or intonation, strain and time from written text perhaps the much hard problems arrived.

Poonam S. Shetake in this paper demonstrating a text to speech gathering form of linguistic imparting knowledge can be stored as text or data into speech. Text to speech system made it feasible or done in practice to access textual data over telephone, the synthesizer making speech signals of 16 bits, the sampling rate is resolute by the sampling rate of the diaphone data used. A Text to speech synthesizer is a process that can read text in loud voice automatically, which is decoction from OCR (Optical character recognition). Diaphone include the transitions interval two phones, had been selected as the synthesis unit for concatenative synthesizers. There should be about 1500 to 2000 diaphones in English, and the mapping for phoneme string is straight forward.

Shen Zhang a primary factor on emotional audio visual speech synthesis. This approach foremost and first consist of 3 phases: first we take the text and or goal PAD values as input and employ TTS engine to produce neutral speeches. Prosody word boundaries automatically estimate by the text analysis module. In TTS system, ME (Maximum entropy) model is used for prosody words boundaries Prognosis. Colloquialism of human feeling, affect only has gained favoured attention lately in both audio speech synthesis and talking face animation.

M.Macchi proposes an ultimate goal of Text to speech synthesis is to translate a customary orthographic text into an acoustic signal i.e discriminated from human speech. Originally, synthesis system were professionally designed around a system of guidelines and a model that were based on diligent inquiry on human language and speech creation and vision processes. The quality of speech created by a system which is innately limited by the property of the guidelines and the models. Given that our acknowledgement of human voice process is still defective, the quality of TTS is away from natural-sounding. Hence, today's curiosity in high quality voice for application, is associated with advances in computer material or resource, has caused the point of a conic to shift from guidelines and model-based methods to corpus-based methods.

Masatsure tamura in this paper represents a technique for synthesizing speech with any requested voice. This technique is basis on an HMM-based Text to speech (TTS) system and MLLR adapted algorithm. In this approach, the HMM-based TTS system is performed and the voice distinguishing feature of synthetic voice are changed by condition or conversion HMM parameters of the voice units in the MLLR adapted framework. To produce speech with an individual discretion given goal speaker's voice, we remodel the speaker independent models that is average speech models, to the goal speaker. In this paper, we represent a technique of adapted

voice distinguished feature and prosodic form of HMM based text to speech system to an judgement given target speaker.

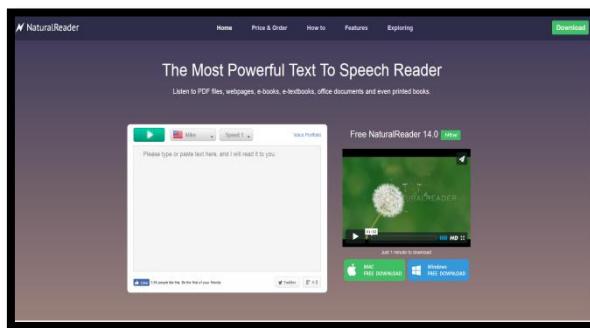
R.Hertz describes a decision of synthesis strategy is the rule-writer's linguistic firmly held belief. For example, the intercessor of an approach based on demisyllables demand that much of the manipulate of contiguous sounds on each other are intervention present in the demisyllable. Even more hard is manipulated phenomena that are not permitted easily be captive in rules.

#### 4. AVAILABLE TEXT TO SPEECH TOOLS

There are various tools available for TTS but most of them are for English language. Some of the tools are :-

A. *Announcify*: Announcify is available online for web browser(for Crome) and for android phones. Announcify reads out various kinds of information, like events or whole documents, for you. If you are tired after study but still want to study more it helps to relax your tired eyes. <http://gdriv.es/announcify> is the link to connect with announcify.

B. *Natural Readers*: It allows you to convert any text for example PDF files, Microsoft word, Emails into spoken form. It is one of the most powerful toll for TTS. It provide facility to change speaker and speed . It is available for Mac and windows system. We can connect with it from the URL <http://www.naturalreaders.com/index.html>



C. *Pediaphon*: It is a free program and allows you to learn during driving and jogging. MP3 files, play lists are producing in a automatic manner from Wikipedia.

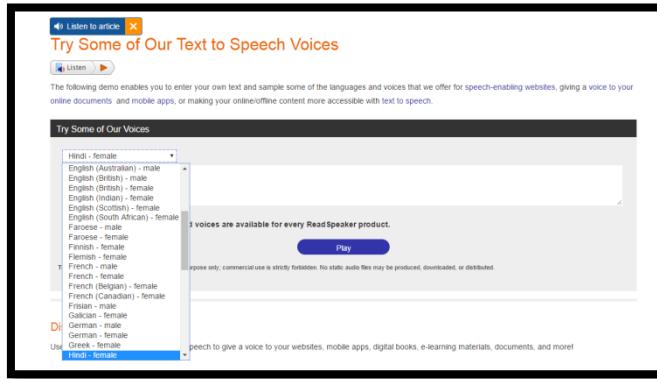
D. *Power Talk*: you have to download and install it and when you open and run your presentation it start speaks text on your slide. It is capable to speak hidden text also.

E. *Select And Speak*: It uses speech human sounding TTS it select text from any website and starting talk. You can also include additional languages by requesting them. Professional versions are also available . User can install and add extension in web browser(Crome)

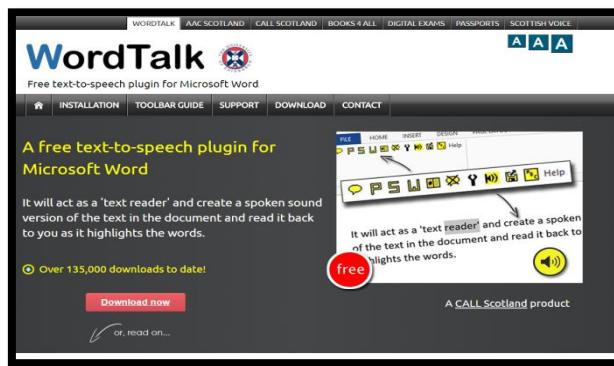
F. *Spoken Text*: It easily translate text into speech. Word files, Microsoft office power point files, recorded PDF files are automatically converted to speech. <http://www.spokentext.net/> provide to login and signup facility to use this.

G. *Oddcast.Com*: it also provides various options for text to speech. User can try it from the URL [http://www.oddcast.com/home/demos/tts/tts\\_example.php?sitepal](http://www.oddcast.com/home/demos/tts/tts_example.php?sitepal)

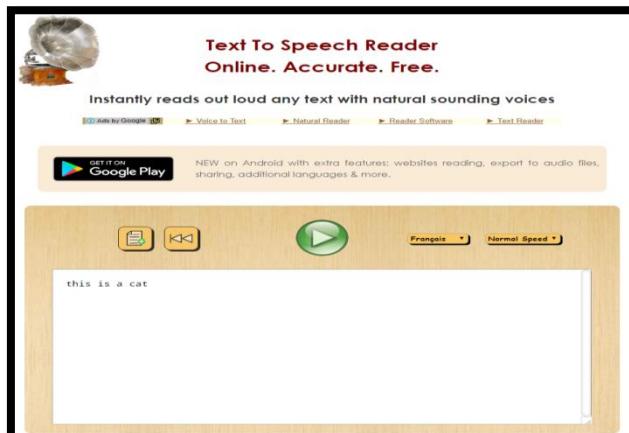
H. *ReadSpeaker*: It text to speech to give a voice to your websites, mobile apps, digital books, elearning materials, documents, and more. It provide facility of various languages and user can use test it from the URL <http://www.readspeaker.com/voice-demo/>



I. *Word Talk*: Word talk speaks the written form of the document and highlight as it goes. It is a free text-to-speech plug-in for Microsoft Word. <http://www.wordtalk.org.uk/home/>



J. *TTSReader.com*: TTSReader is really helpful for busy people who want to be able to listen to written content while doing something else. <http://ttsreader.com/>



## 5. SPEECH SYNTHESIS

TTS describes various high level modules involved in this process:

### A. *Text Normalization*

Input text can be adapted so as to be synthesized. It converts text into a single canonical form. Before processing, normalize text allows for separation of concerns, then input is to be consistent before operations are performed on it.

### B. *Text Segmentation*

It divides written text into meaningful units such as words, sentences. This term has two methods one method is mental processes and second is artificial processes. In Mental processes used by human when reading text and artificial processes implemented in computers.

### C. *Tokenization*

Tokenization substitutes a sensitive data element with a non sensitive equivalent. It divides the text of the sentences at white spaces and punctuation marks. This is completed by a parser.

### D. *Non Standard Words*

Certain abbreviations like Mr. Dr. Phone numbers email or URL addresses need to be expanded into tokens in order to be synthesized correctly. Rules need to be developed for dealing with non standard words.

## 6. SPEECH SYNTHESIS TECHNIQUES-

Different categories of speech synthesis are:

A. *Concatenative Synthesis*: It has 2 phases: a. Offline phase

b. Online phase

Offline phase includes segmentation.

Online phase includes text and analysis.

B. *Articulatory Synthesis*: It uses acoustic and mechanical models to synthesize speech.

C. *Rule Based Synthesis*: It makes the acoustic speech data through rules on the acoustic correlates.

D. *Domain Specific Synthesis*: It is used in calculators and talking clocks. It mainly concatenates pre-recorded sentences to create complete utterances.

E. *Unit Selection Synthesis*: It uses relatively great size data bases of recorded speech and it is dominant synthesis techniques.

F. *Diaphone Synthesis*: It creates a synthetic voice from recording of a particular person.

G. *Sinusoidal Synthesis*: It uses harmonic model and decomposes each frame into a set of harmonics.

H. *Corpus Based Speech Synthesis*: It is able to produce normal speech are generalizations of the combination which are based on dynamic selection of units are based on large quantities of speech data.

#### 7. CONCLUSION

From the various resources we come to know how we can develop TTS for regional languages. We need to develop special rules to create TTS for regional languages. The sound of character is changed according to the position of the character in the word. The planned system will give a easy method to convert text to speech for Punjabi language.

#### REFERENCES

- [1] Angramsing Kayte1, Monica Mundada1 and Dr.Charansing Kayte “A Text-To-Speech Synthesis For Marathi Language Using Festival & Festvox” European Journal of Computer Science and Information Technology Vol.3, No.5, pp.30-41, November 2015
- [2] kaveri kamble, Ramesh kagalar,” translation of text to speech conversion for Hindi language” International journal of sciences and research (IJSR), Vol 3, issue 11, November 2014.
- [3] Kaladharan N, “An English text to speech conversion system” International journal of advanced research in computer science and software engineering, Vol 5, issue 10, October 2015.
- [4] Gurpreet Singh Josan & Jagroop Kaur, “Punjabi to hindi statistical machine translation” International journal of information technology and knowledge management july-december 2011, volume 4, no.2, pp.459-463.
- [5] Amarpreet kaur\* & Er.jyoti rani, “A review on a web based Punjabi to hindi statistical machine translation system” International journal of advanced research in computer science and software engineering august 2014, volume 4, issue 8, ISSN:2277 128X.
- [6] Suhas R. Mache, Manasi R. Baheti, C. Namrata Mahender, “Review on text to speech synthesizer”, International journal of advanced research in computer and communication engineering, vol 4, issue 8, aug 2015.
- [7] Shreekanth.T ,Udayashankara.V ,Arun Kumar.C “An Unit Selection based Hindi Text To Speech Synthesis System Using Syllable as a Basic Unit” IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 4, Issue 4, Ver. II (Jul-Aug. 2014)
- [8] Poonam S. Shetake, S.A.Patil, P.M.Jadhav, ”Review of text to speech conversion methods”, International journal of industrial electronics and Electrical engineering, Vol 2, Issue 8, Aug 2014.
- [9] J.O.Onalapo, T.E.Idachaba, J.Badejo, T.odu and O.I. Adu “A simplified overview of text to speech synthesis”, Vol I, WCE 2014, July 2-4, 2014, London, U.K.

[10] Kaveri kamble, Ranesh Kagalkar, “A review: Translation of text to speech conversion for hindi language”, International journal of advanced research in sciences and Research, ISSN (Online): 2319-7064, Impact factor (2012): 3.358.

[11] Lakshmi Sahu and Avinash Dhole,”Hindi &Telugu text to-Speech Synthesis(TTS) and inter-language text Conversion”, International Journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012, ISSN 2250-3153.

[12] Mrs. S. D. Suryawanshi, Mrs. R. R. Itkarkar, Mr. D. T. Mane “High Quality Text to Speech Synthesizer using Phonetic Integration” International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)Volume 3, Issue 2, February 2014

[13] <http://www.naturalreaders.com/index.html>

[14] [http://www.oddcast.com/home/demos/tts/tts\\_example.php?sitepal](http://www.oddcast.com/home/demos/tts/tts_example.php?sitepal)

[15] <http://www.readspeaker.com/voice-demo/>

[16] <http://www.wordtalk.org.uk/home/>

[17] <http://ttsreader.com/>

## Image Processing Techniques on Agriculture- A Review

**Amrita**

Department of Computer Engineering  
Punjabi University, Patiala  
[kauramrita30@gmail.com](mailto:kauramrita30@gmail.com)

**Dr. Lakhwinder Kaur**

Department of Computer Engineering  
Punjabi University, Patiala  
[mahal2k8@yahoo.com](mailto:mahal2k8@yahoo.com)

### Abstract

The productivity of agriculture crop yield is reducing everyday because of reasons like various kinds of disease, mineral deficiency, etc. To meet the global energy demands of ever growing population the productivity as well as quality of agricultural crop must be improved by using advanced current computer technologies like robotics, computer vision. One technique which is emerging as a useful tool is image processing that has been reported to improve agricultural productivity in number of ways. For agriculture image processing, advanced image processing methods and computer vision techniques are grouped according to specific objective like image acquisition, pre-processing, image segmentation, feature extraction and classification. This paper intends to focus on the survey of application of image processing in agriculture field which covers the detection of diseases, weed, stress severity factors, greenness and crop rows.

**Keywords:** Agriculture, computer technologies, robotics, image processing, feature extraction, classification.

### Introduction

India is an important agricultural country with agriculture being the largest economic sector that plays a major role in its socio-economic development. It has over 210 million acres of farm land where diverse crops are grown by the farmers ranging from cereals (jowar, wheat, barley), pulses, fruits (apples, banana, grapes) and commercial crops (Sugarcane, cotton, chili, groundnut) [1].

However, farmers are facing a lots of problems in the cultivation of these crops due to drastic environmental changes which includes storm, heavy rain, draught besides man-made problems like delay in expert advice of disease, nutrients deficiency problem, limited resources like water, electricity, increasing cost of fertilizers. These problems directly affect the productivity of the

crop yield and consequently cripple the economy of the country. The need of the hour is the timely technological interventions that will help circumvent the problem of crop loss and improves quality of the produce.

The agricultural image processing is one such technique where digital image of agriculture crop is used as input for information extraction from which various decisions can be generated. Various diseases and deficiencies of agriculture crop can be easily detected by just analyzing digital image of infected part like leaves of crop. The concept of precision agriculture that deals with re-organizing the total system of agriculture towards a low input, high efficiency, high profit and sustainable agriculture has received a major boost with the help of image processing technique . The other important applications of image processing are detection of weeds, sorting of fruits in fruit processing, classification of grains, recognition of food products in food processing, medicinal plant recognition, etc. The development of automated system has also helped farmers to avoid consulting divine experts. Therefore computer base technology, particularly image processing has proved to be a boon for the farmers. The paper presents the multiple applications of image processing in the field of agriculture which covers varied aspects like detection of diseases, weeds and severity factors and many more.

### **Digital image processing in agricultural sector**

#### **Disease detection**

The diseases affects different parts of the plant like stem, leaves, stalks, roots and fruits. The diseases are caused by viruses, bacteria, fungi and nematodes. These biological agents invades the plants and causes various types of diseases like molds, rusts, smuts , rots , scabs, blotches and so on. Most of the diseases give conspicuous symptoms which can be easily recognized by the expert of the domain by visual examination or through laboratory tests. The disadvantage however, is the high consultation charges of the experts which are beyond the reach of small and marginal farmers. Hence, there is an urgent need for the automatic detection of the disease at the early stages itself. Table 1. illustrates the various model systems along with their accuracy for the detection of diseases in crops.

**Table 1: Various techniques and their accuracy levels for the detection of diseases in crops**

Techniques	Disease	Applied over	Accuracy and future work	References
Support Vector Machine (SVM) and k-mean Clustering	Bacterial blight (telya)	Pomegranate disease	Accuracy- 82% Training the system to detect diseases for other fruits, increase dataset size to improve the overall system performance to detect diseases more accurately	Bhange <i>et al.</i> (2015) [2]
<b>Fruits-</b> Canny edge detector, Median filter, Grey level co-occurrence matrix, Grey level run length matrix, Block wise nearest neighbor	Anthracnose, Powdery mildew, Downey mildew	Mango, Grape ,Pomegranate,	Accuracy: Grey level co-occurrence matrix (GLCM)- 91.37 % Grey- level run length matrix (GLRM) - 86.715%. The average classification has increased to 94.085% using block wise features	D.Pujari <i>et al.</i> (2014) [1]
<b>Vegetables-</b> Chan- vese, Local binary patterns, k-nearest neighbor, neuro- KNN	Anthracnose, powdery mildew, rust, downey mildew, early	Beans, Bengal gram, Soyabean, Sunflower and Tomato	Accuracy: Artifical neural network (ANN) - 84.11% Neuro K- nearest neighbor (KNN)	

<b>Crops</b> – Grab-cut, Principal component analysis, probabilistic Neural network , Imfilter, k-mean clustering, RGB , HSI boundary descriptors , SVM	blight, late blight Anthracnose, Powdery mildew, Fruit rot, Altemaria leaf spot, Fusarium wilt, Gray mildew, Smut, Red rot	Chili, cotton & Sugarcane	classifier- 91.54%  Accuracy: Mahalanobis distance classifier: 83.17%  Probabilistic neural network (PNN) classifier : 86.48%	
<b>Cereals</b> - Hue saturation intensity (HSI), Color co-occurrence matrix (CCM), SVM	Leaf blight, Leaf spot, Leaf rust, Smut	Jowar, Wheat, Maize	Accuracy: HSI color features : 85.33%  HS color features : 91.16%	
Color filtering , Edge detection using homogeneous technique , RGB color feature image segmentation	Cotton leaf spot disease – Foliar, <i>Fusarium</i> built, <i>Verticillium</i> wilt, Root rot, Boll rot, Gray mildew, Bacterial blight, Leaf curl	Cotton leaf spot disease	Accuracy- 89.5%	Revathi <i>et al</i> (2012) [3]

Support vector machines (SVM), Spectral vegetation indices	Powdery mildew, Sugar beet rust, <i>Cercospora</i> leaf spot	Sugar beet disease	Accuracy- 97% Robust and economic sensors for practical use may be developed in future	Rumpf <i>et al</i> (2010) [6]
Partial-least-square discrimination analysis	<i>Ganoderma</i> disease	Oil palm canopy	Accuracy- 92% Fitting of new protocols to airborne or satellite-borne hyperspectral data to calibrate a dedicated model that would take into account the imaging specificities	Lelong <i>et al</i> (2009) [4]

### Weed detection

The pressing demand of the world is to use minimum herbicide applications in order to protect the environment from their adverse effects. The solution to this problem lies with the fact that the herbicide applied at the right time and place and according to the distribution as well as the stage of the weed ensures minimum dosage of herbicide. Therefore, weed detection is an integral step for the precision application of herbicides. It has been demonstrated that the correct spraying technology and decision support system for precision application of herbicide can cut upto 75% of its dosage [10]. Table 2 presents the different model systems used for the detection of weeds and their accuracy levels along with the future works.

**Table 2. Various techniques used for the detection of weed**

Techniques	Applied over	Accuracy and future work	References
Robust crop row	Crop/weed	Accuracy- weed- 95%	Xavier p. Burgos

detection algorithm(RCRD),Fast image processing(FIP)	discrimination in maize fields	Crop -80 %  Direct testing of the systems on real fields, feeding its output to the fuzzy controller developed by the group which determines the optimum herbicide dosage using expert knowledge and controls the spraying	(2011) [22]
Probabilistic neural network classifier (PNN)	Discrimination between corn seedlings and yield	Accuracy- Corn seedlings -92.5%  Weed-95 %  Upgradation of Back propagation (BP) network to match that of PNN network	Chen <i>et al</i> (2010) [10]
Poisson process, Neyman-scott process, Gabor filtering, Hough transform	Measure and compare the effectiveness of two algorithms for the discrimination of crop and weed	Accuracy: Hough transform- 90%  Gabor filter-75%  To upgrade the accuracy of Gabor filter to 90% which is equal to accuracy of Hough transform.	Jones <i>et al</i> (2009) [23]
Double Hough transform (DHT), Region based segmentation using blob coloring analysis,	Discrimination of crop and weed	To optimize the detection step in Hough space in order to reduce the processing and cost allowing of the method for real-time agricultural applications.	Gee <i>et al</i> (2008) [8]

### **Stress severities detection**

The term ‘stress’ is used to signify any effect on the plant which could have detrimental effect on its growth. The plant flourishes, if all the conditions viz. water, nutrition and environment factors are available at optimum level and the plant is free from any type of attacks by disease causing agents. Any deviation from the optimum levels result in stunted growth and poor yield. The literature survey suggests that the remote sensing technique have emerged as potential tool for the retrieval of plant developments, evaluation of growth processes and yield forecasting.. The various useful techniques in this context are described in Table 3.

**Table 3. Various techniques used to detect Stress Severities**

<b>Techniques</b>	<b>Applied over</b>	<b>Accuracy and future work</b>	<b>References</b>
Visible-near infrared (VNIR) and thermal infrared (TIR), tasi and casi, which are regularly operated by the institute Cartografic de catalunya (ICC)	Crop water stress characterization	A quantitative validation of the method based on empirical relationships between leaf water potential and absolute values of crop water stress index (CWSI) represents the focus of future studies.	Pipia <i>et al</i> (2012) [15]
Regression approach using State vector machine	Continual disease severity in wheat	Results show that the regression analysis using support vectors can achieve satisfactorily results but a good variety of the training data is needed in future	Mewes <i>et al</i> (2010) [12]
Acoustic emission technology, 1)pci-2 acoustic emission board and r15 acoustic	Detection of crop disease stress	Determination of relationship between acoustic emission signals, disease degrees and physiological state of the crops that forms the basis of investigation for disease	Yang <i>et al</i> (2009) [13]

emission sensor probes were chosen to construct the hardware detecting system 2) the aewin software and virtual instrument technology were utilized to construct the software system technology		forecasting and spraying of agriculture pesticides.	
Spectral mixture analysis	Crop stress detection	Spectral mixture analysis (SMA) is particularly suitable to estimate disease severity due to satisfactory coefficients of determination up to $R^2=0.64$ .	Franke <i>et al</i> (2009) [11]

### Greenness identification

Crop growth monitoring is an essential task since it plays an indispensable role in precision agriculture. The knowledge of the growth condition will help to determine the relationship between the crop growth processes and the growth conditions to ensure effective applicability of agricultural services such as irrigation, fertilizers, sowing, harvesting and so on. The important way to monitor the plant growth is to determine its greenness. The methods which are commonly used for the identification are based on visible spectral-index, such as the excess green index, the excess green minus excess red index, the vegetative index, the color index of vegetation extraction, the combined index. All these visible spectral-index based methods assume that plants display a clear high degree of greenness, and soil is the only background element. Remote sensing and camera based observation are the important tools employed for crop growth monitoring. The remote sensing uses satellites and aero planes to capture the images while camera based observations use digital cameras for the purpose. It has been reported that remote

sensing is a more effective in capturing large scale analysis of crop growth as it possesses limited spatial resolution of the sensors ( Yang *et al* 2015 ).Table 4 gives a glimpse of model system for the detection of greenness.

**Table 4. Various techniques used to detect Greenness identification**

Techniques	Applied over	Accuracy and future work	References
Hue, saturation and Value decision tree	Greenness identification for maize seedlings	Poor results are obtained in extreme light conditions which impedes the Identification process. Development of technology to obtain accurate results in all weather conditions.	Yang <i>et al</i> (2015) [16]

### Crop row detection

The approach of field navigation with the help of agricultural robots for various purposes like sowing, irrigation, spraying can only be possible if crop rows are detected accurately. The interest in this approach has widened greatly with the advent of machine vision based detection systems. Various methods of robust recognition of crop row detection systems are presented in table 5.

**Table 5. Various techniques used to detect crop rows**

Techniques	Applied over	Accuracy and future work	References
Binarization, Linear least square method, Clustering method , Linear regression method	Crop rows	Detection rate- 93% Accuracy- $0.00023^0$  Extension of color indices beyond green plants; designing of intelligent expert to cover	Jiang <i>et al</i> (2015) [17]

		combination of all the categorized algorithms; The fusion of Global Positioning System (GPS) and inertial sensors to help with tasks such as end-of-row detection.	
Least square regression method	Crop rows	Detection of row positions with higher accuracy and speed	Zheng <i>et al</i> (2014) [18]
Vertical projection method, Hough transform	Crop rows	Effective extraction of crop rows based on too close or too narrow interspace row wheat canopies	Jiang <i>et al</i> (2010) [24]
Least square regression method	Crop rows	Algorithm can be extended for the selection of size and shape of structural element relating to the practical application scenarios in order to obtain the best results.	Jinlin <i>et al</i> (2010) [20]
Gradient based Hough transform algorithm, randomized Hough transform	Crop rows	Improvement in the current speed of HT (Hough Transform), i.e. 1.715s detection system to detect a 400* 300 color image which is less than RHT (Random Hough Transform) i.e. 0.802 s.	Ji <i>et al</i> (2011) [21]

## Conclusion

The emergence of precision agriculture is ascribed to the vast expansion in computer and sensor based technologies. These novel technologies have paved the way for the modernization of agricultural sector. The most important of such techniques is digital image processing. The paper presents the review of digital image processing techniques in the field of agriculture that includes its application in detection of weeds, diseases, stress severity, greenness and crop rows. All these applications helps in the efficient implementation of agricultural practices. While the existing technologies suffice the needs of current times, there still are certain aspects that have to be

looked upon. The major set back is the absence of online images databases on food quality assessment, fruit defects detection or weed/crop classification. So there is a dire need of agricultural database that will expedite the development process and thus contribute towards achieving the wider goal of sustainable agriculture and food security.

## References:

- 1) D. Pujari *et al*, “ Image Processing based detection Fungal Diseases in plants”, *International conference on information and communication technologies (ICICT)*, pp. 1802-1808, 2014
- 2) Bhange *et al* , “Smart Farming: Pomegranate Disease Detection Using Image Processing”, *Second International Symposium on Computer Vision and the Internet* , pp.280-288, 2015
- 3) Revathi *et al*, “Advance Computing Enrichment Evaluation of Cotton Leaf Spot Disease Detection Using Image Edge detection”, *IEEE*, 2012
- 4) Lelong *et al*, “Discrimination of fungal disease infestation in oil-palm canopy hyperspectral reflectance data”, *IEEE*, 2009
- 5) Barbedo *et al*, “Digital image processing techniques for detecting, quantifying and classifying plant diseases” *Springer*, pp.1-12,2013
- 6) Rumpf *et al*, “Early detection and classification of plant diseases with Support Vector Machines Based on hyperspectral reflectance”, *Elsevier*, pp.91-99,2010
- 7) Camargo *et al*, “An image-processing based algorithm to automatically identify plant disease visual symptoms”, *Elsevier*, pp.9-21, 2009
- 8) Gee *et al*, “Crop/weed discrimination in perspective agronomic images”, *Elsevier*, pp.49-59,2008
- 9) Artizzu *et al*, “Real-time image processing for crop/weed discrimination in maize fields”, *Elsevier*, pp. 337-346, 2011
- 10) Chen *et al*, “Weed identification method based on probabilistic neural network in the corn seedlings field” *Ninth International Conference on Machine Learning and Cybernetics, IEEE*, pp. 1528-1531

- 11) Franke *et al*, “requirements on spectral resolution of remote sensing data for crop stress detection”, *IEEE*, pp.184-187,2009
- 12) Mewes *et al*, “Derivation of stress severities in wheat from hyperspectral data using support vector regression”, *IEEE*, 2010
- 13) Yang *et al* , “Study on detecting system of crop disease stress with acoustic emission technology”, *IEEE*, pp.107-111, 2009
- 14) Kancheva *et al* , “Plant Spectral Signatures as Growth Stress Indicators”, *IEEE*, pp.355-360,2006
- 15) Pipia *et al*, “Simultaneous usage of optic and thermal hyperspectral sensors for crop water stress characterization” *IEEE*, pp. 6661-6664, 2012
- 16) Yang *et al*, “Greenness identification based on HSV decision tree”, *Elsevier*, pp.149-160, 2015
- 17) Jiang *et al*, “Automatic detection of crop rows based on multi-ROIs”, *Elsevier*, pp. 2429-2444, 2015
- 18) Zheng *et al*, “Multi-crop-row detection based on strip analysis”, *International Conference on Machine Learning and Cybernetics, IEEE*, pp.611-614, 2014
- 19) Jiang *et al*, “ A machine vision based crop rows detection for agricultural robots” , *International Conference on Wavelet Analysis and Pattern Recognition, IEEE*, pp.114-118, 2014
- 20) Jinlin *et al*, “Vision-based Guidance Line Detection in Row Crop Fields”, *International Conference on Intelligent Computation Technology and Automation, IEEE*, pp.1140-1143
- 21) Ji *et al*, “Crop-row detection algorithm based on Random Hough Transformation”, *Elsevier*, pp.1016-1020, 2011
- 22) Xavier P. Burgos *et al*, “Real-time image processing for crop/weed discrimination in maize fields”, *Elsevier*, pp. 337-346, 2011
- 23) G. Jones *et al* , “Modelling agronomic images for weed detection and comparison of crop/weed discrimination algorithm performance, *Springer*, pp.1-15, 2009
- 24) Jiang *et al*, “A machine vision based crop rows detection for agricultural robots” *IEEE*, pp. 114-118, 2010

# Swarm intelligence based routing algorithm

Anandika Sharma<sup>#1</sup>, Dr. Amardeep Singh<sup>#2</sup>

<sup>#1</sup>Department Of Computer Engineering  
Punjabi University Patiala, Ucoe, India

<sup>1</sup>anandika.sharma7@yahoo.com

<sup>2</sup>amardeep\_dhiman@yahoo.com

## Abstract

**Swarm intelligence is a field that emerged from biological research, sometimes referred to as an group of individual agents (example - ants, bees, termites, bacteria, fish, birds) for the optimization. Swarm intelligence has one of its applications in network routing. Due to the network complexity it should be self organized, self configured such as ants in nature are self organized and are adaptive to changes, this kind of system is referred with the term swarm intelligence .In this review paper we have explained the routing algorithms for wired and wireless communication based on swarm intelligence such as ANT-NET , ANT BASED CONTROL (ABC) and MOBILE AD HOC NETWORK (MANET). This paper has done comparative analysis of different network routing approaches, in swarm intelligence based on ant colony optimization routing algorithms which increase the performance of network example: end-to-end delay and delivery ratio, and the reliability of wireless communication.**

**Keywords:** ant-net, ant based control, ant-hoc net and ad-hoc networks.

## I. INRODUCTION

Routing is the important aspect of computer overall performance of the network system. There are many routing algorithms proposed for the network whether in wired or wireless communication network. Due to the increasing network facilities swarm intelligence take place in the field of networking. It is a soft-computing technique that gain attention over the last couple of years (Schuster, 3|2005.). Swarm intelligence is the nature inspired algorithm, defined as the collective behavior that emerges from a group of social insects like ants, bees, termites has been dubbed “SWARM INTELLIGENCE” (Meyer). Swarm intelligence tackle, the problems in network routing due to increasing size, changing topology and complexity of communication network. Traditional routing techniques either fail or face complexity to challenge the new features of routing , so swarm based routing algorithms has been developed to cooperate the challenges. A number of swarm based algorithms has been developed but in this paper we considered ‘ant’ based routing algorithm for our study which uses stigmergy (stigmergy is at work system protocol plays a prominent role with respect to modules like agents) (Csete ME, 2002). In section-1 we explain the swarm intelligence characteristic and general principle of ant based routing that how ant based algorithm can work. In second section we give an overview of ant first routing algorithm Ant-net, in section three we explain the ABC (ant based control) which is similar to Ant-net. In section four we explain the ad-hoc network (MANET) which is wireless in nature and is most popular in communication network these days. And at last we conclude this paper in short and give the overview of future scope.

## II. SWARM INTELLIGENCE

As explained earlier swarm intelligence is the group of agents called ants which communicate with each other to find the optimal solutions. Swarm intelligence is the method to solve the static and dynamic optimization problems (Stojmenovic). Ant inspired routing algorithms have been developed both for wired and wireless channels to find the superior results. The algorithm performed better than OSPF and a distributed Bellman Ford with dynamic metrics. Swarm is considered as the group of insects called ants having following characteristic are (Stojmenovic, “swarm intelligence for routing in ad-hoc wireless network”):

*Scalability:* The ants can change their network size according to the desired level.

*Fault tolerance:* The ants cannot follow the central control mechanism hence it is independent of each other movement.

*Adaptive:* Ants can easily adapt the new path or the shortest path from source to destination.

*Speed:* Ants move faster to find the destination and alert the other neighboring ant's to move communicate faster.

*Modularity:* Ant's follow high intensity value of pheromone, as they work independently to accomplish the task.

These are the advantages to use the ant based colony optimization algorithm. This algorithm work efficiently with main principle of depositing pheromone. Pheromone is the chemical substance deposited by the ant's when move from source to destination. The ants independently move from source to destination, different ants can follow different routes, the shortest path covered by the ant can followed by the other ant's due to the pheromone deposited by the ant's. More ants can follow the shortest path and increase the intensity of pheromone as described in the figure. Pheromone gets evaporate due to time so the pheromone on other routes gets evaporated with time and hence shortest path is detected with greater intensity of pheromone value laid by the ant's in network.

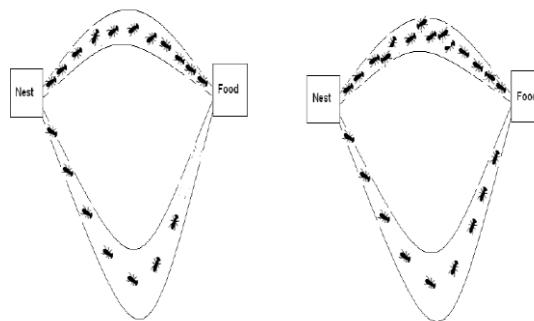


Figure1: Ant's find the shortest path from source to destination

The different algorithms can be formed based on ant colony which increase the performance of the network system and help the network system to control the group behavior of data packets. Swarm intelligence network routing algorithms will enhance the reliability and timeliness of a data transfer in a wireless communication. It also reduce the overhead in network due to scalability feature of ants.

**1. ANT-NET**

Ant-net is the first routing algorithm based on ant colony routing for wired networks. This is proactive algorithm approach based on packet-switched networking. In ant-net there is a forward and backward ant in which ant can traverse the network node to node and updating the routing metrics based on probability distribution function from the trip times of the route measured by the ant. In this each node send a forward ant (FANT) packet to the destination. The FANT record the path as well as the time needed to arrive each intermediate or neighboring nodes. When FANT reach their destination with same priority of all nodes, the backward ant (BANT) can move from destination to source with highest priority node. Each intermediate node can update the routing table with the information given by the BANT. Routing tables can contain per destination next hop biases information for the faster routes and greater efficiency (Rajeshwar Singh).

**2. ANT BASED CONTROL**

Ant based control is the second routing algorithm for wired network used for routing in telephone networks. ABC is also proactive in nature and is based on circuit-switched network. ABC is same as Ant-net but dissimilar in some issues as ABC has only one kind of ant known as forward ant (FANT). In ABC the routing table can be updated at the time of forward ant reaches the intermediate nodes and are based on the life of ant at the time of visit of FANT to each node (Stojmenovic, “swarm intelligence for routing in ad-hoc wireless network”). This algorithm is not based on pheromone as laid by the ant but call setup messages. These messages do not based on the probability distribution function for the selection of path but choose the path having higher intensity value. This algorithm has the same philosophy of routing table but the updation of routing table is done by FANT for each visit. This algorithm is highly adaptive and robust under various conditions and therefore it increases the discovery of new routes and hence used for the telephone network to choose the different path to overcome the traffic occurred in routes.

**3. ANT-HOC NET**

Ant-hoc net is the another swarm based routing algorithm for wireless or ad-hoc network. It is a hybrid algorithm which contain both the proactive and reactive components. For the last 20years ad-hoc networks are the fast developing engineering. The MANET consists of nodes which are dynamically self- organized and network topology are infrastructure less. As they self organize having no centralized behavior, their main aim is to minimize delay, increased throughput, maximize energy efficiency and lifetime of network (G.Dicaro, 9|1998). So it is based on the idea of ant colony optimization as it is adaptive, robust and provide automatic load balancing and self healing.

In MANETS there are no routers, all nodes can act as routers and data packets are sending from source to destination in a multi hop fashion.

Due to the challenging behavior of MANETS like robustness, constantly changing behavior of topology and the restricted flow of control ant based routing algorithm should be used because of the same properties of self-organization, no central or infrastructure less. Ant hoc net routing algorithm is designed for MANETS based on self organizing behavior of ant colonies , the shortest path discovery and ant colony optimization (Gianni Di Caro). The shortest path is detected by using the key element called pheromone trail produced by the ants.

The path having high probability of pheromone is the shortest path covered by ants from their food source to destination. More ants can be attracted towards high intensity of pheromone field which in turn increases the pheromone level and hence this way of sampling the path by ants measure the goodness of each move. This form of indirect communication among the ants with the social environment with all modifications lead to the phase of global co-ordination of agents action is called “stigmergy” (Theraulaz G, 1999). Stigmergic co-ordination obtained the self organized behavior of social insects which is the second key element of ant based routing.

#### *How routing take place in MANETS:*

MANET routing algorithm can be classified as proactive, reactive and hybrid(combination of proactive and reactive) algorithms. In *proactive algorithm* ,where the routing information will updated between all the pair of nodes in a network at all the times. The advantage of proactive algorithm ,it always available whenever needed or without any demand whereas its disadvantage is that it always track all the routing Information all the time which decrease the overhead and is difficult for large networks to update the information of all pair of nodes (example- DSDV Destination sequence distance vector routing & OSLR Optimized link state routing (Manjula poujary, 2011)). In *reactive algorithm* routing information between the nodes are update only on demand or it is maintained for only between those pair of nodes having source and destination of data packets. They set routes on demand or whenever needed for the new communication setup or any communication failure. The advantage of this type of algorithm , it is more efficient and is more scalable. The disadvantage is overhead delay that is when needed not available (example: DSR dynamic source routing, AODV ad hoc on demand distance vector routing algorithm). The *hybrid algorithm* is the combination of proactive and reactive algorithm try to maintain the best of them.

MANETS have some limitation in network like battery power, changing topology and high mobility etc (Neha patil, may 2015). Due to these limitations new class of algorithm can be set up based on swarm intelligence. The ant colony optimization (ACO),at earlier it was used to solve the travelling salesman problem, set partition problem and multi- knapsack problem which was based on behavior of ants for searching the shortest path between source to destination in communication network which was further implemented for the ad-hoc networks. Several algorithms are introduced based on ant colony optimization in ad-hoc network like ARA and PERA. The ARA is proposed by mesut crunes , udo sorges and I.bouazizi (scientists) in 2002 for multipath network. It is purely reactive in nature. Due to limited efficiency of proactive algorithm, reactive algorithm ARA is introduce to reduce the overhead for routing. It consists route discovery, maintenance and failure handling. The AODV can have the same mechanism but

it does not prevent the traffic of nodes as ARA prevents the loops by maintaining the traffic at nodes. If nodes receive the backward node and then deactivate the node so that it cannot be further send in same direction. This loop prevent mechanism is problematic because further backtracking is not resolved but it is based on traffic memorization. In case of link failure pheromone value tends to be 0 and the node will send packet to another link and so on the best link. Enhanced version of ARA protocol routing is based on probabilistic rather than selecting path with maximal pheromone trail. It uses data prioritized queue rather than ordinary data packet handling experiment data shows, the flooding of data packets is a big disadvantage in ad-hoc network rather in case of mobile ad-hoc network , sensor based algorithms , satellites algorithms and so on.

### III. CONCLUSION

In this review paper we explain the swarm based routing algorithms and the protocols used in these algorithms. The Ant-net and ABC are used for wired network whereas Ant-hoc net is used for wireless networks based in ant colony optimization. We give an overview of ARA (ant colony based routing algorithm) is used for the prevention of loops and enhanced version of ARA explain the probabilistic routing. Instead of this the flooding packets can delay the network delivery of packets. So in future we give an overview related to control the flooding of data packets in network and flooding techniques with less overhead delays.

### REFERENCES

- [1] Peter Dempsy and Alfons Schuster, “swarm intelligence for network routing optimization”, Journal of Telecom and Information and Technology,3|2005.
- [2] Eric Bonabeau and Christopher Meyer, “swarm intelligence: a whole new way to think about business”.
- [3] Csete ME, Doyle JC, “reverse engineering of biological complexity”, science 2002; 295(1):1664-1669.
- [4] Milos Stojmenovic, “swarm intelligence for routing in ad-hoc wireless network”.
- [5] Rajeshwar Singh, D. K.Singh and Lalan Kumar, “Swarm intelligence based approach for routing in mobile Ad Hoc networks”.
- [6] G.Dicaro, M.Dorigo, “ant-net: distributed stigmergic control for communication network”, Journal of artificial intelligence research, 9|1998, 317-365.
- [7]. Gianni Di Caro, Frederick Ducatelle and Luca Mario Crambardella, “special issue on self-organization in mobile networking”.
- [8] Theraulaz G, Bonabeau E, “a brief history of stigmergy artificial life, special issue on stigmergy”, 1999;5:97-116.
- [9] Manjula poujary, B.renuka, “ ant colony optimization routing to mobile ad-hoc network in urban environment” , “international journal of computer science & information technology”, vol.2(6),pp.2776-2779,2011.
- [10] Neha patil, S.B takale, “a review of development of routing protocol using ant colony”, International journal of advance research in computer science and software”,vol.5,issue 5,may 2015.

# An Investigation of Distributed Denial of Service Attacks at Application-layer

Parneet Kaur\* and Abhinav Bhandari

\*M.Tech Scholar, Department of Computer Engineering, Punjabi University, Patiala, Punjab, India

[kparneet471@gmail.com](mailto:kparneet471@gmail.com)

Assistant Professor, Department of Computer Engineering, Punjabi University, Patiala, Punjab, India

[bhandarinitj@gmail.com](mailto:bhandarinitj@gmail.com)

**Abstract:** In recent years, the online services and web applications offered by the Internet have become very popular. But Distributed Denial of Service (DDoS) attacks is atrocious threat to the application amenities. DDoS attacks are typically carried out at the network-layer and are easily defeated by the current defence mechanisms. Nowadays, DDoS attacks launched against the application-layer are most devastating in nature and are not detected by the network-layer defence methods. Therefore, detecting DDoS attacks at the application-layer has become a serious problem for the availability of application services. In this paper, a current scenario of the application-layer DDoS (ALDDoS) attacks has been presented along with its defence mechanisms. Defence mechanisms can be effective only if the attack signatures are represented in irregular user behavior.

**Keywords:** *DDoS Attacks, ALDDoS Attacks, Defence Mechanisms*

## I. INTRODUCTION

In the present era, the society has been swayed by information and telecommunication sciences such as Internet and e-commerce [1]. With the increase in the number of Internet users, the amount of information has increased tremendously on the interconnections of the networks. However, the open nature of Internet architecture results in frequent opportunities for attackers to commence Distributed Denial of Service (DDoS) attacks. These assaults are the most sophisticated attacks with botnets as their engines behind them [2] that consign a scads of normal or malformed packets apropos the target to enervate the victim's caches or exploit the protocol bugs (or vulnerabilities) [1]. The main objective of such dreadful attacks is to make the online amenities inaccessible to the permissible traffic that can extend from the petulance to website users, serious financial or business losses that relies on the online availability of services. Moreover, attackers mold or spoof the source IP address (i.e. IP spoofing, subnet spoofing) to dissemble their locale and identity in order to defeat the traceback schemes against DDoS attacks. Countering DDoS attacks is shearing onerous due the presence of vast stratagems and techniques available to assailants [3]. In recent literatures, the researchers have posited many glosses to avert the DDOS attacks from separate OSI layers [4] where the defence schemes have mainly cynosure on the IP and TCP layers in lieu of the high (Seven) layer. Therefore, defence methods design to defend network or transport layer DDoS attacks [5] are not decorous to grasp new sort of attacks at the application-layer [6]. Application-layer DDoS (or ALDDoS) attacks have become the pressing problem for today's Internet because the attacks are legitimate in

packets as well as protocols [7] against web servers and are indistinguishable from the flash events (or crowds). The remainder of this review paper includes the following section: The background and attack strategy of ALDDoS attacks has been presented in section II. Section III outlines the various defence mechanisms against these scandalous attacks. Section IV concludes the paper.

## II. BACKGROUND AND ATTACK STRATEGY

Nowadays, DDoS attacks are moving away from network-layer to application-layer that pretence as flash crowds [8]. It has been mentioned in the security report of Arbor Networks that there is a stabbing upsurge in the occurrence of DDoS attacks against the well liked websites [9]. Recent studies show that there exist more than 60000 bots for instigating these strapping attacks [3]. DDoS attacks are targeted to shatter the network resources (such as domain name servers, routers, or web clusters) or to bring the application server down (i.e by consuming the server resources like CPU or I/O bandwidth) with outward communication pleas, so that it cannot repose to genuine traffic [8] [10]. DDoS attacks are classified into two categories: *intrusion attacks* (to exploit the software openness such as buffer overruns) and *protocol attacks* (that squeeze protocol fickleness to proffer servers unreachable e.g., confiscate DNS entries or revamping routing). With ALDDoS attacks, the intruder attacks the victim server through bots with normal TCP connections [8] where the bots send the packets with a legitimate format which are not easily exposed by current detection mechanisms [11]. These attacks have become more subtle and harder to detect when the attacker uses only a few resources for conducting an attack. In order to hedge the detection of ALDDoS attacks, the attacker tries to be non-invasive as well as protocol-acquiescent and overwhelm the server resources (to increase the application workload) after feigning the identity of legit patrons of the application service [3]. The different classes of ALDDoS attacks are divulged in [10], namely *Session flooding, Request flooding and Asymmetric attacks*. *Session flooding attacks* occur when the session connection request rate increases than the normal request rate from legitimate clients. *Request flooding attacks* happen when the session contains a huge number of requests than it usually handles. *Asymmetric attacks* arise when high-workload requests are sent to different sessions. The detection of such attacks is a new challenge due to indistinguishable nature of attacker requests to the normal requests [8]. It has been revealed in [3] that SYN flooding attacks give a way in generating the ALDDoS attacks. ALDDoS attack such as “Mydoom” has been witnessed in [8]. An ALDDoS attack, namely Index reflection attack has been disclosed in [12] that is verified on many peer-to-peer applications such as Gnutella, BitTorrent, Overnet, FastTrack, ESM and so on. According to [13], 90% of tenable clients’ seizures the 10% of the WebPages of a website, which are termed as “hot WebPages”. Therefore, it behoves serene for the detection methods to recognize the bots as the human user can identify that which WebPages are hot WebPages and which are not whereas the machines (zombies or bots or attack tools) do not.

## III. DEFENCE MECHANISMS

This section introduces the various defence approaches against ALDDoS attacks. Most of the defence approaches against ALDDoS attacks are hinged on packet type and rate perusal, but such schemes become ineffective in case the attackers use a new type of packet that has not been yet defined in the attack signatures [1]. When the detection tools use the traffic logs to expose the attack traffic pattern, it is very time-consuming to look for or scans the huge amount of regular traffic. Kandula et al. [14] and Ahn et al. [15] develop CAPTCHAs based automated detection system using artificial intelligence methods in which humans can recognize the

distorted pictures and fathom the paradoxes to retrieve the services, but the robots (or zombies or bots) cannot. But it is very troublesome for the clients solve the puzzle again and again. Therefore, CAPTCHA tests are not so well founded [11]. In [16], the authors propose an Adaptive Content Distribution Networks (CDN) approach to distinguish the flash events from DDoS traffic based on cluster overlapping i.e. requests from the flash crowd are less scattered as compared to DDoS events. Moreover, flash events can be distanced on the basis of client distribution, per-client request rates, and requested file types. In [17], the authors broach that ALDDoS attacks with damper erratic request attacks can be defended by the statistical methods. In [18], the authors develop a matrix based on source IPs (either spoofed or non-spoofed) to validate the legitimate clients. The detection system generates rank of the matrix and uses XOR and AND operations based on cluster overlapping in order to distinguish the flash events from DDoS flows. Ranjan et al. [3] introduce an obverse method based on statistical measures to identify the features of HTTP bouts and rate limiters are employed to filter the venomous traffic. Xie et al. [6] propose a detection strategy based on record vogue in which extended hidden semi-Markov model dissects the user web browsing behavior and detects ALDDoS attacks. The sequence of User's HTTP requests accustoms as an important measure to check the user's normality. After analyzing the system logs of different websites, [19] use a human behavior modelling concept to distinguish the DDoS bots from human users. Yu et al. [20] propose Trust Management Helmet (TMH) method that assigns a trustee license to the legitimate clients on which detection is made. Wang et al [13] discusses relative entropy and clustering based detection scheme that defines the click ratio of web objects and distils the click ratio traits where the extracted features are used to compute the relative entropy of new instances and to identify the suspicious sessions. In [21], the authors present a detection stratagem based on information theoretic measures that use the span of the package distribution etiquette of the different sceptical flows to distinguish the satirizing flooding attacks from plausible accessing. Beitollahi et al. [11] present a novel approach ConnectionScore against ALDDoS attacks on the basis of statistical analysis of regular traffic where the connections get a score based on their behaviour. The connections having low scores are treated as anomalous and the bottleneck resources are taken from such connections. This technique can identify the legitimate connections with high probability and can effectively handle both common and meek attacks. Devi et al. [10] propose entropy based detection scheme that surveils the user's browsing behavior (e.g. HTTP request rate, page deeming measure and solicited series of objects and their order) from the system log in the past and reckons the licit request rate (entropy) and the user's trust level by looking up their history. Then the computed entropy of arriving requisitions juxtaposes with the venial rate or threshold level. If the deviations exceed the threshold the sessions are treated as vindictive. Moreover, rate-limiter is employed to filter the session (based on user's trust score) and a scheduler is used to organize the requests (based on system workload). Durcekova et al. [22] presents the different ALDDoS attack detection methods. The authors have introduced two detection architectures to monitor web traffic and discover dynamic changes in regular burst traffic. Yu et al. [2] propose a flow correlation coefficient based detection approach that is used as a metric to find the resemblance among chary flows. The authors have mentioned that DDoS attacks flows have high similarity than the normal flash crowds. Zargar et al. [23] has discussed various defence mechanisms against DDoS attacks that can be deployed in any network (source-end, victim-end and core-end networks) in order to detect, prevent or response the DDoS attacks. In [24], the authors propose a detection mechanism based on data mining procedures in order to reveal the abnormal behavior of the web pages (i.e. both in static and dynamic web domains). Xu et al. [25] present a random walk graph model to expose the

asymmetric DDoS attacks. The graph has been constructed on the basis of sequences of page requests from the legitimate clients and predicts the subsequent page request sequence. If the predicted behavior is not similar to the observed one earlier, then it is declared as freakish. Zhou et al. [9] develop a detection scheme that consists of a Real-time Frequency Vector (RFV) and an entropy calculation module to differentiate the ALDDoS attacks from flash crowds.

#### IV. CONCLUSION

ALDDoS attacks are being the fate of a hot issue around the world that degrade or disrupt the application services and mimics or occur in the flash crowd events of popular websites which are difficult to be addressed. Because of the seriousness and ruinous of these attacks, researchers have found many solutions for defending ALDDoS attacks, but none of the methods give best results in terms of high detection accuracy and low false alarm ratio. Therefore, the emerging ALDDoS attacks are still an open problem.

#### REFERENCES

- [1] T. H. Kim, D. S. Kim, S. M. Lee and J. S. Park, "Detecting DDoS attacks using dispersible traffic matrix and weighted moving average," in *Springer*, 2009.
- [2] S. Yu, W. Zhou, W. Jia, S. Guo, Y. Xiang and F. Tang, "Discriminating DDoS attacks from flash crowds using flow correlation coefficient," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 6, pp. 1073-1080, 2012.
- [3] S. Ranjan, R. Swaminathan, M. Uysal, A. Nucci and E. Knightly, "DDoS-shield: DDoS-resilient scheduling to counter application-layer attacks," *IEEE/ACM Transactions on Networking (TON)*, vol. 17, no. 1, pp. 26-39, 2009.
- [4] S. B. Ankali and D. Ashoka, "Detection architecture of application-layer DDoS attack for internet," *International Journal of Advanced Networking and Applications*, vol. 3, no. 1, p. 984, 2011.
- [5] C. Ye and K. Zheng, "Detection of application-layer distributed denial of service," in *IEEE*, 2011.
- [6] Y. Xie and S.-Z. Yu, "Monitoring the application-layer DDoS attacks for popular websites," *Networking, IEEE/ACM Transactions on*, vol. 17, no. 1, pp. 15-25, 2009.
- [7] J. Yu, Z. Li, H. Chen and X. Chen, "A detection and offense mechanism to defend against application-layer DDoS attacks," in *IEEE*, 2007.
- [8] Y. Xie and S.-Z. Yu, "A novel model for detecting application-layer DDoS attacks," in *IEEE*, 2006.
- [9] W. Zhou, W. Jia, S. Wen, Y. Xiang and W. Zhou, "Detection and defense of application-layer DDoS attacks in backbone web traffic," *Future Generation Computer Systems*, vol. 38, pp. 36-46, 2014.
- [10] S. R. Devi and P. Yogesh, "Detection of application-layer DDoS attacks using information theory based metrics," *CS & IT-CSCP*, vol. 10, 2012.
- [11] H. Beitollahi and G. Deconinck, "Tackling application-layer DDoS attacks," *Procedia Computer Science*, vol. 10, pp. 432-441, 2012.
- [12] J. Yu, C. Fang, L. Lu and Z. Li, "Mitigating application-layer distributed denial of service attacks via effective trust management," *IET communications*, vol. 4, no. 16, pp. 1952-1962, 2010.

- [13] J. Wang, X. Yang and K. Long, "A new relative entropy based app-DDoS detection method," in *IEEE*, 2010.
- [14] S. Kandula, D. Katabi, M. Jacob and A. Berger, "Botz-4-sale: Surviving organized DDoS attacks that mimic flash crowds," in *USENIX Association*, 2005.
- [15] L. V. Ahn, M. Blum, N. J. Hopper and J. Langford, "CAPTCHA: Using hard AI problems for security," in *Springer*, 2003.
- [16] J. Jung, B. Krishnamurthy and M. Rabinovich, "Flash crowds and denial of service attacks: Characterization and implications for CDNs and web sites," in *ACM*, 2002.
- [17] W. Yen and M.-F. Lee, "Defending application DDoS with constraint random request attacks," in *IEEE*, 2005.
- [18] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer networks*, vol. 51, no. 12, pp. 3448-3470, 2007.
- [19] G. Oikonomou and J. Mirkovic, "Modeling human behavior for defense against flash-crowd attacks," in *IEEE*, 2009.
- [20] J. Yu, C. Fang, L. Lu and Z. Li, "A lightweight mechanism to mitigate application-layer DDoS attacks," in *Springer*, 2009.
- [21] S. Yu, W. Zhou, R. Doss and W. Jia, "Traceback of DDoS attacks using entropy variations," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 3, pp. 412-425, 2011.
- [22] V. Durcekova, L. Schwartz and N. Shahmehri, "Sophisticated denial of service attacks aimed at application-layer," in *IEEE*, 2012.
- [23] J. J. Saman and D. Tipper, "A Survey of Defense Mechanisms Against Distributed Denial of Service (DDoS) Flooding Attacks," *IEEE Communications Surveys and Tutorials, Accepted for Publications*, 2013.
- [24] M. Stevanovic and J. M. Pedersen, "An efficient flow-based botnet detection using supervised machine learning," in *IEEE*, 2014.
- [25] C. Xu, G. Zhao, G. Xie and S. Yu, "Detection on application-layer DDoS using random walk model," in *IEEE*, 2014.

# SURVEY OF DIFFERENT KEY FRAME EXTRACTION TECHNIQUES

Riya<sup>#1</sup>, Sumandeep kaur<sup>#2</sup>

<sup>#</sup> Department of Computer Engineering, Punjabi University  
Patiala, India

<sup>1</sup>Khuranariya43@gmail.com

<sup>2</sup>sumandhanjal@gmail.com

## Abstract

Now a day's social networking become a trend and through social networking we can share videos, audios, images and many more. But sometimes there is huge amount of unnecessary or redundant data is present in our videos so to reduce these types of data we use a method called key frame extraction method and to summarise that extracted data is done by a method known as video summarization .In last few years lot of work has been done on this area many new techniques and different algorithms are used. This paper does comparison of different key frame extraction techniques and also mentions their advantages and disadvantages and their limits.

**Keywords:** key-frame, video summarization histogram, epitome, sparse, on the fly extraction, shot detection.

## 1. INTRODUCTION

Video is basically a recording of moving data or images. We analyse different objects moving and we record them and that footage is called video. But in actual that recorded data is not that simple it contains scenes, shots and frames. When we do recording there is huge amount of data recorded which is not superfluous it only amplify the volume of videos that type of data is called redundant or surplus data and to remove that surplus data we use the method called key frame extraction. This method is an active method which is used to remove the repetitive or disused information in video retrieval application (Azra Nasreen).Shapes, edges, optical flow, colour histograms, motion descriptors are the features which we use for key frame extraction. This method is also obliging in reducing memory space and intricacy of the videos. These key frames are extracted by two ways -Locally and globally by using various features such as audio features or video features (Miss A.V.Kumthekar, 2013) (Sachan Priyamvada Rajendra, 2014).Key frame extraction method is further classified as sequential based and cluster based .We can use visual features and temporal information to determine key frames in sequential based approach and basic ideas of production of key frames are used in cluster based approach (Sheena, 2015).

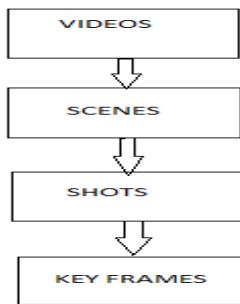


Fig1.1. Different components of Video (Miss A.V.Kumthekar, 2013)

**A. Video:** Recording of reposition or shifting of data or object.

**B. Scene:** A scene is an amalgamation (combination) of shots .Video is first divided into scenes and then that scenes are further divided into shots (Zhao, 2009).

**C. Shot:** A shot is an assembly (collection) of frames .The difference between these frames are used to define the shots (Zhao, 2009).

**D. Frame:** A frame is a grouping of pixels. Pixel is defined as the tiny constituent of an image (Zhao, 2009).

**E. Precision and Recall:**"Precision" defines how much our algorithm is consistent or reliable while "Recall" Defines the carrying out (performance) of the algorithm from the starting to the end (Zhao, 2009).

#### **F. Video Detection process**

**1. Color Histograms.** In color Histograms there are different frames are collected on the basis of similarity of their colours. A Threshold value is used to indicate the shot boundaries after detection of inter frame similarities in the frames (Chandra Sehkar Muthlesh, March ,2016).

**2. Edge Detection:** An object and its background are separated by a single line or boundary and that boundary is called edge. An edge also separates two overlapping objects. Edge Detection is the step where two images are detected to find out the edges . (Chandra Sehkar Muthlesh, March ,2016).

**3. Macro blocks:** The frames are partitioned in blocks of pixels (e.g. macro blocks of  $16 \times 16$  pixels in MPEG) (Sachan Priyamvada Rajendra, 2014). And these macro blocks further divided into three types forward prediction, backward prediction or no prediction at all (Chandra Sehkar Muthlesh, March ,2016). We can classify the different blocks by encoding the data based on the efficiency and motion estimation. (Chandra Sehkar Muthlesh, March ,2016).

## **2. SHOT DETECTION**

The foremost step of video extraction method is shot detection method. In this procedure we expose the transition between different shots each shot consist of a number of frames and can be represented by one or more frames based on their different temporal differences. The shot detection method is further categorised into two categories: uncompressed and compressed .The detection method generally categorise into abrupt transition detection and gradual transition detection (Azra Nasreen).The commonly used methods are: a)To detect motion based objects we use Pixel-based comparison. Template matching which is used for error detection .c).Histogram matching which is used for the location information of pixels.

## **3. KEY FRAME EXTRACTION**

Number of frames are combined together to make a video. These frames amuse large cosmos in memory. Frame rate is about 20 to 30 frames per second. To start the extraction process, the first frame is declared as a key frame, and then the variation between various frames are computed between the current frame and the last extracted key frame (Sachan Priyamvada Rajendra, 2014). If the variation between the frames complies with a certain threshold condition, then the current frame is selected as key frame. There are many algorithms used for key frame extraction .There are numerous techniques used for key frame extraction certain of them are:

- 1).Key frame extraction pedestal on shot activity. Gresle and Haung estimate an activity indicator by performing different operations on the intra and reference histograms and depend on that activity curve, the local minima is pick as the Key frame (P.Gresele, 1977).
- 2).Key frame extraction method build on macro-block statistical characteristics of MPEG video stream. Janko and Ebroul spawn the analysis statistics by finding out the difference between different frames of macro blocks which are extracted from the MPEG compacted stream and the key frame extraction method is implemented using difference metrics curve simplification by discrete contour evolution algorithm (J.Calic, 2002).

3).Key frame extraction build on motion analysis. Wolf computed the optical flow by using a simple motion metric by evaluating the alterations in optical flow along the series (W.Wolf, 1996).

4).Key frame extraction from consumer videos using epitome.N.jojic and Frey use image Epitome as a feature vector and to measure the dissimilarity between the frames of the given video we apply an information divergence which is completely based on the feature vector which is used as a distance measure. (C.T .dang).

#### **4. SURVEY OF DIFFERENT RESEARCH**

**Sheena CV, N.K Narayan** purposed an automated extraction of Key frames method for video summarization. To lead the removal of feature vector for classification this method is used because this method is useful for removing the non informative and unuseful frames..In this method object recognition and classification methods are used .These methods are useful for extracting the visual features of the videos and they compute key frames by using simple calculations of histograms of absolute difference or threshold value of consecutive frames (recognition, Dec2013). This method is highly efficient and feasible. The worth of compression ratio and fidelity criteria shows that the results obtained are reasonably accurate (Sheena, 2015).

**Miss.A.V.Kumthekar, Prof.Mrs.J.K.Patil** purposed a method for extraction of frames from the video summarization by using histogram difference. This method having these features it is useful in removing repeated frames and reduces the computational complexity of the key frames and helps to improve recognition. The compression ratio of this purposed method is 98% which is very high and the error rate is very low as compare to other algorithms (Miss A.V.Kumthekar, 2013).

**C.T Dang, M.kumar, H.Radha** purposed a method for key frame extraction of consumer videos by using Epitome which is small than the size of the image. This method works on two applications video summarization and video Processing. This approach is completely depend on the size of the epitome and the image epitome is used as the feature vector A minima and maxima method is used to find out the dissimilarity scores of the given video. It does not require shots detection, or semantic understanding. The only problem of this method is that sometimes it quantitatively misses the frames. (C.T .dang).

**Mrs.Poonam S. Jadhava, Prof. Dipti S. Jadhav** This approach is worn for video summarization using image block mean, standard deviation, skewness and kurtosis histogram has been planned. The projected technique best confine the shot boundary detection and key frames using the higher order color moments of the frames as well as removes redundancy. (Mrs.Poonam S. Jadhav, 2015).

**Guozhu Liu et & al.** proposed a method in order to origin correct information from video, is to process video data in efficient manner and reduce the transfer stress of network, more alertness is given to the video processing technology. Key frame extraction and video segmentation methods are used to minimize the data of the video. So, these are the only two technologies which becomes the concentration of the research. By using the features of MPEG compressed video stream, a new method is presented for extracting key frames. We use an improved histogram matching method which is used for video segmentation .For each sub-lens Key frames are extracted by using different frames like I-frame-frame and B-frame. Fidelity and compression ratio are used to measure the validity of the method. Experimental results show that the extracted key frames can ssum up the salient data of the video and the used method is highly feasible and having high efficiency, and high robustness. It is not only highly feasible but also have high efficiency, with low error and high robustness (Chandra Sehkar Muthlesh, March ,2016)

**Walid Barhoumi and ezzeddine Zagrouba** purposed a method for on the fly extraction of key frames for efficient video summarization. In this an efficient object based method for key frame extraction is presented by maintaining convenient memory requirements, even under loop-closure situation. This method is mainly based on detection of significant events while analysing the spatial-temporal behaviour and visual

appearance of salient objects. The detection on the fly of key frames allows integrating implicitly the temporal content within the input shot without having to process the whole slot (Walid Barroom, 2013).

**Magda B.Fayka, HebaA.ElNemrb, and MonaM.Moussab** In this paper, an algorithm for key frame selection is presented. This technique is based on dividing the video into equal segments and then opt the key frames for each segment using PSO. A post-processing stage compensates for the uncompromising initial segmentation into equal segments by performing inter- and intra-merging operations. A widespread optimum segmentation size has been agreed that can be used to give acceptable results for most video types. This general segmentation size can be used as an initial value that can be further tuned in a learning stage applied on video samples. The Segmenting video momentary results in sinking the processing time, while the presented post-processing task enhances the results by decreasing the false rate. Dividing the video into equal segments has trim down whole processing time by almost 70% in spite of the over-head needed for the post-processing task that compensates for this simple segmentation approach (Magda B.Fayka, 2010).

REVIEW OF DIFFERENT TECHNIQUES, THEIR ADVANTAGES AND LIMITATIONS.						
S.N	TITLE	AUTHOR	AREA/APPLICATION	METHODOLOGY USED	DESCRIPTION	
					ADVANTAGE	LIMITATION
1.	Key-frame Extraction from MPEG video stream.	Guozhu Liu, and Junming Zhao[5].	Video Processing Technology.	1. Histogram Matching method.  2. Features of I-frame p-frame and b-frame are extracted for each Sub lens.	1. Good Feasibility.  2. High efficiency  3. Low error.  4. High Robustness.	
2.	Key frame extraction from consumer videos using Epitome.	C.T.Dang,M.k umar, H.Radha [1].	Video Processing and Video Summarization	1. Exploit image epitome to measure dissimilarity between frames.  2. Use min-max approach to extract the desired no of key frames.	1. Does not require Shot detection.  2. Does not require semantic understanding.  3. Good feasibility.	1. Quantatively misses frame sometimes
3.	Key frame extraction and shot boundary detection using Eigen values.	Vijeekumar Benni, R.Dinesh, Punitha P, and Vasudeva Rao[12].	Content-based Video retrieval.	1. Covariance Matrix is used.  2. Eigen values are used for measuring the dissimilarity between the consecutive	1. Highly Efficient.  2. Accurate in detecting the abrupt cuts.	1. Fails to detect the Fades.  2. Threshold is set Empirically.

				frames.		
4	Key frame extraction by analysing histograms of videos frames using statistical methods.	Sheena V, N.K Narayan[3]	Object recognition and classification	1. Extract visual features.  2. Compute key frames using simple calculations of histograms of absolute difference or threshold value of consecutive frames.	1. Highly efficient.  2. Highly feasible.	
5.	On-the-fly extraction of key frames for efficient video summarization.	Walid Barhoumi and Ezzeddine Zagrouba[11]	Video Summarization and object based.	1. This method allows a properly capturing of the underlying dynamics of the input frames.  2. It allows to integrate implicitly the temporal content within the input shot without having to process the whole shot.	1. Avoid Complexity and highly efficient.	
6.	A Novel Approach Towards Key frame Selection for Video Summarization.	Chitra A. Dhawale and Sanjeev Jain[19]	The method has, been tested on various video sequences like news programs,sports,academic etc.	The improved algorithm for Histogram based approach.	This method gives better results in much less memory and time.	
7.	Videos key frame extraction from Recognising hand drawn human face.	T. Judes Divya and A. Rama[17].	Face Recognition	Principle Component Analysis algorithm.	The forensic department and criminal investigations.	
8.	Key Frame extraction from consumer Videos using sparse representation .	Mrityunjay Kumar, Alexander C.loui[20].	Video processing and video Summarization	Video frames are projected to a low dimension feature space using a random projection matrix and sparse representation.	1. Does not requires shot detection, segmentation or semantic understanding.2.ComputationallyEfficient.	
9.	Design of video summarization in the wavelet domain using statistical feature	J. Kavitha, Dr .P. Arockia Jansi Rani.J[16].	Marine Research	Wavelet Domain Using Statistical Feature Extraction	1. Effectively detect even the movements.  2. Detect fast moving animals	

	detection.				
--	------------	--	--	--	--

## 5.CONCLUSION

In this paper, assessment of various techniques has been done. The above evaluation shows that the technique Key frame extraction from consumer videos using Epitome and using sparse representation does not require shot detection, segmentation and semantic understanding and are computationally efficient. The technique key frame selection for videos summarisation is a histogram based approach and works in less memory and time. Features for good quality of video summarization algorithm are threshold free; application range is wide, low error, high robustness, and good feasibility, less wipe effect. The technique Wavelet Domain Using Statistical Feature Extraction detects the movement of fast moving animals which is not only specific for animals even detect movements of anything. There are many other techniques which are still under working this paper is just an outline of those techniques which is recently used.

## References

- [1].C.T.Dang, M.kumar and H.Radha "Key frame extraction of consumer videos by using Epitome"IEEE, 2012.
- [2]. Azra Nasreen, Dr ShobhaG "Key frame extraction from videos-A survey" International journal of Computer Science &Communication Networks, Vol3 (3), 194-198.
- [3].Sheena, N.K.Narayanan "Key -frame extraction by analysis of histograms of video frames using statistical methods" 4<sup>th</sup> International Conference on Eco-Friendly computing and Communication system, Procedia Computer Science 70(2015) 36-40.
- [4].Miss.A. V.Kumthekar, Prof.Mrs.J.K.Patil, "Key frame extraction using color histogram method" International Journal of Science Research Engineering& Technology (IJSRET), Volume2 Issue4 pp207-2014, july2013.
- [5].Guozhu Liu, Jumming Zhao, "Key Frame Extraction From MPEG Video Stream" Proceedings Of The Second Symposium International Computer Science And Computational Technology(ISCSCT)Huangshan,P.R.Chih,PP.007-011,26-28,Dec2009.
- [6]. Chandra Shekhar Mithlesh, Dolly Shukla "A Case Study of Key Frame Extraction Techniques" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol.5, Issue3, March2016.
- [7]. P.Gresle, T.S.Huang "Gisting of Video Documents: A Key Frame Selection Algorithm Using Relative Activity Measure" The 2<sup>nd</sup> International Conference in Visual Information System, 1977.
- [8]. J.Calic, E.Izquierdo "Efficient Key Frame Extraction and Video Analysis Information Technology: Coding and Computation" International Symposium on Information Technology,pp28-32,2002.
- [9].W.Wolf "Key Frame Selection By Motion Analysis" Proc IEEE International Conference Acoust Speech Signal Proc, Vol2, pp1228-1231, 1996.
- [10].N.Jojic, B.Frey, A.Kannan "Epitomic Analysis Of appearance and Shape" International Conference On Computer Vision (ICCV), 2003.
- [11]. Walid Barroom, Ezzeddine Zagrouba "On The fly key Frames For efficient Video Summarization" AASRI Conference on Intelligent Systems and Control, 2013.
- [12]. Vijeetkumar Benni, R.Dinesh, Punitha, Vasudeva Rao "Key Frame Extraction and Shot Boundary Detection Using Eigen Values" International journal of Information and electronics engineering, Vol.5.No.1, January 2015.
- [13].Li.Liu. "Learning Discriminative Key posses For Action "IEEE Transaction On Cybernetics,Vol43,No6,Dec2013.
- [14]. Mrs Poonam S.Jadhav, Prof Dipti S.Jadha, "Video Summarization Using Higher Order Color Moments "International Conference On Advanced Computing Technologies and Applications(ICACTA-2015)Published by Elsevier B.V,Procedia Computer science,Vol.45,pp275-281,2015.
- [15].Lei Pan, Xin Shu and Ming Zhang "A Key Frame Extraction Algorithm Based on Clustering and Compressive Sensing" International Journal Of Multimedia and Ubiquitous Engineering, Vol.10, no.11.
- [16]. J.kavitha,Dr.P.Arockia Jansi Rani "Design Of A Video Summarization Scheme in the Wavelet Domain Using Statistical Feature Extraction" I.J,Image graphics And Signal Processing,MPECS,Vol.4,pp60-67,2015.
- [17]. T.Judes Divya, A.Rama "Video Key frame Extraction for Recognising Hand Drawn Human Face" Middle-east journal of Scientific Research, Vol.20, issue4, pp537-541, 2014.
- [18]. Sachan Priyamvada Rajendra, Dr.Keshaveni N "A Survey of Automatic Video Summarization Techniques" IJEECS, Vol.3, 2014.

[19].Chitra A.Dhawale and Sanjeev Jain “A Novel Approach Towards Key Frame Selection for Video Summarization” Asian Journal of Information Technology, 2008.

[20]. Mrityunjay Kumar, Alexander C.loui “ Design of video Summarization in the wavelet domain using statistical feature detection”.

# Protege: Review of a powerful Ontology tool

Nancy Bhalla  
(Research Scholar)  
(Department of Computer Engineering,  
Punjabi University)  
nancybhalla00.nb@gmail.com

Navjot Kaur  
(Assistant Professor)  
(Department of Computer Engineering,  
Punjabi University)  
navjot\_anttal@yahoo.co.in

**Abstract:** *Ontology is the new trending topic in the present era of technology. This concept originated in the early 90's. The concept of ontology represents the domain concepts in the form of classes and object and data properties. The data contained in the paper represent the overview of ontology domain, its architecture and the brief introduction about the languages used to create and query the ontology. The next section analysed the different ontology visualization and editing tools. It concluded that the protégé is the most advanced and trending visualization tool used by present day researchers. Its architecture and components are also discussed in the paper below. In the final phase, the ontology created in the visualization tool has been briefly explained along with its constructed object properties.*

**Keywords:** *Ontology, Knowledge base, Reasoner, query, retrieval, semantics, domain.*

## 1. INTRODUCTION:

In 1990's, the research community introduced the new term "Ontology" in the field of Information Science. It aimed at the representation of world in a Program Code. It has been implemented recently in numerous fields of Information Technology such as intelligent information integration, Knowledge Management and Information retrieval on Internet. An ontology is used to define the terms and concepts that are used to represent the knowledge area. In previous years,[6] lots of development has been undertaken in the establishment of ontology tools and ideas but the still the area needs more development for future work. From the viewpoint of many researchers, ontology is considered as a backbone to support various types of information management including information retrieval, storage and sharing on web. Semantic web is based on Ontologies. An ancient Greek philosopher Aristoteles (384-322 B.C.) was the very first person to define this term Ontology. He described the ontology as "the metaphysical study of the nature of being and existence". The definition of ontology for Computer science field was firstly introduced by Willen.N.Borst:-"Ontology is the formal specification of shared conceptualization".

The word conceptualization refers to the abstract model of ongoing process in the world that identifies the phenomenon's relevant concepts. The term formal means that ontology is presented in well defined semantics. The term shared shows that an ontology captures consensual knowledge so that it is not constrained to an individual but accepted by a group of people. As mentioned above, ontology contains classes, also known as concepts. Concepts/classes describe the common properties of a collection of individuals. These classes are arranged in hierarchy using the is\_a relationship. The relationship between two objects can also be shown in the form of roles (Binary relationships). The glossology of the term concept and classes is almost same but the only difference is that the classes are mainly used in ontologies whereas concepts are used in Description Logic.

Most of the ontologies are developed for a particular domain to facilitate a controlled vocabulary of terms with an explicitly defined and machine processable semantics.

## 2. LITERATURE REVIEW

**Thaket Slimani**[1] attempted to review and compare some ontological development tools with the special attention to interoperate between them. The basic criterion followed for the comparison is the user interest and its application in different types of real world problems.

If ontology is build from scratch, the following conditions must be taken care of:

- ✓ Selection of a particular tool that is capable of a specified ontology development in an efficient manner
- ✓ Choosing a particular ontology storage type and cross checking the type of inference engine being used by that tool.
- ✓ Priorly indicating the language to be used for the ontology construction.
- ✓ To analyse all the languages that are supported by that specific tool and conformed if the particular language chosen is appropriate for information exchange or not.
- ✓ Evaluation of compatibility and integrity of that particular tool with other languages and tools.

In the concerned paper **Vishal Jain** and **Dr. Mayank Singh**[2] the explicit concept of ontology and its corresponding methodology. Semantic web is the result of ontology development world and also extension to the present internet. In semantic web, information is shown in a well defined meaning that converts unstructured data into computer understandable format. Ontology helps not only in information visualization but also in information gathering and reusing the same information for number of applications. The difference between RDBMS and knowledgebase is shown in table 1below:

TABLE 1  
Comparison of RDBMS and knowledge base [3]

Feature	RDBMS	Knowledge Base
Structure	Schema	Ontology statements
Data	Rows	Instance statements
Administration Languages	DDL	Ontology statements
Query Language	SQL	SPARQL
Relationships	Foreign keys	Multidimensional
Logic	External of Databases/Triggers	Formal Logic Statements
Uniqueness	Keys of table	URI

**Dr. Sunitha Abburu, G.Suresh Babu**[3] tried to sum up their research as ontologies can be built from scratch using ontology development tools or reusing the previously developed ontology. Different types of ontology tools are

- ✓ Ontology development [protégé, Neon, Swoop etc]
- ✓ Ontology merge and alignment tools [Protege with PROMPT, Chimaera etc]
- ✓ Ontology evaluation tools [OntoAnalyser, Ontoclean, RADON]
- ✓ Ontology based annotation tools [Neon with Cicero and OWLDoc]
- ✓ Ontology querying tools and inference engine [Pellet, Racer etc]
- ✓ Ontology learning tools [Neon with ODEMapster, Protege with OntoLT]

### 3. ONTOLOGY TYPES AND LANGUAGES USED TO CREATE AND QUERY .

#### 3.1 Types of Ontologies:

Ontologies has been categorised in number of different types. Some of its types are mentioned in the table 2 below.

TABLE 2  
Different types of ontologies

S.no.	Type	Description
1.	Domain ontology	Designed to present knowledge relevant to the domain.
2	Generic ontology	Can be applied to a variety of domain types.
3	Representational ontology	Responsible for defining general representational entities without defining what should be presented.
4	Task ontology	They provide specific terms for particular task
5	Method ontology	They provide particular problem solving methods

#### 3.2 Languages used to create and Query ontology:

- ✓ **OWL:** Owl stands for web ontology language and was recommended by World Wide Web consortium (W3C) in the month of February 2004[4]It shows compatibility with Extensible Mark up Language (XML) and various W3C standards.
- ✓ **F-Logic:** F-Logic stands for frame logic. [6]It is a knowledge representation ontology language that combines the leverages of conceptual modelling with object oriented frame supporting languages. It delivers declarative, compact and simplest syntax for logic based languages. F-Logic is supported mainly by all Protege versions prior to 4.x. It was originally designed for deductive databases but people started using it frequently for semantic web.
- ✓ **DAML+OIL:** DAML is expanded as DARPA Agent Mark-up Language. DARPA can also be spelled as Defence Advance Research Project Agency. OIL represents ontology interchange language. Description Logic is used by this language to express its original use and means.
- ✓ **SWRL:** It is abbreviated for semantic web rule language. [4]It usually tots up rules to the OIL+DL.
- ✓ **RDF:** It stands for resource description framework. The data model of DF is much similar to conceptual models like entity-relationship and class diagrams, on the point of making statements about web resources following the form subject-predicate –object as expression. This expression is referred as triplet in RDF vocabulary. The resource is denoted by object. Traits are defined by the predicate. Predicate is also responsible for defining the relationship between subject and object.
- ✓ **RDF Schema:** Formally it is set of classes with concrete properties using RDF extensible knowledge representation model. It provides elementary features for the presentment of ontologies.W3C released its first version of RDFS in February 2004.
- ✓ **SHOE:** It stands for simple HTML ontology extensions. These are trivial sets of HTML extensions that give meaning to web pages through information like class, property, subclass, and relationships. SHOE was first released in 1996.
- ✓ **CLIPS:** It is an acronym for ‘C language integrated production system’. [5] Its first version was developed in 1985 at NASA for developing expert system technology. It was the most widely used expert system tool. It is a complete object oriented language (written in C, extended with C and can be

called from C). CLIPS mainly contain rules and facts. Jess and Fuzzy-CLIPS are the descendants of CLIPS language.

- ✓ **HTML:** It stands for Hypertext Mark up Language. [4] It is a set of Mark up tags and HTML tags. Each tag depicts a different document content. HTML tags are the keywords surrounded by angle brackets. For example <p> and </p> denotes the start and end of a paragraph in a html document. Html language is used by web browsers to display and read HTML document.
- ✓ **SPARQL:** It stands for ‘Simple Protocol and RDF Query language’. It is similar to SQL language generally applied for querying RDF data. SPARQL helps us in retrieving triples from triple store (RDF database). It does not make use of foreign and primary keys either. SPARQL uses URI as standard reference format for www. [3]As triple store has enormous size and contains bulk of triples, therefore SPARQL uses graph pattern to query these triples. For example if anyone asked question that they need the list of all subjects that ‘plays guitar’. Then the query will be of the format:

```
PREFIX: <http://aabs.purl.org/music#>  
SELECT? instrument  
WHERE { : Andrew: plays instrument ? instrument}
```

#### **4. ONTOLOGY BUILDING TOOLS**

An extensive outline of different ontology building editors and environments are illustrated below

**4.1 PROTEGE:** It is an open source, java based platform. [6] It footholds creation, manipulation as well as visualization of ontologies in number of different representation formats. It can be personalized to provide domain friendly medium for creation of knowledge models and adding data to it. Protege can be outspreaded via Plug-in Architecture. Protege is mainly used for editing. The earlier versions of protégé used to support only Frame based ontologies. However the recent versions like protégé 3and 4 are much useful for creating OWL and OWL2 ontologies respectively.

OWL consists of object properties, data properties that connects individuals to basic data and supports the functions like reasoning, annotation. Protege is user friendly open source software with number of familiar tabs named active technology, entities, classes, object properties and data properties.

In Active ontology tab, one can get information about the important metrics of current ontology like number of classes, object properties etc. and the respective description logic expressions that the current ontology is uses. Next, the entities tab assists you in exploring all the classes, properties and individuals in an ontology. Each tab is capable of showing number of views that can be resized, removed, splitted, and layered. Protege allows the classification of ontology with the built in reasoner called Hermit that will automatically classify your ontology. After reasoner is done with its work, the additional sub tabs appear to show the inferred class hierarchy. Only unsatisfiable classes are shown in red colour under nothing and everything appears as it is in the hierarchy under their inferred super classes. There are numbers of reasoners available for protégé.

Owl-Viz: Installation of Graph-viz is mandatory for its use. It shows the graphical illusion of class sub assumption hierarchy.

DL-Query: It requires the pre classification of ontology before getting started with queries. With the help of this tab, the reasoner is queried for the sub/super class, inferred members, instances or depending upon what is selected.

**4.2 APOLLO:** Apollo is an easy to learn knowledge application that permits user to create ontologies with basic structures such as instances, classes, functions and many more.[2] It can inherit other ontologies as well. The class system of Apollo toolkit is modelled according to the OKBC. In this, one ontology can inherit the properties and classes of other ontology and then use as its own. Every ontology inherits atleast default ontology that contains all the primitive classes like integer, float, string list etc. Apollo has two types of class slots. These are named as non template and template class slots. Currently Apollo does not support non template class slots. Each class is capable of creating a number of instances. All the slots of one class are inherited by an instance and each slot contains a set of facets.

**4.3 SWOOP:** It is an ontology editor released in 2004. It is also an open source, web based editor for ontology and editor. It also facilitates the comparison of relationship and entities across different ontologies. Swoop also allows the editing and merging of ontologies. Swoop supports various OWL presentation syntax views. [5] It provides reasoning support. A variety of domains can be compared on the basis of their associated properties and instances. Swoop exhibits hyperlink capability to ease the navigation and understanding. Using swoop, users can easily reuse external ontology data. We can either purely link to the external entity or should entirely import the external ontology but the partial import of OWL is not possible.

Swoop allows the use of ontology search algorithm that associates the keywords with description logic with so as to find the related concepts. The search is carried among all the ontologies residing in the swoop knowledge base.

**4.4 TOP BRAID COMPOSER:** This tool is used for modelling and creating the ontologies. It comes in three different editions

- a) Free edition
- b) Standard edition
- c) Maestro edition

It is useful for creating RDF and OWL ontologies.

##### **5. PROTEGE (*an efficient ontology tool*)**

Protege is mainly an ontology visualization tool that is most popular these days for ontology creation and visualization. [6] It facilitates the better understanding and visualization of knowledge. Now day's researchers are focussing on development of visualization tools to graphically visualize the built ontologies and analysing its structure. Protege is one of such a useful tool that assist in intuitive editing of ontologies and also helps in browsing the content of ontology for an easy modelling procedure. In the core, the ontology is the specification of concepts in the domain and describes the relationship between them for a better understanding. Protege is capable of editing the knowledge bases and ontologies in an interactive manner that can be concurrently accessed using GUI and Java API's.

The functionality of protégé can be extended with the help of additional plug-ins offering a variety of surplus tasks such as multimedia support, querying and reasoning engines, ontology management etc. Frame based ontologies are also supported by protégé earlier versions that make use of F- Logic on harmony with open knowledge base connectivity protocol (OKBC). Some of the other ontology formats supported by protégé versions are RDF, OWL, XML, HTML schemas in which protégé is capable of exporting the ontology. [2] Some of the widely used plug-ins supported by Protege is mentioned as Jess Tab, Algernon, PROMPT, J save, Data Genie, OWL S-Editor, Word net, XML schema, Doc gen. The different plug-ins are explained as: Java

Class defines protégé classes are generated by J -save Plug-in. Protege web browser plug-in permits the protégé tool to share its ontology over internet. Word net plug-in acts an interface to the word net knowledge base. Protege knowledge is transformed into XML backend with the help of XML schema plug-in and UML plug-in. The creation of knowledge model from RDBMS with the use of JDBC is carried out using Data Genie plug-in. Doc gen Plug-in is used to create reports of ontologies. While running protégé we need to interact with Jess, this action is carried out using Jess stab plug-in which facilitates the use of Jess console window for interaction. The forward and backward inferencing of knowledge bases (Frame based) is supported by Algernon plug-in tab. In protégé multiple ontologies can only be managed using PROMT plug-in. Creation, management and visualization of OWL-S services can be carried out only with the support of OWL-S Editor plug-in.

### 5.1 VERSIONS OF PROTEGE

Till date Protege is released in the five successive series named Protege 1.x, 2.x , 3.x, 4.x, and 5.x. Protege 5 is the latest release by Stanford University in the month of June 2016. In the proposed thesis , the properties and functions of version Protege 2.1.2, protégé 3.3.1 and Protege 4.3 has been compared at its best. In every successive version, some new features had been summed up to upgrade the performance and visualization of the tool

### 5.2 PROTEGE ARCHITECTURE

Protege is a multi platform software means it can run on a broad range of platforms like Windows, Linux, Mac OS and Unix. Protégé is a java based tool. Its component list consist of ontology storage layer, graphical user interface, Application programming interface, a knowledge model layer and a set of users. Users can deal with ontologies in an easy way using graphical interface while all the application programs can interact with ontologies in protégé using API's. The protégé GUI finds a use of API to work with protégé knowledge model of ontology. Therefore all other components like plug-ins, GUI and user applications govern ontologies through same API. This makes the protégé architecture modular in nature and adds on to its flexibility. As shown in the figure 1 the plug-in architecture of protégé allows developers to modify its source code and thereby adding extra functionality to it. Hundreds of protégé plug-ins are available till date for import/export, validation and improved visualization. Protégé also facilitates the user with relational database backend that proves usefulness in storing huge ontologies.

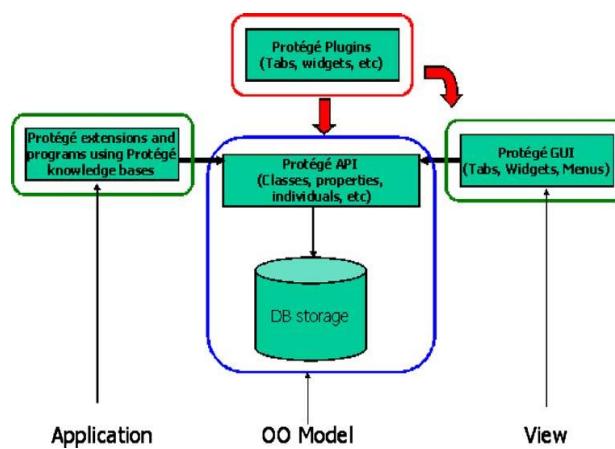


Figure 1: Protege architecture [7]

### **5.3 COMPONENTS OF PROTEGE TOOL**

Classes also called concepts .Classes/Concepts describe the common properties for the set of individuals. The classes follow a particular hierarchy and order using is-a relationship. The role of a class is interpreted in the form of binary relations between the objects and a class. The main difference in the class and concept is that classes are used in ontologies whereas concepts are used in description logic. Reasoner is that part of ontology [6] that allows the user to make logical inferences or we can call it reasoning. It is the only component that differentiates between two confusing terms called ‘machine readable’ and ‘machine understandable’. The major tasks performed by the DL reasoner are sub assumption, classification, instance checking, and satisfiability. For example if any concept defined both as human and non human can never be an instance.

### **6. OVERVIEW OF ONTOLOGY CREATED**

The main goal of research is to create an ontology related to tourism named ‘Make my trip.owl’. In order to achieve this, in the very first step tried to identify the main concepts of domain (listed below Entities widget). Basically protégé 4.3 consist of 11 widgets or tabs called as: Active Ontology, Entities, Classes, Object Properties, Data Properties, Annotation Properties, Individuals, Owlviz, DL Query, OntoGraf, Ontology differences.

Active ontology tab shows an overview of current ontology, also presents the annotations on the ontology as whole and also clears mentions if there is any other import of ontology(if any). This tab is useful when we are working with more than one ontologies. The next one is Entities tab. This tab is the sparkplug of the ontology editor from where you can explore all the classes, properties and individuals of an ontology. Each tab is made up of multiple views that can be resized and removed. In the protégé tool, backward and forward access is possible through arrow buttons on the top for an easy navigation similar to any browser. When anyone attempts to expand any concept, just a single click on that concept is required and the tree like structure appears to present the hierachal view. In the screenshot shown below (Figure 6), Extra-Events, Hotel\_Rating, Places, Room service and Packages are the main subclasses of the superclass Thing. Extra events have its own subclasses named camping, Bonfire, sports, water sports etc. where water sports and sports are equivalent classes. Places have the three four main sub classes name Deserts, Mountain area, Beaches and Plains. Beaches and deserts form disjoint classes means no place of beach can be a desert. The three buttons shown just above the main super class Thing are meant for creating a new super class, addition of a sibling class and third one is for deletion of any class. The annotation view allows the developer to add the description of a class. If we click on any of the property, like in this ontology Has\_Activity has been made symmetric in nature. This means if x is property of y then y is also property of x. Similarly, other characteristics that make the constructed ontology efficient are need to be checked in the check boxes shown in the interface. They are explained as follows:

- ✓ Functional and Inverse Functional: Functional means it can have only member. For example Hotel\_Rating can have only one value. It cannot be both silver and premium. Inverse functional is same as the antonym of functional. It implies that single member of domain can only correspond to single member of range.
- ✓ Transitive: This term in ontology mean that ‘if x has property y and y has property z then x has property z’.
- ✓ Asymmetry: It implies opposite mean to symmetry. This shows if x is property of y the it is not true in reverse manner as well.
- ✓ Reflexive: This describes that ‘x is property of x’.

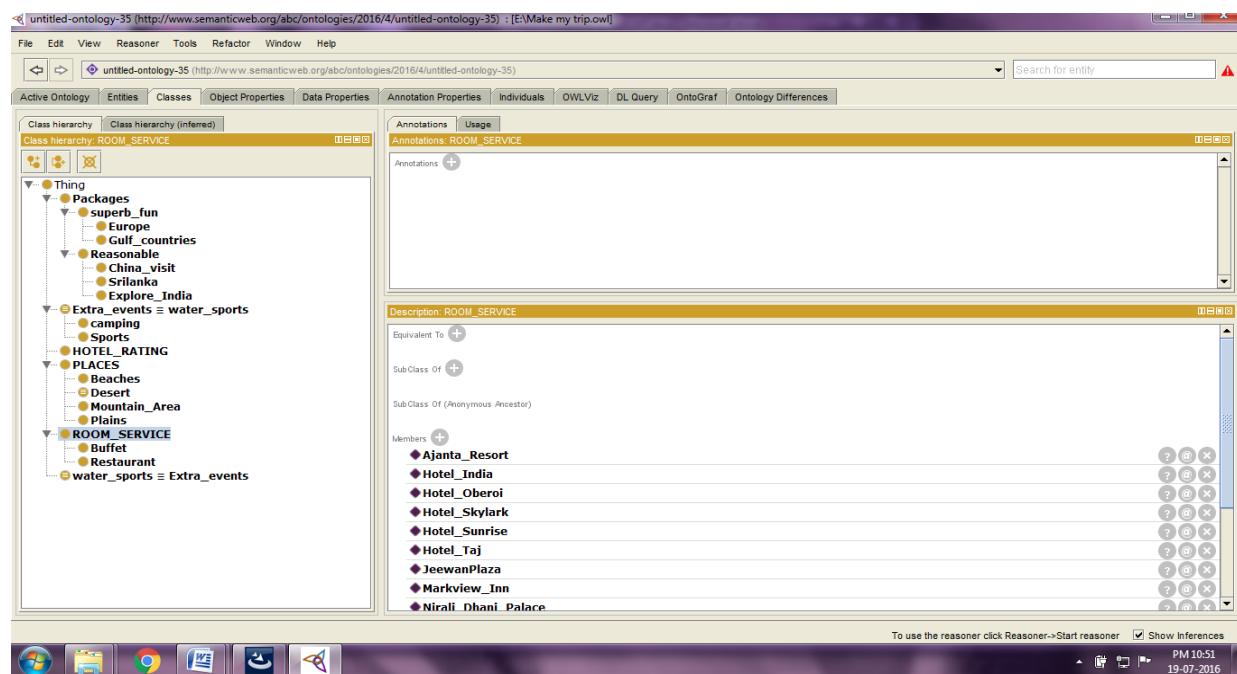


Figure 2: screenshot of protégé 4.3 showing the class hierarchy

Object properties connect one individual to another .These are the joining threads of an ontology. The constructed ontology consists of the following object properties as shown in table 3:

TABLE 3: Description of object properties

S. no.	Object Property	Description
1	Top object property	It is the main parent property class which contains all the children object properties.
2	Has_Activity	Each destination offers a specific activity. This object property connects the activities and destinations.
3	Has_Rating	This describes the ratings of every hotels(either silver , gold or premium)
4	HasPart	The option of Buffet, restaurant and room service are connected to every destination through this property.
5	Is_Offered_at	Some activities can be at specific destinations like rafting and paragliding.
6	Has_cost	This describes the extra cost incurred for every activity.
7	Contains_destinations	Destinations included in special packages are described by this object property.
8	Available_at_time	Some seasonal activities are for short period and available at time.
9	Suitable_for	This property describes the package cost suitable for different categories.
10	Has_distance	Distance metrics of each destination is explained through this object property.

## CONCLUSION:

In the above paper, the use of protégé has been mentioned in full detail and its concern with ontology creation and visualization. The use of ontology has been fully implemented in other advance applications related to semantic web where the domain ontology serves as building blocks of the basic constructed foundation. Protege is one of the effective editing and visualization tool when compared with others on the basis of numerous performance metrics. Querying data and retrieving relevant information can be the most efficient metrics of ontology as compared to other retrieving phenomenon and applications.

## REFERENCES

- [1] Catherine Roussey, Francois Pinet, Myoung Ah Kang, and Oscar Corcho, “An Introduction to ontologies and ontology Engineering”. *Verlag London Limited 201 , Advanced Information and Knowledge Processing 1, DOI 10.1007,© Springer.*
- [2] Thabet Slimani, “Ontology development –A comparing study on tools, languages and Formalism”..*Indian Journal of Science and Technology, September 2015, Vol8(24),DOI:10.17485/ijst/2015/v8i34/54249.*
- [3]Vishal Jain, Dr. Mayank Singh,“Ontology Development and Query Retrieval using Protege Tool”. *I.J. Intelligent Systems and Applications, August 2013, 09, 67-75 Published Online.*
- [4] Dr. Sunita Abburu, G.Suresh Babu, “Survey on ontology construction Tools”, *International Journal of Scientific & Engineering Research, June-2013, Volume 4, Issue 6.*
- [5] Sunitha Abburu,“A survey on ontology reasoners and comparison”, *International Journal of Computer Applications November 2012, (0975 – 8887) ,Volume 57– No.1.,*
- [6] R.SivaKumar and P.V.Aviroli, “Ontology visualization and protégé tool- A Review”. *August 2011, IJAIT.*
- [7] Bhaskar Kapoor and Savita Sharma(Department of Information Technology, MAIT, New Delhi, India). “A comparative study Ontology Building Tools for Semantic Web Applications.” *International Conference on Web Intelligence, 2005, IEEE (WI'05), 0-7695-2415-X/05 .*
- [8] Oscar Corcho, Mariano Fernandez Lopez, Asuncion Gomez- Perez, “Methodologies, tools and languages for building ontologies. Where is their meeting point?” *Data & Knowledge Engineering, 30 October 2002, 46 (2003) 41–64.*
- [9] James J.Wang, Farha Ali. “An Efficient ontology comparison tool for semantic web applications.” *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence 2005 IEEE.(WI'05)0-7695-2415-X/05.*

# Twins Biometrics: A Comparative Analysis of Various Methods

**Sandhya Vats**

Assistant Professor, Guru Nanak College, Budhlada

Email: profsandhyavats@gmail.com

**Abstract-**Automatic biometric systems based on human characteristics for identification have attracted the today's generation. The system is based on reliable and accurate verification of human physical characteristics such as iris, palm, fingerprints etc. As the population of twins increasing quickly and they have the closest genetic based relationship as well as the maximum similarity between their physical and behavioral characteristics. So, identical twins are a challenging problem with these biometric systems. The biometric systems that overlook the twins' problem are presenting a serious security hole.

**Keywords:** Identical twins, biometrics, face, fingerprints, face, iris, hand

## I. INTRODUCTION

Biometric Systems measures behavioral and biological characteristics for personal identification. Biometrics is more reliable than traditional token-based (access card) or Knowledge based(password). Traditional methods are easily identifying because password may be guessed and token may be stolen, lost and disfigured in use. On the other side biometrics, biological characteristics cannot be easily shared, forget and misplaced. For a physical or behavioral characteristics to be used for verification in automatic system, it must have the following properties:(i) universality(everyone possesses the characteristic),(ii) permanence (the characteristic remains invariant over life time),(iii) collectible (the characteristic is easy to capture), and (iv) distinctiveness (the characteristic is different for everyone)[4].

Twins have progressively increased in the past decades. Twins Birth rate has risen to 32.2 per 1000 birth with an average of 3% growth per year since 1990. It has been heard that there are two villages in India(Kerala) where in general only the twins born and no medical research have so far resulted the reason There are basically two types of twins:-

**Identical (Monozygotic) twins:** When one fertilized egg breaks and develops two babies with exactly the same genetic information that forms the identical twins.

**Fraternal (Dizygotic) twins:** When two eggs are fertilized by two sperms and produce two genetically unique children.

Identical twins are effected by many factors such as facial features which are very similar. On the other hand, some identical twins which don't share similar facial features but still have very much similarity. Confusion over their identities has made it difficult for others to make sure who owns and what and who does. What most importantly if

one of the identical twins commits a serious crime, their unclear identities create confusion and uncertainty in court trials to identify the culprit exactly.

There are some DNA based methods, to identify the identity of those identical twins like DNA fingerprinting technique. In DNA based method, first obtained a sample of cells, such as skin, hair or blood cells which contain DNA, then DNA is extracted from the cells and analyzed in the Lab and results are compared with DNA known samples.

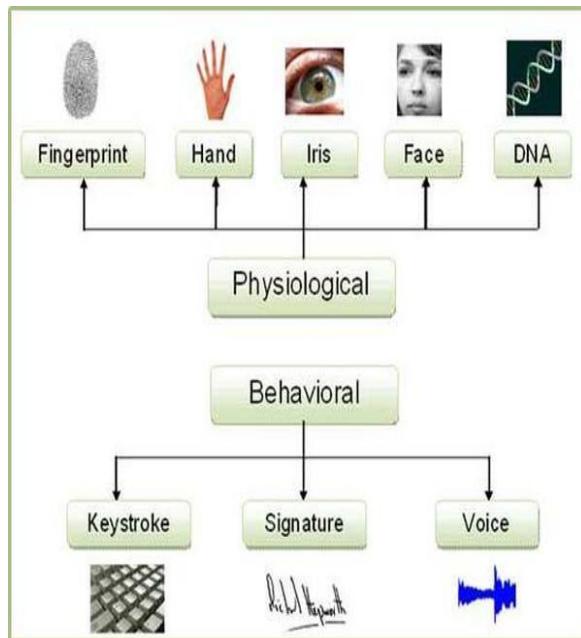


Figure 1. Various Type of Biometrics

## II. METHODS OF RECOGNIZING

A. *FACIAL MOTION*: Face Recognition is known to be a great challenge in identical twins. Psychologists concluded that faces with changing facial expression are significantly easier to recognize than static faces. Mainly the wrinkles of forehead, nose impression just to be to upper side and furious looking are the self speaking aspects of one's temper. Furthermore, research in Psychology conducted by Fraga et al. reports that identical twins may exhibit differences associated with different environments and lifestyles [4]. The test on face recognition is based on two real life scenarios.

- (a) *ACCESS CONTROL SCENARIO*: In this scenario, one twin is an authorized user and another is unauthorized. It is a great challenge for security to give access to the right twin and deny the unauthorized twin without prior knowledge of twin siblings. In this, we evaluate the performance in terms of Twins-Error.
- b) Social Party scenario: In this, the pair of twins who look and dress alike. We know that the pair is twins in social party but in access control it is not known. Correct guess about twins is evaluated by measuring accuracy.

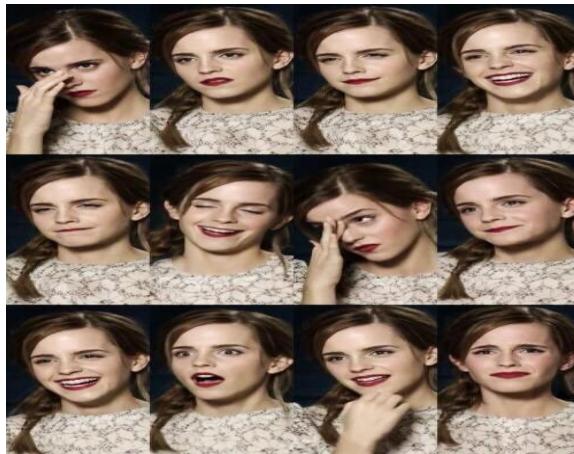


Figure 2: Face expression

There are two tests based on the facial motion:

- Simple Displacement Algorithm: It uses the sparse displacement as a feature. In this several key points are tracked for different face expression.
- Dense Displacement Algorithm: It uses the dense displacement and deformation on the entire face as a feature. In DDA, face is tracked by the video and eye used for center position alignment.

**B. IDENTIFICATION OF IDENTICAL TWINS FROM FINGERPRINT:** “Fingerprints cannot lie but liars can make fingerprints”. Fingerprints are taken into count as proof of evidence in courts of law all over the world and are taken as unchangeable. As now it is the progressive biometric technology. Identical twins have the closest genetic relationship but they have no same fingerprints as it is universal truth that fingerprints cannot match with each other. Finger print Formation: Fingerprints are fully formed about 7 months of fetus development and finger ridge configurations do not change throughout the life except due to accidents such as bruises and cuts on the finger tips[5]. In the criminal investigations, finger prints are routinely used by forensic science labs.

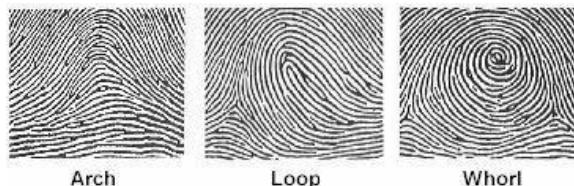


Figure 3: Finger prints

Finger print is classified into three parts:

- Whorls:-In this pattern the ridge are usually circular
- Arches:-In this pattern from one side the ridges entered make a rise in the center and opposite side generally exit.
- Loops:-In a loop pattern from the other side ridges enter ,re-curve and pass out the same side.

Classifications of Finger print matching:

- a) Rooted finger print matching
- b) Partial finger print matching

Cells on the fingertip are slightly differing from hand to hand and finger to finger. Fingerprints are used in forensic science labs for criminal investigations. In the fingerprints case the pattern of the general characteristics determine by the genes. Like the growth of blood vessels and capillaries finger print formation is similar. During the formation of finger prints there are so many changes so it would be impossible two fingers to be same. Finger print formation process is a chaotic system not a random one. Overall finger print pattern are classified into five major classes: whorl, left loop, right loop, arch, tented arch. Based on ridge depth, ridge thickness and ridge separation fingers are identified at a finer level. The two points on the finger e.g., delta and core measure the number of ridges this is called as the ridge count feature. By using the most widely method minute details finger print similarity can be distinguish then the minutiae similarity is high when the two fingers are same. Unrelated persons have very little generic similarity in their fingerprints. But parent-child have some similarity because half of the genes shared by the children. The monozygotic (identical) twins have more similarity due to the closest genetic relationship. Fingerprints of identical twins show genetic similarity because their development from the same DNA. However during development identical twins are situated in any different part of the womb. Only the expert can differentiate the fingerprints based on minutiae (dis)similarity. An iris-based biometric system performance evaluated by Daugman. It is not easy for automatic system to accurate the position of the ridge in image of fingerprint. Ridge location depends on the image quality of the finger print.

*C. IDENTIFICATION OF IDENTICAL TWINS FROM EAR IMAGES:* The shape of the ear does not change after the adulthood. Its surface has uniform color. The most well-known pioneer seems to be Iannarelli (1989) in which he performed manual identification over 10,000 ears and found no indistinguishable ears.[3] On the basis on Psychological model it's suggested that face perception in human being known as Expectation Report Model (ERM). ERM described by two main suggestion.

1. Do accurate face recognition by focusing less on normal features and more on abnormal features.
2. The optimality of the use of only 105 of the features (only abnormal features) for rapid and accurate recognition (ERM optimality) [2] ERM used in some automatic method for face recognition but not for ear recognition.

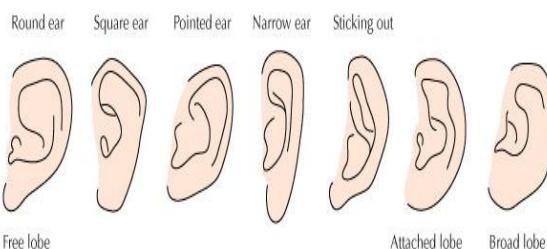


Figure 4. Structure of ears

There are two parts consist in our proposed system.

- (a) In the first part shape and appearance of the ear recognition.
- (b) In the second part weight points in the ear shape according to their abnormality level is recognized.

The ERM suggests that the importance of a feature has a direct relationship with the abnormality of that feature and the further a point is from its related mean value, the more abnormal it is.

*D. IDENTIFICATION OF IDENTICAL TWINS FROM IRIS:*-Identical twins can be distinguished by automated iris biometrics. Dr. John Daugman advertises that “No two irises are alike”. There is no detailed correlation between the iris patterns of even identical twins or the right and left eye of an individual. Iris is an important privacy concern in the degree of similarity between twins. An image of iris is highly protected because it is not labeled with a name. The iris is physically small in size but well designed optical systems. The systems work to magnify a high-resolution image. The diameter comes to 200 to 300 pixels. There are also many minute features that are i.e crypts, furrows, coronas, freckles and strips, etc. These are contained in this reason which forms unique iris texture for each eye. The visual pattern of a human iris consists both color and texture. The color has limited discriminating power for recognition. The grey-level iris images captured under near infrared illumination. This texture pattern is for personal identification.



Figure 5. Iris expression

The iris texture pattern takes formation to become stable after the eighth month of gestation and this very pattern formation is determined by the gestation environment. Iris is a photographic biometric trait. The identical twins can be determinate using suitable iris features as sometimes we find right and left iris of the same person are different. Moreover it has been observed in many cases that twins linking are the same, their movements also fully agree with each other as sitting, standing, walking, gait and talking. Even their sound is not recognizable in such cases. Sometimes, we stress on mind and call by them name to have exact identification. Everyone is deceived in such cases and the Nature it seems has put the human being on test.

### III. CONCLUSION

After going through the given circumstantial evidences regarding twins one main thing of this subject seems to be outcome of forensic science. If we go in detail whatever we have taken regarding advantages of twin identification and its need it goes to criminology science. Many times it becomes difficult to recognize the real culprit on the resemblance and characteristics stand by us for correct identification. It is the most useful and unchangeable accept of the subject and facilitation to curb the crime with correct identification of the accused. Twin biometric theory is

entirely based on the use of forensic science which has played a vital role in today word of crimes and this theory has given a great success to maintain the law and order for the masses to live peacefully and fearlessly. When we have no crime then there is only the peace for which today's world is going to achieve and it is the demand of every human being on earth

#### REFERENCES

- [1] Adams Wai-Kin Kong,David Zhang,Guangming Lu, "A study of identical twins' palmprints for personal verification,"Elsevier Volume 39,Issue 11,Nov 2006.
- [2]K.Arthi,N.M.Nandhitha,S.EmaldaRoslyn, "A Study and Evaluation of Different Authentication Methods and Protocols, "IJCSMR, Volume 2,Issue 1 January 2013.
- [3]Hossie Nejati,Li Zhang,Terence Sim,Elisa Martinez-Marroquin,Guo Dong,"Wonders Ears:Identification of Identical Twins from Ear Images,"21<sup>st</sup> International Conference on pattern Recognition(ICPR 2012)Nov 11-15.2012.Taskuba,Japan
- [4]Li Zhang,Ning Ye,Elisa Martinez marroquin,Dong Guo,Terence Sim,"New Hope for Recognizing Twins by Using Facial Motion,"
- [5]Anil K.Jain,Sali Prabhakar,Sharath Prankanti,"On the similarity of identical twin fingerprints,"www.elsevier.com/locate/patcog.Patten Recognition 35 (2002) 2653-2663.
- [6] Swati Y. Dhote, A. D. Gotmare, M. S. Nimbarate,"Differentiating Identical Twins by Using Conditional Face Recognition Algorithms," International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.
- [7] Kong, A. W.-K., Zhang, D., and Lu, G., \A study of identical twins' palmprints for personal veri~cation,"*Pattern Recognition* 39, 2149{2156 (2006).
- [8] Srihari, S. N., Srinivasan, H., and Fang, G., \Discriminability of ~ngerprints of twins," *Journal of Forensic Identifi~cation* 58(1), 109{127 (2008).
- [9] A. W. Kong, D. Zhang, and G. Lu, "Study Of Identical Twins Palmprints For Personal Verification ", *Pattern Recognit.*, vol. 39, no. 11, pp. 2149–2156, Nov. 2006.
- [10] U. Park and A. Jain, "Face Matching And retrival Using Soft Biometrics", *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 3, pp.406-415, Sep-2010.
- [11]<https://www.comp.nus.edu.sg/~tsim/documents/wonder-ears.pdf>
- [12]<http://vision.unipv.it/VA/VA-En/TEST/Stereo-baseline%2065,%20focal%20distance%2035/Twin%20ears.pdf>

# GENETIC ALGORITHM WITH ELITISM FOR VEHICLE ROUTING PROBLEM

Sandhya  
CSE Dept., M.M.E.C  
Maharishi Markandeswar  
University,Mullana, India  
Sandhya12bansal@gmail.com

Rajeev Goel  
CSE Dept.  
ACE, Mithanpur  
Ambala,India  
rcse123@gmail.com

Virat Rehani  
CSE Dept.  
CT group of Institues  
Jalander,India  
vrehani@yahoo.com

## *Abstract*

**Vehicle Routing Problem (VRP)** is an optimization problem in operational research where customers of known demands are supplied by a central depot. The aim of the problem is to minimize total route cost while satisfying the capacity constraint. Vehicle Routing Problem with Time Windows is an important variant of VRP. In this every customer should be served within its time windows along with capacity constraint. This is an NP-hard problem, thus many meta heuristic approaches have been proposed to find the optimal solution. In this paper Genetic algorithm (GA) has been used to solve the problem. An attempt has been made to improve already existing GA. The proposed algorithm (EGA) uses new representation scheme and elitism mechanism to solve the problem. The proposed approach is validated on standard Solomon's benchmark problems and computational results shows that the proposed approach is effective and efficient.

**Keywords:** Vehicle routing problem, Time Windows, Genetic algorithm, Elitism.

## I. INTRODUCTION

The Vehicle Routing Problem (VRP) is a complex combinatorial optimization problem that aims to provide services to a number of customers with a given number of vehicles. The problem was first introduced by Dantzig and Ramser [1] in 1959. VRP arises in the fields of transportation, distribution and logistics. It has attracted considerable attention in recent years due to its rich applications in problems that involves routing and scheduling in constraint environment like transport of goods, persons, bank deliveries, solid waste management and many more. The objective of the VRP is to deliver a set of customers with known demands on minimum-cost vehicle routes originating and terminating at depot. Till now many variants of VRP are proposed. VRPTW is one of them. In VRPTW, a set of  $K$  homogeneous vehicles having fixed capacity  $Q_i$  are required to serve  $N$  geographically scattered customers having fixed demand  $q_i$  in such a way that all customers are served within their time windows  $[e_i, l_i]$ . A vehicle may arrive before opening time  $e_i$  but in that case it has to wait. No services can be made after closing time  $l_i$ . The objective of VRPTW is to minimize the number of vehicle used and to minimize the total travel distance such that all customers are served once and only once, all routes originates and

terminates at depot while preserving the capacity and time window constraint. Figure 1 shows the pictorial representation of VRPTW.

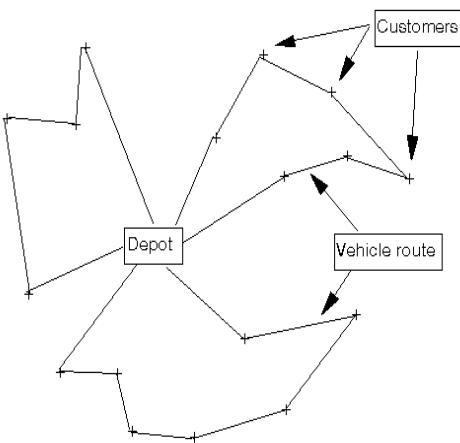


Figure 1 Vehicle Routing Problem

As VRPTW is NP hard problem, many methods are proposed for solving it. There are mainly two categories: Exact and meta heuristics. Exact methods are used for providing solution to small instances of VRPTW. Significant improvements in Solomon's benchmark problem instances were established by Kohl et al. [2], Larsen[3] and Chabrier [4] using exact methods. However their performance degrades when numbers of customers are increased. For solving large size problems meta heuristics are used. These methods provide optimal solution within a time bound. Survey of solving VRPTW using heuristics and meta-heuristics approaches has been presented in [5, 6, 7, 8]. In [9] [10] and [11] Taillard et al.; Chiang and Russell and Cordeau et al. repectively proposed a tabu search based solution for VRPTW. Gambardella et al. [12] proposed an ant colony optimization approach for solving the problem while [13] Shaw presents a large neighborhood search for finding the solution. Many hybridized meta heuristics [14, 15, 16] were also proposed for solving the problem. The GA approach was introduced in 1975 by Holland. GA is an effective stochastic approach that is based on the principal of survival of fittest genes that compete each other for resources. Moreover it is an effective technique that produces efficient results for NP hard problems like VRP. The first application of hybrid GA to VRPTW was given by Blanton and Wainwright. They hybridized it with greedy technique. Many hybridized algorithms have been presented for solving the problem by many researchers [17, 18]. In [19] Gehring and Homberger presented a parallelization of a two-phase meta heuristic for solving VRPTW. In previous approaches [17, 19] best solution is lost. In this paper a new representation scheme for GA and a new approach for preserving the best solution have been presented.

Rest of the paper is organized as follows: section II will present the proposed representation scheme of individuals and elitism based GA (EGA) for solving the problem. Section III presents the experimental results obtained in solving some of the Solomon's benchmark problem by using EGA. Finally conclusions are drawn in section IV.

## II. ELITISM BASED GENETIC ALGORITHM (EGA)

Among various heuristics and meta heuristics approaches proposed for solving VRPTW, GA has been widely used. There are three main GA operators: Selection, Crossover and Mutation. GA is an adaptive heuristic which roots its ideas from process of natural selection and genetics. Selection operators derives

the GA towards the better solution, crossover operators are used for deriving new population for next generation and finally mutation operators are used to escape from local optima. However previous researches on VRPTW with GA [17, 19] loses best solutions during crossover and mutation. Often the proposed solution will rediscover these lost improvements in further iterations but there is no guarantee. To overcome this problem, this paper uses elitism based GA (EGA). This approach preserves the fittest candidate for next generation and replaces the worst candidates of new population by best candidates preserved.

The general algorithm of EGA can be shown as:

### *Algorithm EGA*

**Step 1:** Initialize population P randomly according to distance.

**Step 2:** Calculate fitness of each population.

**Step 3:** Repeat step 4-7 until termination condition is met.

#### **Step 4:** Selection (Tournament Selection).

**Step 5:** From these find fittest candidate ( $S_{best}$ ) and preserve it for next generation.

**Step 6:** For remaining perform crossover (route exchange) and mutation(Swap, Flip and Slide) and obtain a population P temp.

**Step 7:** Calculate fitness of each candidate of Ptemp. Compare it with Sbest. If the fitness of each candidate of Ptemp is less than fitness of Sbest, then replace the worst string of Ptemp with Sbest.

**Step 8:** If termination condition is not met goto step 3. Otherwise goto step 9.

### **Step 9: Generate best route.**

Finally the flowchart of overall algorithm is depicted as shown in Figure 4 in Appendix I.

**Initial Population:** An initial population is build randomly on the basis of distance in such a manner that time window and capacity constraint of the problem are met.

**Chromosome and Individual Representation:** Chromosome of GA is represented by an integer identifier  $i$  where  $i < N$ . Collection of chromosomes called an individual represents a complete routing sequence. While binary strings are generally used to represent individuals, where each chromosome represents a customer. A two part individual is used to represent solution as shown in figure 2. The individual is divided into two parts by a zero number. The first part is the customer- part. This contains the several routes, each of them representing a sequence of customers that must be covered by vehicle. The second part is the vehicle part. The number in each of chromosome represents the length of corresponding route. For e.g in figure 2 total two vehicles are used and first vehicle will cover first two customer's i.e 1 and 5. Second vehicle will cover next three customers

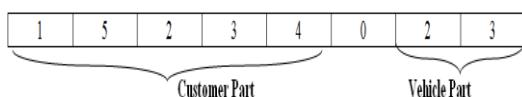


Figure 2 Individual Representation scheme

**Selection:** We used tournament selection. In this we divide the entire population into distinct set of four members each. All the four members are kept sorted. This means the fittest member is on the top position

and the weakest member is at the bottom. This ensures that the fittest member is directly inherited with its original form. The tournament selection scheme for choosing fittest member is used with a probability of  $p_s=0.75$ .

**Elitism:** In this process the fittest individuals are preserved for reproduction in next generation. This guarantee that the fittest solution obtained from the current population is copied unchanged in the next population and guarantees that solution quality obtained by GA will not decrease from one generation to the next rather the solution quality increases after every iteration. Remaining three worst populations undergoes crossover and mutation operations. Finally the fittest individuals and the mutated individuals are combined to obtain the solution.

**Crossover and Mutation:** Single or double point crossover is not suitable for VRPTW because of duplication of customers that may result into in-feasible routes. In this paper route exchange crossover is applied. In this operator an effort has been made to interchange the chromosome of individual having minimum number of customers in each of the remaining chromosomes of the individual such that time window and capacity constraints are observed. Any duplication of customers is deleted to ensure feasibility of routes.

Mutation is done so that search does not confine to limited area. In this paper 3 mutation operators namely: swap, flip and slide are performing on the worst individuals with equal probability. The swap operator randomly selects two customers and swaps them if they result into a feasible route. Flip operator selects two randomly points and reverses the order of visting the customers between these points. Finally the slide operator slides two randomly selected sequences. All the three mutation operators are shown in figure 3 (a-c).

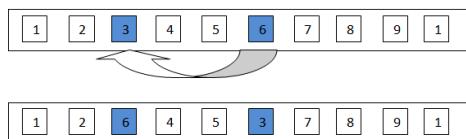


Figure 3(a) Swap Operator

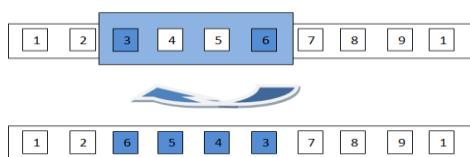


Figure 3(b) Flip Operator

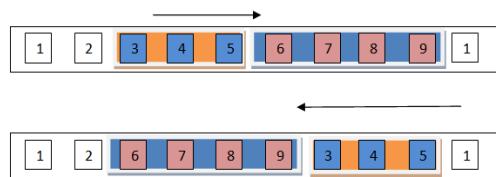


Figure 3 (c ) Sliding Operator

### III. COMPUTATIONAL RESULTS

In this section we summarize the results obtained by using EGA for solving VRPTW. The proposed algorithm is coded in MATLAB and implemented on a core i3 processor. The EGA has been tested on all the instances of Solomon benchmark problem sets involving 100 customers. The problems vary in available fleet size, vehicle capacity, traveling time of vehicles, spatial and temporal distribution of customers to be served. In classes R1 and R2 customers are randomly scattered, while the customers are clustered in C1 and C2 datasets. RC1 and RC2 are combination of R and C datasets. Each customer  $i$  has a time window  $[e_i, l_i]$ , representing the earliest as well as the latest service time at  $i^{th}$  customer. Each instance is executed ten times and obtained results are compared with best known results and OCGA [17]. Following simulation parameters have been used: Population Size = 100; Generation size = 1000; Crossover rate = 0.75; Mutation rate = 0.25.

Table 1 presents the summary of results obtained by using EGA. Here in first column results available in literature are shown. Second column of the table presents result obtained using OCGA and third column presents result obtained using EGA. Here NV and TD denote the number of vehicles used and total distance traveled respectively.

Table 1 Comparison of EGA, Best Known and OCGA for 100 customers

S.No.	Problem	Best Known [20]		OCGA [17]		EGA	
		NV	TD	NV	TD	NV	TD
1	C101	10	827.3	10	828.4	10	828.2
2	C102	10	827.0	10	828.94	10	827.39
3	C201	3	589.1	3	591.6	3	590.3
4	C202	3	589.1	3	591.5	3	591.5
5	R101	20	1637.7	19	1650.8	20	1642.8
6	R102	18	1466.6	17	1486.1	17	1486.6
7	R201	8	1443.2	4	1252.4	8	1447.4
8	R202	4	1088	6	1079.3	4	1191.7
9	RC101	15	1619.8	14	1697.0	15	1619.2
10	RC102	14	1457.4	14	1496.2	10	2221.2
11	RC201	9	1261.8	7	1406.9	9	1266.6
12	RC202	8	1092.3	7	1176.5	8	1146.3

The proposed algorithm produced improved results in clustered problem with both small and large time windows whereas in case of mixed problem there is a tradeoff between NV and TD. If TD increases the NV decreases and vice versa. On the other hand comparable results are obtained for random problem.

#### IV. CONCLUSION

This paper proposes elitism based genetic algorithm for solving VRPTW. A new representation scheme of individual and three mutation operators have been proposed. The proposed algorithm has been tested on a number of benchmark problems available in literature and compared with OCGA as well as other available best solutions. Finally the computational results demonstrated that the presented approach is competitive with other available complex meta heuristics in terms of the quality of obtained solution. In future work, it may be interesting to enhance EGA with other variants of VRPTW such as rich VRP and/or stochastic VRP etc.

#### References

- [1]. Dantzig, George B., and John H. Ramser. "The truck dispatching problem." *Management science* 6.1 (1959): 80-91.
- [2]. Kohl, Niklas, et al. "2-path cuts for the vehicle routing problem with time windows." *Transportation Science* 33.1 (1999): 101-116.
- [3]. Larsen, Jesper. *Parallelization of the vehicle routing problem with time windows*. Diss. Technical University of DenmarkDanmarks Tekniske Universitet, Department of Informatics and Mathematical ModelingInstitut for Informatik og Matematisk Modellering, 1999.
- [4]. Chabrier, Alain. "Vehicle routing problem with elementary shortest path based column generation." *Computers & Operations Research* 33.10 (2006): 2972-2990.
- [5]. Bräysy, Olli, and Michel Gendreau. "Vehicle routing problem with time windows, Part II: Metaheuristics." *Transportation science* 39.1 (2005): 119-139.
- [6]. Kumar, Suresh Nanda, and Ramasamy Panneerselvam. "A survey on the vehicle routing problem and its variants." *Intelligent Information Management* 4.3 (2012): 66.
- [7]. Sandhya, Vijay Kumar. "Issues in Solving Vehicle Routing Problem with Time Window and its Variants using Meta heuristics-A Survey." *International Journal of Engineering and Technology* 3.6 (2013): 668-672.
- [8]. Katiyar, V. "Relative Performance of Certain Meta Heuristics on Vehicle Routing Problem with Time Windows." *International Journal of Information Technology and Computer Science (IJITCS)* 7.12 (2015): 40.
- [9]. Taillard, Éric, et al. "A tabu search heuristic for the vehicle routing problem with soft time windows." *Transportation science* 31.2 (1997): 170-186.
- [10]. Chiang, Wen-Chyuan, and Robert A. Russell. "A reactive tabu search metaheuristic for the vehicle routing problem with time windows." *INFORMS Journal on computing* 9.4 (1997): 417-430.
- [11]. Cordeau, Jean-François, Gilbert Laporte, and Anne Mercier. "A unified tabu search heuristic for vehicle routing problems with time windows." *Journal of the Operational research society* 52.8 (2001): 928-936.
- [12]. Gambardella, Luca Maria, Éric Taillard, and Giovanni Agazzi. "MACS-VRPTW: A multiple ant colony system for vehicle routing problems with time windows." (1999).
- [13]. Shaw, Paul. "Using constraint programming and local search methods to solve vehicle routing problems." International Conference on Principles and Practice of Constraint Programming. Springer Berlin Heidelberg, 1998.
- [14]. Marinakis, Yannis, and Magdalene Marinaki. "A hybrid genetic–Particle Swarm Optimization Algorithm for the vehicle routing problem." *Expert Systems with Applications* 37.2 (2010): 1446-1455.
- [15]. Lau, Henry CW, et al. "Application of genetic algorithms to solve the multidepot vehicle routing problem." *IEEE Transactions on Automation Science and Engineering* 7.2 (2010): 383-392.
- [16]. Lee, Chou-Yuan, et al. "An enhanced ant colony optimization (EACO) applied to capacitated vehicle routing problem." *Applied Intelligence* 32.1 (2010): 88-95.
- [17]. Nazif, Habibeh, and Lai Soon Lee. "Optimized crossover genetic algorithm for vehicle routing problem with time windows." *American journal of applied sciences* 7.1 (2010): 95.

- [18]. Derbel, Houda, et al. "Genetic algorithm with iterated local search for solving a location-routing problem." Expert Systems with Applications 39.3 (2012): 2865-2871.
- [19]. Homberger, Jörg, and Hermann Gehring. "A two-phase hybrid metaheuristic for the vehicle routing problem with time windows." European Journal of Operational Research 162.1 (2005): 220-238.
- [20]. <https://www.sintef.no/projectweb/vrptw/solomon-benchmark/100-customers/>.

## Appendix I

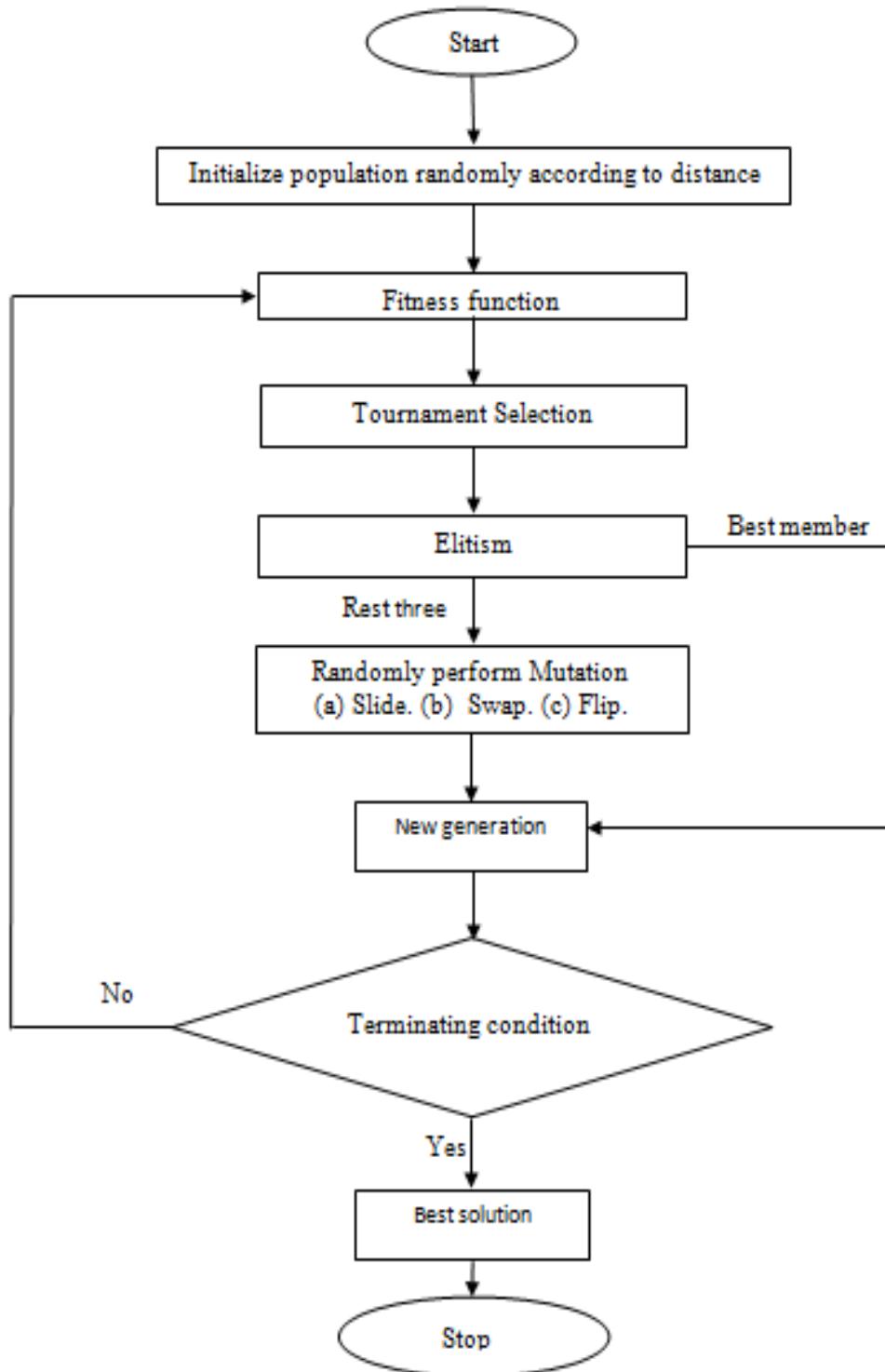


Figure 4 FLOW CHART OF EGA ALGORITHM

# Extensive Survey of Simulators available in Distributed Environment

Subhash Chander <sup>1</sup>, Sunil Kumar <sup>2</sup>

<sup>1</sup> (Assistant Professor, Department of Computer Science, University College, Jaito, Punjab)

<sup>2</sup> (Assistant Professor, P.G. Department of Computer Science, G.N. College for Girls, Muktsar, Punjab)

<sup>1</sup>sumit.mittal@mumullana.org, <sup>2</sup>sarora66078@yahoo.co.in

**Abstract:** Cloud computing is emerging technology in the present scenario, which is gaining more popularity day to day. As its user base is increasing at rapid speed it leads to more complexities to fulfill the demands of users. Cloud computing has to face performance, security, load balancing, energy efficiency and cost management related problem to maintain its sustainability in the present competitive environment. Cloud Simulators plays the vital role in the analysis of cloud related problems with lowest cost with the help of designing the virtual environment without using the real hardware and software resources due to its high cost. In the present era, there are various type of Cloud simulators has been proposed by various eminent authors, In this paper, an extensive survey has been performed to find the suitable platform for the further research work.

**Keywords:** Simulators, Cloud Computing, Grid Computing, Distributed Computing

## 1. INTRODUCTION:

Cloud computing has become imperative element in the field of large set of computational activities in the real world. Cloud offers variety of services like SaaS, PaaS and IaaS. All of these technologies provide the services to the users as per their demand and as per their usage. In present scenario various types of cloud deployment has been performed private, public and hybrid to manage the user needs. As per recent survey private clouds are increasing at swift pace. To achieve the efficiency in the cloud computing, there are various prime elements like security, cost modeling, energy management and virtual machine migration have become the major concern for the researcher under review [1, 2]. Researchers are working for the betterment of cloud in arena of security, cost and energy, but experiment for betterment of these parameters in real cloud environment can lead to high cost, Cloud simulators are only the prominent solution for this problem. Simulators provide the virtual environment for the experiment at minimal cost with actual parameters of cloud. In the present scenario, Simulators are considered as the third pillar of science. A simulator provides the virtual environment which can lead to controllable as per requirement and can be repeated as requirement at minimal cost [3]. Various eminent authors have proposed various cloud simulators to tackle the cloud related problem. These Simulators can be broadly classified as GUI based/ Non-GUI based as shown in figure 1.

In this paper, an extensive survey has been done to present motivation behind the designing of some of renowned simulators with its features so that it can be help to choose the accurate simulator for the specific research problem.

## 2. RELATED WORK:

In the arena of huge computational activities on the cloud environment, it is impossible to make the experiment for the betterment of cloud services, directly on the real cloud environment due to its high cost. In the real world various cloud simulators are existing to manage this problem such as CloudSim[4], CloudAnalyst[5], iCanCloud[6] and many other. Most of the Simulators are compatible to java coding but some of simulator like CloudAnalyst are GUI based they doesn't require any coding to give the results but they are limited in use. In [7] authors have given the complete survey of twelve prominent simulators used for the cloud computing related issues. Under the survey authors also had given the complete comparison of these simulators in term of cost modeling, communication model like nine different parameters. Authors conclude that most of the simulators are not working for mobile cloud computing so there is great scope to design an simulators which also work better for the mobile cloud computing. In [8] authors given the comparison of fourteen cloud simulators, in terms of their features so that readers can choose the appropriate cloud simulator as per their problems. At the end the authors defined that there is no simulator which can manage as whole there is need to collaborate the various simulator or design the cooperative simulation tools. In this paper extensive survey of simulation tools has been performed to become helpful in selection of appropriate tool or more tools as per problem of user.

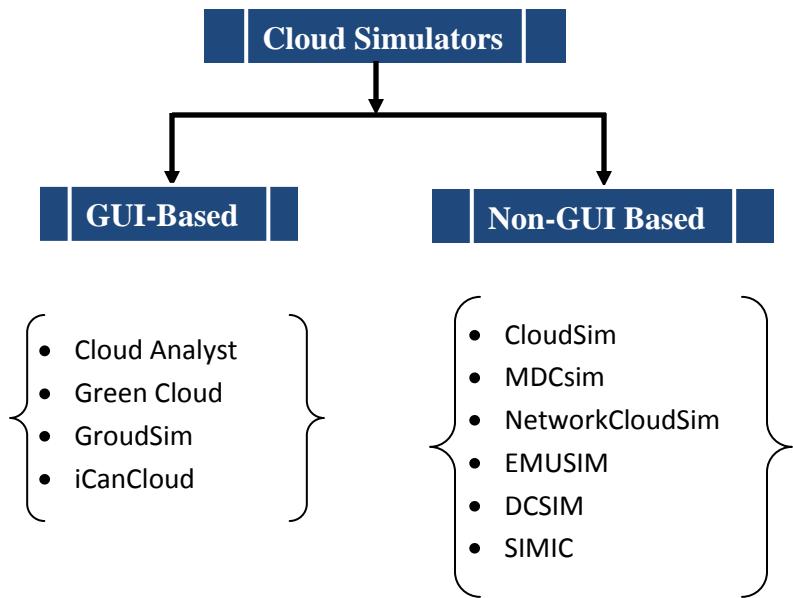


Figure 1: Classification of Cloud Simulators

## 3. CLOUD SIMULATORS UNDER STUDY:

This section deals with the extensive survey of various distinguished cloud simulators, arrangement of simulations technologies based on their popularity and usage.

A. *CLOUDSIM:*

This is one of the most widely used simulators for cloud computing related issues. It is complete toolkit with wide number of classes to design the virtual cloud environment by setting the different data centers, virtual machines and numbers of user base with different policies for scheduling and provisioning. It is not a GUI based environment so the user will setup the complete environment as per their requirement in the java code means user can redefine the classes as per their requirement to get the simulation as per their requirement. CloudSim is helpful to research in study of any specific problem without worrying about the low-level complexities of system. With the help of Cloudsim researchers can proposed various new resource allocation and scheduling policies and have capability to switch between the space shared and time shared services [4].

B. *CLOUD ANALYST:*

It is one of the GUI based popular tool for the simulation in cloud environment. Cloud Analyst was developed in the java to provide the user friendly environment to the user. It gives the complete GUI environmental ease for user to set the parameters for cloud related problems' experiments. It was developed by Wickremansignhe et. al. [5] in 2009 to provide the GUI based environment to the researcher for performing the cloud based simulation deployed on the idea of CloudSim. After the three year of development of CloudAnalyst, Rawat[9] evaluate the complete performance of social networking sites by adding up the new brokering policy to it. This platform gives the flexibility to researcher to set the parameters and achieve the simulation results without bothering about the coding for simulation.

C. *GREENCLOUD:*

Simulators described before this are the general simulators but this is specific simulators which works for the energy aware cloud computing. GreenCloud [10] works like its name it lead to the problems works for the energy management, which has become the prime concern in eminent researchers. Basically it is improvement in the one of renowned simulator NS-2. It gives the better performance in finding the solution for resource allocation, workload management as well as optimization of communication protocol and network infrastructure. The major drawback in this simulator, to run applications on this simulator research should have knowledge of C++ and OTcl.

D. *ICANCLLOUD*

iCanCloud come out to overcome the drawbacks of renowned simulators like CloudSim[4], GreenCloud[10]and MDCsim[11]. Flexibility is the major feature of this simulator [6], in this simulator user can modify the core class named as hypervisor to run the simulation as per requirement. It was developed by the group of researchers named as ARCOS at one of the renowned university of

Spain. It provides the capability to make comparison with a corporate model. The main feature of this simulator is that it works on C++ which is considered as basic language for programming knowledge.

*E. GROUDSIM*

This is one of the simulator which works for the both of eminent technologies of distributed system i.e. Grid Computing and Cloud Computing. GroudSim [12] works on prominent programming language java, which is used by most of the simulators, so that it is easy work on this platform. Extendibility through the probability distribution packages is main feature of Groudsim. Various researchers have used this simulator to design the proposed work in the field of Cloud and Grid computing.

**4. ANALYSIS OF CLOUD SIMULATORS:**

In the above section, brief introduction about the five prominent simulators in the arena of Cloud and Grid Computing. In this section, an extensive analysis of these technologies has been done on the basis of some basic parameters such as motivation, feature, availability and GUI and Programming language used in the specified simulators. Table 1 is giving the complete extensive analysis of these simulators to give the brief idea to user in selection of simulator for their research issue.

**5. CONCLUSION:**

In this paper, an extensive analysis of five prominent simulators has been performed. After the complete study, It has come out no one simulators can be defined as best simulator because all of them has their own capability and drawbacks. So, selection of the simulator should be based on the research area of problem and as per suggestion come out after analysis use of one or more simulator can be more helpful to achieve the accurate results.

**TABLE 1**  
**EXTENSIVE ANALYSIS OF CLOUD SIMULATORS**

Simulator	Motivation	Features	Availability	GUI/ Non-GUI	Programming Language
CloudSim	<ul style="list-style-type: none"> <li>To endow with comprehensive and extensible simulation framework.</li> <li>Make possible the scientific experiment possible with minimal cost.</li> </ul>	<ul style="list-style-type: none"> <li>It is capable to perform the computation of large scale.</li> <li>Capable to customize policies for resource provisioning, task scheduling.</li> <li>Provide the capability to perform energy aware computational resources.</li> <li>Dynamic in nature.</li> </ul>	Open Source	Non-GUI	Java
CloudAnalyst	<ul style="list-style-type: none"> <li>Lack of tool which provide the user friendly GUI environment for simulation with minimum number of code line.</li> </ul>	<ul style="list-style-type: none"> <li>GUI based environment and gives the capability to set parameters like CloudSim.</li> <li>Ease of repeatability for experiment.</li> <li>Due to open source availability there is lot of scope for extension.</li> </ul>	Open Source	GUI	Java
GreenCloud	<ul style="list-style-type: none"> <li>Due to lack of simulator which is capable to work especially for energy aware environment.</li> </ul>	<ul style="list-style-type: none"> <li>Especially works for energy aware data centre based environment.</li> <li>Improvement over the renowned simulator NS2.</li> <li>Complete communication model with support of TCP/IP.</li> <li>Based on the Energy saving models (DVFS and DNS)</li> </ul>	Open Source	Limited	C++/Octel
iCanCloud	<ul style="list-style-type: none"> <li>Due to drawbacks in CloudSim, CloudAnalyst and MDCsim.</li> <li>Predict the trade-offs between cost and performance of application on specific hardware.</li> </ul>	<ul style="list-style-type: none"> <li>Flexible in nature and authorized user to customize the core file hypervisor or improvements.</li> <li>Provide GUI based environment</li> <li>New components can be added for betterment of results</li> </ul>	Open Source	GUI	C++
GroudSim	<ul style="list-style-type: none"> <li>Due to lack of platform which can work both of computing techniques Cloud and Grid.</li> </ul>	<ul style="list-style-type: none"> <li>Extendibility through the probability distribution packages.</li> <li>Provide the limited level of GUI interface.</li> </ul>	Open Source	Limited	Java

**REFERENCES:**

- [1] Q. Zhang, L. Cheng, R. Boutaba, “Cloud computing: state of the art and research challenges”, Journal of Internet Services and Applications, vol.1, no.1, 2010, pp.7-18.
- [2] Dillon, T., Wu, C., & Chang, E., “Cloud computing: Issues and challenges”, 24th IEEE International Conference in Advanced Information Networking and Applications (AINA), , 2010, pp. 27-33.
- [3] R. Buyya, R. Ranjan, R.N. Calheiros, “Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities”, Proceedings of the 7th High Performance Computing and Simulation Conference (HPCS 2009), Leipzig, Germany, IEEE Press, New York ,2009, pp. 21-24..
- [4] R. N. Calheiros, R Ranjan, A Beloglazov1, C A. F. De Rose, R. Buyya, “CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms”, Published online 24 August 2010 in Wiley Online Library.
- [5] B. Wickremasinghe, R. N. Calheiros, R. Buyya, “CloudAnalyst: A CloudSim-based Visual Modeller for analysing Cloud Computing Environments and Applications”, 24th IEEE International Conference on Advanced Information Networking and Applications, 2010.
- [6] A. Núñez, J. L. Vázquez-Poletti, A. C. Caminero, G. G. Castañé, J. Carretero, I. M. Llorente “iCanCloud: A Flexible and Scalable Cloud Infrastructure Simulator”, Journal of Grid Computing, vol. 10, Issue 1, 2012 pp. 185-209.
- [7] Arif Ahmed, Abadhan Saumya Sabyasachi, “Cloud Computing Simulators: A Detailed Survey and Future Direction”, 2014 IEEE International Advance Computing Conference (IACC), 2014, pp 867-872.
- [8] Adil Maarouf, Abderrahim Marzouk, Abdelkrim Haqiq, “Comparative Study of Simulators for Cloud Computing”, International Conference on Cloud Technologies and Applications (CloudTech), 2015, pp: 1-8.
- [9] P. S. Rawat, G. P. Saroha, Y. Barthwal, “Performance Evaluation of Social Networking Application with different Load balancing policy across Virtual Machine in a Single Data Center using CloudAnalyst”, in 2nd IEEE International Conference on Parallel, Distributed and Grid Computing, 2012
- [10] D. Kliazovich, P. Bouvry, S. U. Khan, “GreenCloud: A Packet level Simulator of Energy-aware Cloud Computing Data Centers,” Journal of Supercomputing, vol. 62, no. 3, pp. 1263-1283, 2012.
- [11] S. Lim, B. Sharma, G .Nam, E. K. Kim, and C. R. Das “MDCSim: A Multi-tier Data Center Simulation Platform” Cluster Computing and Workshops, 2009.
- [12] S. Ostermann, K. Plankenstein, R. Prodan, Th. Fahringer, “GroudSim: An Event-Based Simulation Framework for Computational Grids and Clouds”, Euro-Par 2010 Parallel Processing Workshops Lecture Notes in Computer Science Volume 6586, 2011, pp 305-313.

# A Comprehensive Review of Various Issues And Challenges In Wireless Sensor Networks

Anupama Khatak,Raman Maini

MTech Student,Department Of Computer Engineering,Punjabi University,Patiala,Punjab,India

Professor,Department Of Computer Engineering,Punjabi University, Patiala,Punjab,India

**Abstract-**Over the past decades Wireless Sensor Networks WSN have shown a great change due to its advancement in technology and adaptability nature. It is that breed of organization that can be found in number of small devices namely sensors nodes. These are usually arranged in remote and distant areas. Every sensor node consists of radio interface to communicate with each other, also these nodes are limited because of the batteries embedded in them. When various sensors mutually monitor large sensible environments, they present a wireless sensor network (WSN). They are adaptable to various environments so they maintain vast range of functions. In this work various issues, challenges, and characteristics related to wireless sensor network are discussed. Since wireless devices should be bandwidth limited and small, a few main challenges in wireless networks are Signal fading, mobility, data rate, power and energy, security and Quality of service (QoS). Also, it has been observed that Wireless sensor networks have motivated various operations in the plots like health, environment, human activity monitoring etc.

Index terms- WSN, sensors, networks, nodes, issues, challenges.

## I. INTRODUCTION

Network is described as a bunch of two or more computer systems linked to one another to exchange the information wherein, wireless network is called as the sort of computer network that make use of wireless data interconnections for linking network nodes. A object whose goal is to analysis the changes in its environment, and then produce a related output is called as sensors. Summing up of all the above terminologies forms wireless sensor network. They are the partially scattered self-determining sensors to invigilate the physical or environmental circumstances, most likely temperature, sound, pressure, and so on. and so collectively handover their information throughout the network to the head destination. Such achievement of wireless sensor networks was inspired from army applications such as battleground control; recent days wireless sensor networks are consumed in the fields like industrial and user applications, such as industrial process monitoring and control, machine health check. This paper summarize the basics related to the wireless sensor networks, issues, challenges and its advantage to the environment or to the humans.

In the section 2 wireless sensor networks are explained along with standards and characteristics.

Section 3 comprises issues being faced. Further sections includes beneficiaries provided to the environment and challenges it accepts.

## II. WIRELESS SENSOR NETWORK

To built a WSN, enormous count of sensors nodes are required in which single node is linked with one or sometimes a lot of sensors. Number of sector in individual sensor network node are: radio transmitter with an internal antenna or connection to an external antenna, a microcontroller, an electronic circuit for intermixing with sensors and an power

sources. The configuration of wireless sensor networks varies from network to network. The movements can be in the form of routing or overflowing and it happens between the hops of the network. A wireless sensor network has three segments: Sensors Nodes, Sink Node (User) and Target Nod, as shown in Figure below.

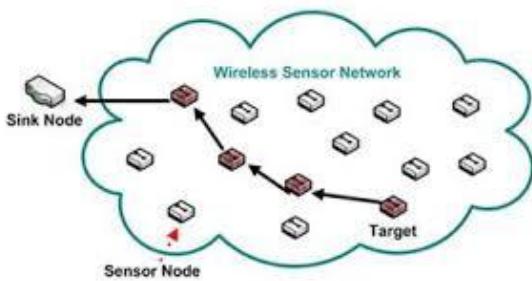


Fig: 1 Wireless Sensor Network [II]

The data is collected in sensor node which contains number of sensors and actuators, that also control various physical processes. Internet segments present in the network are used to transfer collected data to the user. Another function of a node is to operate computation of the systematic data. Individual nodes need face to face interaction among them. The tasks are performed by the node of task manager in data storage, search and exhibit.[ii]

### 2.1 Standards of WSN

Effective standards mainly consumed in WSN communications favors:[iv]

- WirelessHART
- IEEE 1451
- ZigBee / 802.15.4
- ZigBee IP
- 6LoWPAN
- LEECH
- TinyOS

## III. ISSUES

Various issues faced by wireless sensor networks are as follows:

### 3.1. Data Integrity

To solve the problem of data integrity attack, data integrity is must. It is gained by authenticating the data content. Message digest and MAC are the techniques required to maintain the level of data integrity. Wireless sensor networks require data integrity to make sure that the reliability of the information that is packet information received should be same as that of send by the sender. And if in any case any one in between cannot alter that data packet.[vi]

### 3.2. Data Confidentiality

Data Confidentiality is the most significant issue in the field of security of network. Confidentiality is needed in wireless sensor network to protect the data transfer between receiver and sender so that third party do not read, write or update the information. Data Confidentiality is maintained by the term known as Cryptography. Symmetric and Asymmetric key are the two techniques of cryptography used to gain the data confidentiality. [vi]

### 3.3. Data Availability

Availability is the primitive service for the functionality of the network. Many phenomena has been proposed to achieve this aim. Availability of the network resources is maintained even under the pressure of attacks such as the denial of

services attacks (DOS). To endure the functionality of the network, sensor nodes should remain active and this is enhanced by the availability of the data. Data Availability analysis as if sensor node has the capability to use the services and manages the path to communicate the message.[vi]

### *3.4. Data Authentication*

Sensor node require data authentication to make sure that the receivers data is not altered while the transmission proceed. It is gained via symmetric or asymmetric components where sending and receiving nodes share private keys. Digital Signatures are used in asymmetric cryptographic communication wherein in symmetric key, MAC (Message Authentication Code) are used.[vi]

### *3.5. Data Freshness*

Data freshness is the most significant issue in wireless sensor networks among all others. Since human do mistakes and so an attacker can send an out-dated packet of data to ruin the network belongings and decrease the life time of the network. Freshness ensures that the information received by the receiver is not expired along with lifetime energy is maintained.[vi]

## IV. CHALLENGES

Research challenges in wireless sensor networks are as follows:

### *4.1 Signal Fading :*

The singles that travel via wireless network are not reliable as they may carry a lot of noise along with them. This is so because there is no protective shield as in wired network and the wireless networks are highly vulnerable to attacks or malfunctions. The receiver may receive the signal from some unpredicted and unexpected path due to reflection and diffraction. These reasons makes the received signal due to which at times the receive is unable to recognize the genuine signal and lot of packets are lost in this confusion.[vi]

### *4.2 Mobility :*

The devices with wireless connections are free to move without restriction of boundaries and are thus mobile in nature. To continue with mobility the signals or the connections that connect the devices must always be kept in a live status. The well built structures allow the devices to leave one base stations and join another letting handoff to take place. The process of handoff follows proper procedures, rules and regulations. The network must be such that the routing protocol gives maximum efficiency.[vi]

### *4.3 Power and Energy :*

The devices that are mobile are portable and easy to carry everywhere and capable of fulfill the needs of the users though these devices have power issues as compared to the stationary devices. This is so because the power sources can not be carried along as mobile sources. Therefore the mobile devices must be power efficient in order to make the use of devices very efficient. The receiving and transmitting capabilities must be very strong for smooth functioning.[vi]

### *4.4 Data Rate :*

Data rate is essential factor to determine the speed efficiency of multimedia applications. Data rate is dependent on many factors such as data compression efficiency, power considerations, protocols that govern data transfer. Thus the manufactures should keep these factors in mind while considering to get higher rate of data transfers. Data compression plays an important part in multimedia applications when used in wireless connectivity. Current compression methods offers 75-100 compression ratio. The challenge arises to improve these data compression algorithms so that it is possible to produce audio and video of very high quality at present confining rates.

### *4.5 Security :*

The main concern in commercial applications such as e-commerce deals with the security of the mobile users in wireless environments. Wired equivalence privacy (wep) governs as the security standard of wireless medium as defined by IEEE801.11. it provides a means of data encryption while it is transferred between access point and computers. VPNs can be used as well to ensure the safety of networks. To ensure safety one must continuously monitor and regularly update the measures.[vi]

#### 4.6 (QoS) *Quality of Service :*

Quality of Service analysis the performance of the network that symbolizes the quality transmission of the network and availability of services. Single traffic flow of the network for which QoS is categorized by quadruple[vi] specifications:

- o Reliability
- o Delay
- o Jitter
- o Bandwidth

## V. APPLICATIONS

motivated applications of wireless sensor networks as following:

5.1. *Area Monitoring:* in monitoring of the area wireless sensor networks are spread out as in case of military to find the enemies encroachment.[ii]

5.2. *Health Care Monitoring:* this medical monitoring is explained in two individual devices: wearable and implanted. Monitoring of the human body from outside is done by wearable devices and monitoring health from inside is done by implanted devices[ii].

5.3. *Environmental Monitoring:* there are number of applications to monitor the environment parameters[ii] naming:

5.3.1. *Air Pollution:* various cities now days have been deployed with wireless sensor networks to monitor the harmful gases for the citizens.

5.3.2. *Forest Fire Detection:* wireless sensor nodes let know the fire brigade when fire is started and how it spreads. They sense the temperature, dangerous gases, humidity produced by the fire in nature.

5.3.3. *Landslide Detection:* before landslide takes place it can be detected with help of wireless sensor networks and destruction can be avoided. Movement of the soil before and during landslide can be detected by the landslide detectors.

5.3.4. *Water Quality Monitoring:* quality of water rivers, dams, underground water, oceans contains are all measured by wireless sensor network.

5.4. *Industrial Monitoring:* with rapid growth of wireless sensor networks monitoring of industries can be done.[ii]

5.4.1 *Machine Health Monitoring:* machines like rotating machinery and untethered vehicles where wired systems is hard to maintain, wireless sensor networks are used which favors less cost high functionality.

5.4.2 *Data logging:* this is as simple as monitoring of temperature of the fridge. Wireless sensor nodes are used to assemble the data for the environment data monitoring.[ii]

## VI. CONCLUSION

This paper discuss basic facts related to the wireless sensor network, its advantage to the environment and society along with the issues, challenges faced to create definition in present world. WSN are designed to cluster information from the environment. They have huge count of sensor nodes and one or more Base Stations. The nodes in the network is joined via Wireless communication medium. Every node posses the ability to notice the information, work over the information and pass it to other nodes or to Base Station. Such networks are finite by the node lifetime period. To face the issues it has been analyzed that the security ,freshness, reliability of data should be maintained. For smooth functioning transmitting capabilities needs to be strong, continuously measure the updates for the data are the research challenges . In the future work one of the clear-cut field can be security or power capabilities, which is always going to promote score itemized result.

## REFERENCES

- [1] Antonio-Javier Garcia-Sanchez, Felipe Garcia-Sanchez, Joan Garcia-Haro, "Wireless sensor network deployment for integrating video-surveillance and data-monitoring in precision agriculture over distributed crops", Computers and Electronics in Agriculture volume:75 (2011) pages: 288–303.
- [2] Soledad Escolar Díaz, Jesús Carretero Pérez, Alejandro Calderón Mateos," A novel methodology for the monitoring of the agricultural production process based on wireless sensor networks", Journal of Computers and Electronics in Agriculture volume: 76 (2011) pages:252-265.
- [3] Bara" a A. Attea, EnanA.Khalil, "A new evolutionary based routing protocol for clustered heterogeneous wireless sensor networks" journal of Applied Soft Computing (2011).
- [4] Raimo Nikkilä , Ilkka Seilonen, Kari Koskinen, "Software architecture for farm management information systems in precision agriculture" ,journal of Computers and Electronics in Agriculture volume:70 (2010) pages:328–336.
- [5] A. Matese , S.F. Di Gennaro, A. Zaldei, L. Genesio, F.P. Vaccari," A wireless sensor network for precision viticulture: The NAV system" Computers and Electronics in Agriculture volume:69 (2009) pages:51–58.
- [6] Jenna Burrell, Tim Brooke, and Richard Beckwith Intel Research, "Vineyard Computing: Sensor Networks in Agricultural Production", PERVASIVE computing Published by the IEEE CS and IEEE ComSoc 1536-1268/04© 2004 IEEE.
- [7] Zheng, Yugui Qu, Baohua Zhao, "Data Aware Clustering for Data Gathering in Wireless Sensor Networks", International Conference on Networks Security, Wireless Communications and Trusted Computing, 2009, Volume.1, Page(s). 192-214.
- [8] N. Ramanathan, M. Yarvis, "A Stream-oriented Power Management Protocol for Low Duty Cycle Sensor Network Applications," in Proc. IEEE EmNetS-II Workshop, May 2005
- [9] A. Boulis, M.B. Srivastava, "Node-level Energy Management for Sensor Networks in the Presence of Multiple Applications," in Proc. IEEE PerCom Conf., Mar. 2003.
- [10] C-Y. Wan, S. B. Eisenman, A. T. Campbell, J. Crowcroft, "Siphon: Overload Traffic Management using Multi-radio Virtual Sinks in Sensor Networks," in proc. ACM SenSys Conf., Nov. 2005.
- [11] J. Zhang, E.C. Kulasekere, K. Premaratne, P.H.Bauer, "Resource Management of Task Oriented DistributedSensor Networks," in Proc. IEEE ICASSP Conf., May 2001.
- [12] M. Perillo, W.B. Heinzelman, "Providing Application QoS through Intelligent Sensor Management," in Proc.IEEE SNPA Conf., May 2003.

## WEB REFERENCES

- I.<http://www.ni.com/white-paper/7142/en/>
- II.[https://en.wikipedia.org/wiki/Wireless\\_sensor\\_network](https://en.wikipedia.org/wiki/Wireless_sensor_network)
- III.<http://searchdatacenter.techtarget.com/definition/sensor-network>
- IV.[http://www.ijarcsse.com/docs/papers/Volume\\_4/2\\_February2014/V4I2-0175.pdf](http://www.ijarcsse.com/docs/papers/Volume_4/2_February2014/V4I2-0175.pdf)
- V.[www.scirp.org/journal/wsn/](http://www.scirp.org/journal/wsn/)
- VI.[www.sciencedirect.com/science/article/pii/S1389128608001254](http://www.sciencedirect.com/science/article/pii/S1389128608001254)
- VII.[https://www.researchgate.net.../283205038\\_Issues\\_and\\_Challenges\\_in\\_Wireless\\_Sensor](https://www.researchgate.net.../283205038_Issues_and_Challenges_in_Wireless_Sensor)

# Data Hiding In 3D Barcode Image Using Steganography

Authors

Rama Rani<sup>1</sup>, Er. Gaurav Deep<sup>2</sup>

1. Department of computer Engineering, Punjabi University ,Patiala, Punjab, India

2. Assistant Professor , Department of Computer Engineering , Punjabi University Patiala, Punjab, India

Email: Rama.gargcse@gmail.com , deepgaurav48@pbi.ac.in

**Abstract-** Steganography is a method hiding information in an information carrier in such a way that only the sender and the intended receiver knows the confidential information exists. In image steganography , image is used as an information carrier to hide the data. Today Barcodes system is very popular for protecting sensitive information. This paper introduces the concept of hiding data in 3D Barcode image using Discrete Wavelet Transformation. 3D Barcodes do not use any labels. They are embossed or engraved directly on the product during the manufacturing process. 3D barcode use the same basic principal as 1D or 2D barcode use. The performance evaluation is done by using statistical parameters.

**Keywords:** Steganography , barcodes, data hiding , PSNR , MSE , cover image , stego image , Discrete wavelet transformation.

## 1 . INTRODUCTION TO STEGANOGRAPHY

Steganography is defined as the art and science of making communication invisible. It is a technique of hiding information in a cover medium. Steganography is of Greek origin. [1] .It is made up of two words, Stegano which means secret and graphy which means writing. In steganography we can hide the information in various cover mediums such as data files, images, audio files, video files [2]. Steganography means embedding a secret message or encrypted message within a large source cover in such a manner that the third person does not come to know about that some sensitive information is hidden [3]. Redundant or useless bits of the carrier are used to hide the secret information. This hidden information can be plain text or an image. It is mainly used in situation where our main focus is on the confidentiality of information. [4]

### *1.1 General Specifications Of Steganography*

In today's modern era there are a large number of methods are available that can be used to implement steganography but the basic terminology that is used in each implementation is as follows [4]:

*Cover Media:* It is defined as the media in which secret message is to be embedded. It can be an image file, audio file, or video file.

*Hidden Data:* It is defined as the data that is to be embedded or the data that is to be sent in such a way that it must not be known to anyone other than intended receiver.

*Stego-Key:* This key is used to provide security by implementing cryptography. It is useful in such a way that even if hidden is found, than it will not be in a human readable form but in a scrambled form.

*Stego-Media:* This is the same as the cover media but it contains hidden information. This media is sent over the communication channel to the intended receiver.

### 1.2 Main Aims Of Steganography

1. To hide the existence of information from an unauthorized person by embedding the information in some other medium.
2. Secure Data transmission over networks.

### 1.3 Classification Of Steganography

Steganography techniques are classified on the basis of the type of the domain: whether it is spatial domain or transformational domain, type of cover medium used and the type of the embedding method used. Classification on the basis of cover medium [5]:

1. Text Steganography.
2. Image Steganography.
3. Audio Steganography.
4. Video Steganography.

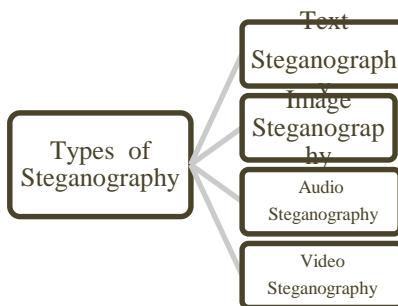


Fig.1 Types of steganography

### *Image Steganography*

It is the most widely used technique for secret communication. Images acts as main source of hiding secret data. An image is usually defined as a collection of number of different light intensities. In image steganography pixel intensities are used to hide data. Image steganography is classified into two categories: Spatial Domain Steganography and Transformational Domain Steganography [4].

#### *1. Spatial Domain Method:*

In spatial domain method the secret messages are embedded directly in the intensity of pixels. In this the most widely used method is least significant bits (LSB) insertion method. In this technique , the least significant bits are replaced by the secret message bits [6].

#### *2. Transform Domain Method:*

The Transform domain method is used for hiding a large amount of information, provides high security, good invisibility and no loss of secret message. In this method the bits of the image are transformed into some coefficients and then data is embedded into these coefficients. This method can be used by applying various transformations such as Discrete Cosine Transformation (DCT), Discrete Fourier Transformation (DFT), Discrete Wavelet Transformation (DWT) [6].

## 2. INTRODUCTION TO BARCODES

A Barcode is an optical machine readable representation of data relating to an object to which it is attached [7]. It is simply a series of lines (usually black) on a light background (usually white). It simply represents data by using spaces and widths between parallel lines. We use special type of scanners called Barcodes scanners to decode barcodes . Barcodes can be represented in various forms like hexagons ,dots, rectangles and other graphical patterns.

### *2.1 Types Of Barcode*

There are three types of barcodes: one dimensional (1D) barcode, two dimensional (2D) barcode , three dimensional (3D) barcodes.

*One Dimensional Barcodes:* In one dimensional barcodes we can store information only in horizontal direction. we also call 1D barcodes as linear barcodes. In these barcode numbers and letters are encoded in parallel lines. These type of barcodes are used everywhere such as in product transportation, business , medical industries. e. g universal product code, code 128.



Fig.2 One Dimensional Barcode

*Advantages of 1D Barcodes:*

1. Data is captured very fastly.
2. It's printing cost is very less.
- 3.These generate very less errors, so are more reliable.

*Disadvantages of 1D barcodes:*

- 1.The main disadvantage of 1D barcode is that it stores less data .

*Two Dimensional barcode:* In comparison to 1D barcodes 2D barcodes store quiet large information. We can store information in both vertical as well as horizontal direction. E.g. 2D Stacked barcodes, 2D Data Matrix barcodes.



Fig.3 Two Dimensional Barcode

*Advantages of 2D Barcodes:*

1. We can read and write information easily .
2. These type of barcodes are more durable
- 3.More secure than 1D barcodes.

*Disadvantages of 2D barcodes:*

- 1.we need special type of scanners known as barcode readers to decode information from 2D barcodes

*Three Dimensional Barcode:* These barcodes do not make use of any symbols or labels. During the production or manufacturing process, we directly emboss them on the product. We use the same principal in 3D barcodes as we use in 1D or 2D barcodes.



Fig. 4 Three Dimensional Barcode

*Advantages of 3D Barcodes:* These barcodes are read based on the height of the barcode not on the ratio of black to white lines.

### 3 . IMPLEMENTATION

There are a number of methods or techniques available that can be used for hiding data in 3D Barcode images.

3D Barcodes are usually represented by matrix. In the matrix we can store large amount of information .we are thus able to hide a large amount of information in barcodes.

*At Sender Side:*

Step 1: Overflow of data or information is the major issue in the data hiding process. This overflow occurs when the gray scale intensity value of the secret message pixel exceeds the lower bound(0) or upper bound(255). This overflow is controlled by filtering the cover image. There are a number of filters like high pass filters and slow pass filters are used to filter the cover image.

Step 2: The filtered image is then transformed into the colored image using HSI( hue, saturation , intensity) model.

Step 3.In this step clustering is applied on the colored image. Clustering is a technique used to group the objects with similar properties.

Step 4: Next, we have to recognize the largest cluster in which we can embed the secret message in an efficient way. The cluster can be selected on the basis of the size or color.

Step 5: In this step the secret information is converted into encoded text by implementing cryptography. The secret information can be converted into the encoded form either by using public key or private(secret )key.

Step 6: Discrete Wavelet Transformation is applied on the identified cluster. This transformation divides the cluster into four frequency sub- bands- LL, LH, HL, HH sub bands. Then the Most significant bits of high frequency sub-band are selected.

Step 6: The encoded secret message (S') is then hidden into the Coefficients of MSB's at high frequency sub bands of DWT.

3D Barcode image as cover image

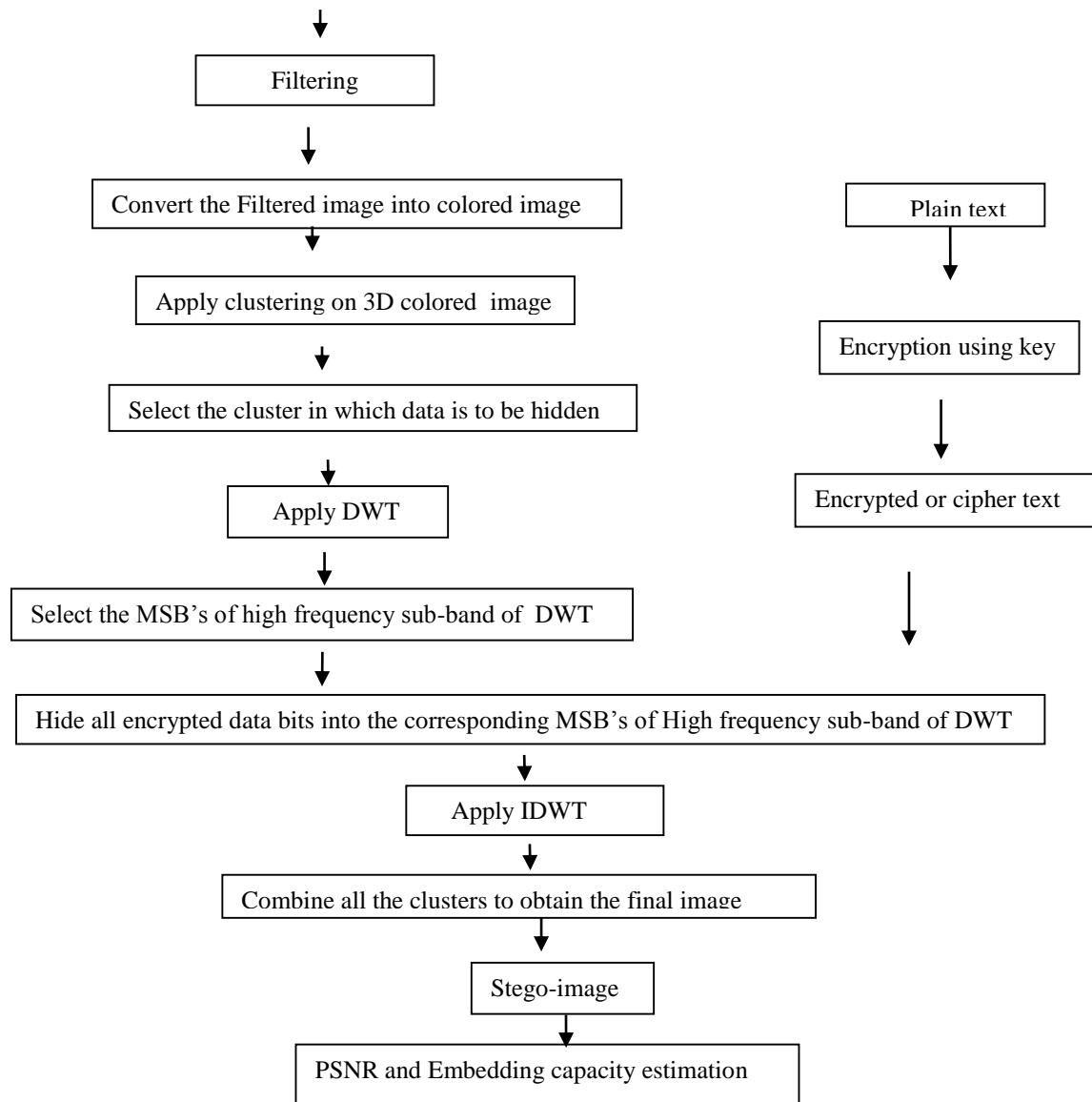


Fig.5. Embedding Phase

Step 7: Finally we apply the inverse Discrete Wavelet Transformation And combine all the clusters to obtain the resulting stego images  $ST(x, y)$ .

Step 8.Two important performance measures i.e PSNR and MSE are estimated.

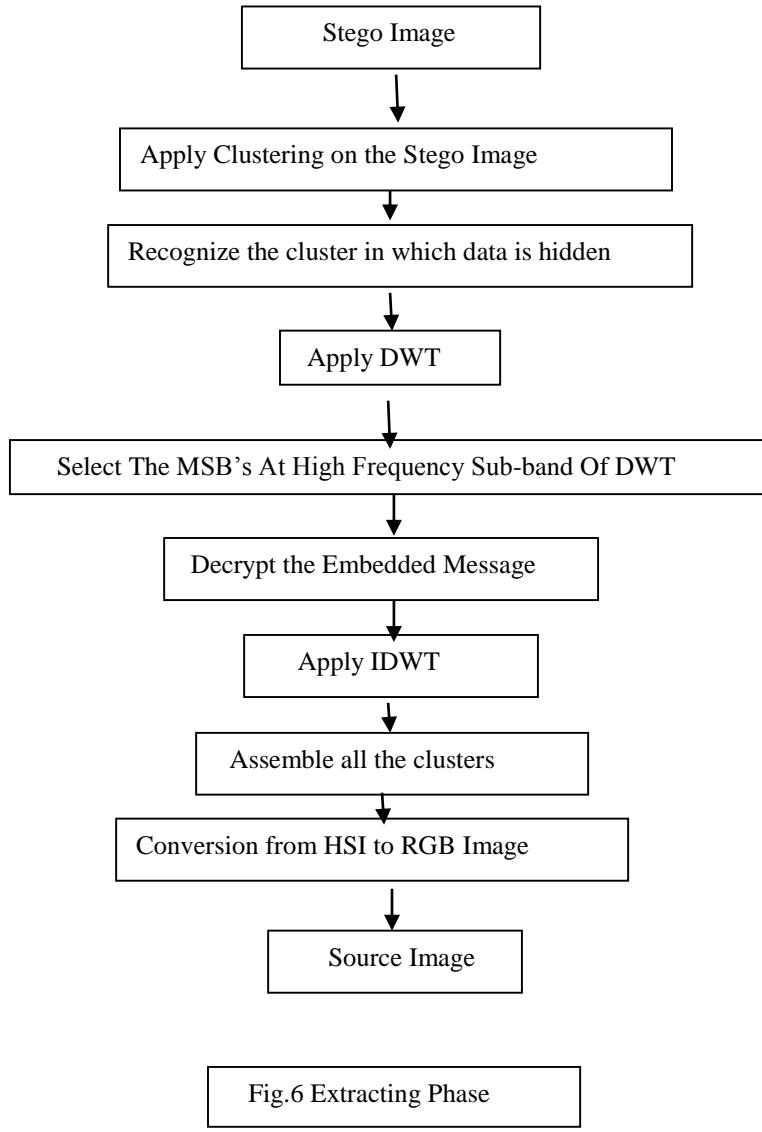
$$MSE = \frac{1}{m*n} \sum_{x=1}^m \sum_{y=1}^n [S(x, y) - ST(x, y)]^2, PSNR = 10 \log_{10} \left( \frac{255^2}{MSE} \right)$$

*At Receiver Side:*

At receiver side , following steps are used to extract the source or cover image and the embedded secret message from the source image.

Step 1. First of all the Stego image is taken as an input. The stego image is the image generated after embedding secret message in the cover image.

Step 2: In this step , a number of clusters of the stego image are created by applying clustering technique like using K-Means clustering.



Step 3: Then we have to recognize the largest cluster in which the sensitive information or secret message is hidden.

Step4: After recognizing the cluster in which the data is hidden , we have to apply discrete wavelet transformation on the selected cluster in order to identify the Most significant bits at high frequency sub band of dwt.

Step 5: In this step , by using the same key as used in the embedding phase we can decrypt the secret message.

Step 6: Then inverse discrete wavelet transformation is applied and all the clusters are assembled.

Step 7: Finally the HSI image is transformed back into the original image or cover image.

#### 4. CONCLUSION

In this paper we described a general methodology for embedding information into image like 3D Barcodes. In This technique first of all a 3D barcode image is taken as input. As we all know in the data hiding process one of the major issue is the overflow of the data, this overflow is controlled by filtering the cover image using filters. Filtered image is then converted into HSI image. Clusters are created based on the color pattern matching . The cluster with the largest number of pixels is selected to hide the data and the encoded message is hidden into the least significant bits of high frequency sub-band of Discrete Wavelet Transformation. The Stego image is generated when embedding is completed. This technique proved to be more secure and effective as it provides more choices of clusters to hide the data and it can not be easy to locate the cluster in which the secret message is hidden. The performance analysis of this algorithm i [6] [8] [9] [2]s done by using two parameters PSNR(Peak Signal To Noise Ratio) and MSE(Mean Square Error).

#### 5..REFERENCES

- [1] G. D. Yogita Puri, "Image Seeded Steganography: Steganography Using Seed Values," *International Journal Of Scientific Research And Education*, no. 3, July-2015.
- [2] D. A. S. K. M. Indra Sena Reddy, "Secured Data Transmission Using Wavelet Based," in *International Conference on Computational Modeling and Security (CMS 2016)*, 2016.
- [3] G. O. B. M Ramesg, "QR- DWT Code Image Steganography," *International Journal of Computational Intelligence and Informatics*, vol. 3, April-June 2013.
- [4] G. Harpreet kaur, "Level's 4 Security in Image Steganography," *International Journal of Computer Applications (0975 – 8887)*, july-2015.
- [5] S. B. ., R. sumeet kaur, "Steganography and Classification of Image," *IEEE*, 2014.
- [6] V. K. Snehal O.Mundhada, "Spatial and Transformation Domain Techniques for Image Enhancement," *International Journal of Engineering Science and Innovative Technology (IJESIT)*, vol. 1, no. 2, NOVEMBER2012.

- [7] I. Š. ,. R. K. ,. S. H. ,. G. O. ,. S. S. Laslo Tarjan, "Automatic identification based on 2D barcode," *International Journal of Industrial Engineering and Management (IJIEM)*, vol. 4, 2011.
- [8] W.-Y. C.-M. T. Chin-Ho Chung, "Image Hidden Technique Using QR-Barcode," in *Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2009.
- [9] S. b. S. M. Z. Alaa A. Jabbar Altaay, "An Introduction to Image Steganography Techniques," in *International Conference on Advanced Computer Science Applications and Technologies*, 2012.
- [10] D. H. S.Uma Maheshwari, "Frequency domain QR code based image steganography using Fresnelet transform," *International Journal of Electronics and Communications (AEÜ)*, p. 6, November 2014.
- [11] S. K. Sahu, "Encryption in QR Code Using Stegnography," *International Journal of Engineering Research and Applications*, vol. 3, no. 4, p. 4, June-Aug 2013.
- [12] A. K. S. Sharu Goel, "A Secure and Optimal QR code," *International Journal of Engineering Research & Management Technology*, September-2014.
- [13] K. Akshara Gailward, "Information Hiding Using Image Embedding in QR Codes for Color Images.," *International Journal of Computer Science and Technologies.*, 2015.
- [14] V. R. a. G. Gopinath, "Hiding Secret Text in Quick Response Code and Transforming the Position of the Secret Data using Rotation Transformation," *Research Journal of Applied Sciences, Engineering and Technology*, 2014.
- [15] G. K. KIMMY GHANAIYA, "new approach Data hiding in 2D data matrix and tilt correction algorithm," *International Journal Of Computational Engineering Research*, vol. 2, May-June 2012.
- [16] E. D. Gagandeep Kaur, "HSI Color Space Conversion Steganography using Elliptic Curve," *International Journal of Innovations & Advancement in Computer Science*, vol. 4, no. 6, JUNE 2015.

# A Survey on Evolution of Cyber-Crime

Amrinder Singh<sup>1</sup>, Amardeep Singh<sup>2</sup>

*Department of Computer Engineering, Punjabi University, Patiala, India<sup>1</sup>*

[Amrindersingh19792@gmail.com<sup>1</sup>](mailto:Amrindersingh19792@gmail.com)

*Professor of Computer Engineering, Punjabi University Patiala, India<sup>2</sup>*

[Amardeep\\_dhiman@yahoo.com<sup>2</sup>](mailto:Amardeep_dhiman@yahoo.com)

**Abstract:-** the survey was done on the number of cyber-crimes committed between its inception with the invention of blue box to making hacking oriented operating systems which help the budding hackers learn penetration into various software technologies. At first the cyber-crimes were committed to spread various virus, worms just to crack down the system. Today the cyber criminals seek to earn major money, highly sensitive data, information, top secret files, etc. The paper includes evolution of cyber-crimes since 1970; the study tries to encapsulate major cyber-crime events which took place during the journey.

**Keywords:** - cyber-crime, DoS, worms, VIRUS, Hackers, Identity thefts.

## I. INTRODUCTION

A few years ago, a threat in the world was considered only on its physical and real world disaster created by a single or group of people. Big threats included bank robberies, burglar, gangsters, underworld, crooks, etc. Starting with 1970, a new threat started to evolve in the society which was meant to break into a computer system named as ‘hacking’ [1]. Since then it has evolved to become one of the biggest and root requirement to accomplish any other attack. Today, it is more often called as cybercrime. A number of criminals are cyber criminals. Cybercrime includes any kind of prohibited activity accomplished on a computer system, server or database using a computer system and internet to perform any specific kind of activity such as downloading a file, spreading virus, worms, phishing schemes, botnets to stealing millions of dollars, etc. The cyber criminals may try to steal the money, useful information, top secret dataset, etc. depending on the reason of their attack. The most dangerous thing about cybercrime is that the criminals might be sitting in totally different geographical location while the crime is being done at other geographical location [2]. One of the major kind of cybercrime is identity theft, which is done using fake websites where users are asked to write username, password, address, contact number, email address, bank account number, debit/credit card details, pin number of debit/credit card, etc. details which make up the identity of a person and can be used to slip money from their bank accounts. Cyber-crime has evolved very drastically stealing from hundreds to thousands and now to thousands of millions annually.

## II. EVOLUTION OF CYBER CRIME

**Years 1970-1979**:- One of the biggest electronic crimes in this phase was the creation of blue box by Steve Wozniak and Steve Jobs (Apple founders) which was used to make free calls using a 2600 Hz whistle same as which was used by operator's console. It interrupted the ongoing calls routed the intruders calls for free calling, controlling the switching in long dialing systems [3]. The difficulty faced at that time was that it was impossible to trace calls because of the inability of the technology at that time, Because of this reason blue boxes were most popular among the drug dealers.

**Years 1980-1989**:- This era started a whole new type of computer attacks which included changing the billing clock (enabling discount during normal hours), Creation of first computer virus by a Pakistani intelligent programmer, Apple II boot virus, Creation of CERT (computer emergency response team), First bank computer theft worth 70 million dollars (national bank of Chicago), a self-replicating computer program called morris worm was created by Robert T. Morris, Jr., graduate student at Cornell University which spread over 6000 networked computers[4][5].

**Years 1990-2000**:- The start of 1990 saw the formation of EFF (Electronic Frontier Foundation), there after starting first ever online warfare- jamming phone lines, monitoring calls, trespassing in each other's private computers. In 1992, first polymorphic virus which affected data types and functions was released by a group named by dark avengers. This decade saw new kind of macro attacks from online piracy, identity theft attacks, spam, cyber stacking, cyber-terrorism, Denial of Service (DoS) attacks and botnet attacks. The end of this decade saw the release of Melissa worm which is the most costly malware outbreak till date. Few major extortions also took place towards the end which asked to pay around 100K or the hackers will release the credit card details of all the customers[4][6][15].

**Year 2001**:- This year saw a big increase in the cyber-crimes relative to the previous decades. In the starting month of the year new DoS (Denial of service) attack named 'Servers' attacked Microsoft and led to shut down the access of Microsoft website for around 2 days, Anna Kournikova virus was one of the famous virus asking to click on a link offering sexy pictures of Anna kournikova followed by first polymorphic worm named by code red infecting thousands of machines, EU adopts a new cybercrime treaty which became very controversial was adopted by EU, which prohibits the use of any kind of hacking tools[7][8][9].

**Years 2002-2004**:- This era was the beginning of big cyber-crimes. Things before this era had not led to loss of big money. A sys-admin named by Roger Duronio created a logic bomb which costs more than \$3M in repairs and losses. Klez.H worm becomes the biggest malware outbreak in terms of machines infected. The start of 2003 saw

the fastest spreading worm in history named by SQL slammer. The big offering of \$250K by Microsoft for telling about the makers of MSBlastworm and Sobig virus makes it clear about the seriousness of these viruses. Microsoft saw series of worm attacks named by MyDoom worm, Netsky, Sasser, Bagel, Sober worms [10][11][15].

**Year 2005:-** This was the year of account hacks, with the news of the FBI's email system being hacked was spread in 2005 suddenly everything looked unsecure and vulnerable followed by Paris Hilton T-mobie hack, Choicenpoint145K accounts information hack, 1.2M account information hack of Bank of America, £220M theft by **hacking** Sumitomo Mitsui Bank at London. Trojan horse software is one of the major outburst virus which spreads very fast and changes the format of data files to something unreadable making them just unusable [12][13].

**Years 2006-2008:-** This era saw a time constraint on attacks known by Crime-Dot-9-to-5 which means peak time of attacks was noted between Mondays to Friday between 9am to 5pm. One of the major events of 2006 was centered on NASA when they were forced to block emails with attachments because of the fear of being hacked. In 2007, the email account of US secretary of Defense's was hacked then Estonia is bombarded with massive DoS attack followed by Sept's attack on Bank of India. The databases of both Republican and Democratic presidential campaigns were hacked in 2008 followed by Korean e-commerce site hack, exposing Facebook's private pictures by some URL manipulations, DoS attack in Radio Free Europe[14][15].

**Year 2009-2012:-** Year 2009 begun with intruders attacking around 5 million computers in Israel's internet infrastructure. The end of 2009 saw hacking of twitter page showing an Iranian message. In 2010, the same Iranian message was then shown after a Chinese search engine was hacked. Stuxnet worm was discovered in Iran, Indonesia targeting Iranian nuclear facility. Year 2011 saw a major attack on Bank of America hacking around 85k bank card details. A hacker named TiGER-M@TE made a record in hacking by hacking 70k websites with a single click. In 2012, an open SQLi exploit was found in Facebook. In a gap of few days Marriott and Foxconn were hacked affecting a major shutdown of these websites for few days [15].

**Years 2013-2015:-** Year 2013 started with hacking the twitter account of Burger king and posting McDonald's logo. Following that S. Korean financial institutions as well as the broadcaster YTN were infected and had to shut them down for hours. In year 2014, Bitcoinexchange Mt.Gox was hacked \$460 million followed by news of white house computer being hacked. In year 2015, a cloud service hiked to the peak known as hacking as a service. This year had grossed the most in terms of money theft, spent on securities etc. A survey conducted by Price water house Coopers showed that on an average cost of worst breach costs around £3.14 million [15].

### III. FIGURE OF EVOLUTION OF CYBER CRIME

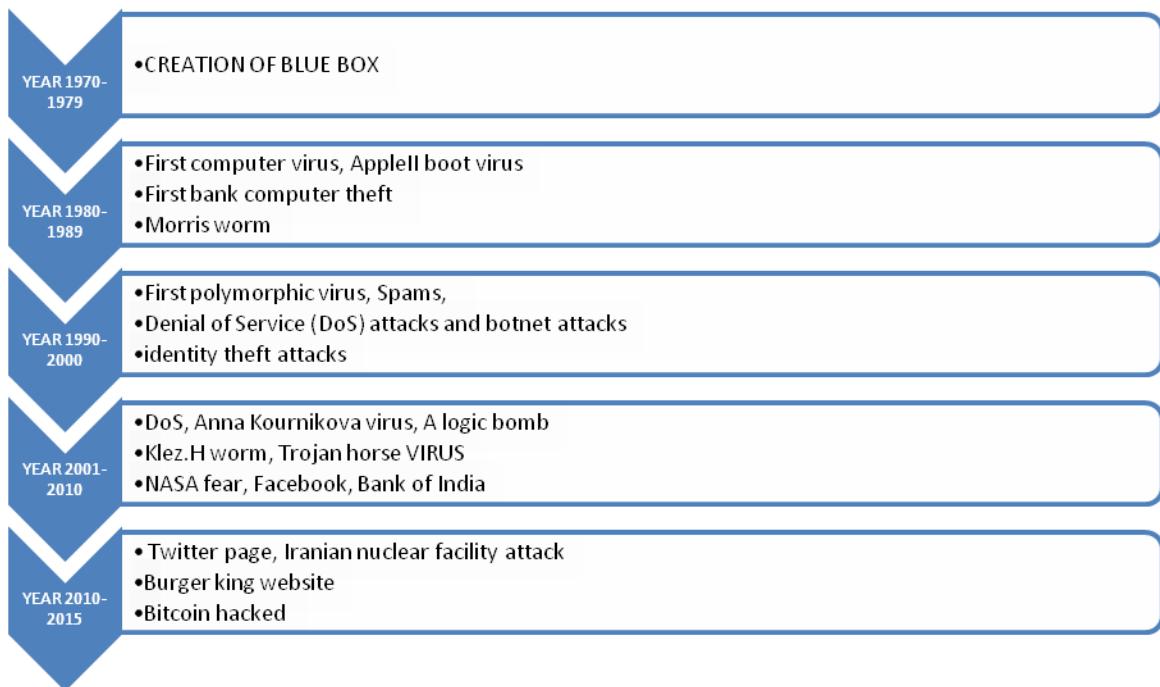


Figure 1: Evolution of cyber crime

### IV. CONCLUSION

The survey concludes that the cyber-crimes have been increasing day by day and year by year. With the increase of technology, the hackers are getting better software as well as hardware. Hacking costs billions of dollars which needs to be checked by introducing better cyber security laws which can be implemented on the current systems so that the rate at which the cyber-crimes are increasing should be stopped and lose of billions of dollars could be stopped and that money might be used for betterment of society. Cyber-crimes should be rated equivalent to the physical crimes and strong laws and punishments should be sentenced to the cyber criminals.

### REFERENCES

- [1] R. Broadhurst, "Developments in the global law enforcement of cyber-crime." *Policing: An International Journal of Police Strategies & Management* 29, no. 3, (2006): 408-433.
- [2] D. Wall, "Cybercrime: The transformation of crime in the information age" polity, Vol. 4, 2007.
- [3] J.T. Draper, "website <http://www.webcrunchers.com/crunch/>"

- [4] D.W. van, Wytske, and P.Wolter. "From Cybercrime to Cyborg Crime: Botnets as Hybrid Criminal Actor-Networks." *British Journal of Criminology* 55, no. 3 (2015): 578-595.
- [5] H. Orman, "The Morris worm: A fifteen-year perspective ", *IEEE Security Privacy*, vol. 1, pp. 35-43, 2003
- [6] Nachenberg, Carey. "Polymorphic virus detection module." U.S. Patent 5,696,822, issued December 9, 1997.
- [7] C. C. Zou, W. Gong, and D. Towsley, "Code red worm propagation modeling and analysis ", *Proc. 9th ACM Conf. Computer and Communication Security*, pp. 138-147, 2002
- [8] Hussain, A., Heidemann, J. and Papadopoulos, C, "A framework for classifying denial of service attacks." In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, pp. 99-110.ACM, 2003.
- [9] Chen, M.Thomas, "Trends in viruses and worms." *The Internet Protocol Journal* 6, no. 3 (2003): 23-33.
- [10] Denning, Dorothy E. "Cyberterrorism: The logic bomb versus the truck bomb." *Global Dialogue* 2, no. 4 (2000): 29.
- [11] Shannon, Colleen, and D. Moore. "The spread of the witty worm" *IEEE Security & Privacy* 2, no. 4 (2004): 46-50.
- [12] M. Devine, "Hacktivism: assessing the damage", *Network Security*, pp.5-13,2011.
- [13] Peluso, Richard, Ashley Haase, Linda Stowring, Mike Edwards, and Peter Ventura, "A Trojan Horse mechanism for the spread of visna virus in monocytes", *Virology* 147, no. 1,pp. 231-236, 1985.
- [14] D.D. Fetterolf and T.D. Clark," Detection of trace explosive evidence by ion mobility spectrometry" ,*Journal of Forensic Science*, 38(1), pp.28-39, 1993.
- [15] Review, NATO. "The History Of Cyber Attacks - A Timeline". *NATO Review*.N.p., 1988. Web. 9 Aug. 2016.

## Image Segmentation Techniques

**Author:** Navdeep Kaur  
Navdeep Kanwal

### I. INTRODUCTION

Segmentation partitions an image into distinct regions containing each pixel with similar attributes. The level to which this partition is carried out depends on the problem being solved, i.e., the segmentation should stop when the objects of interest in an application have been isolated. The applications of computer vision require an image segmentation to extract the meaningful regions of the image. Image segmentation is very useful tool in medical applications. In medical area it is used to extract or region of interest from the background. Image Segmentation simplifies and/or changes the representation of an image into meaningful form and which is easier to analyse.

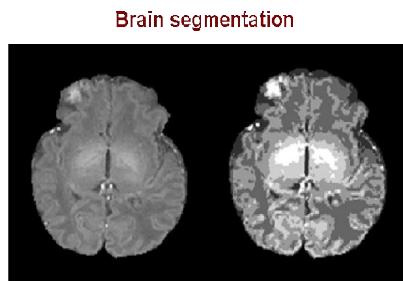


Figure 1. Segmentation

Image Engineering illustrates the level of the image segmentation in image processing. Image Engineering can be divided into three levels. **(a) Image processing** is low-level operations; it operated on the pixel-level. **(b) Image analysis** is the middle-level and it focuses on measuring. **(c) Image understanding** is high-level operation which is further study on the nature of each target and the linkage of each other as well explanation of original image. **Image segmentation** is a key step from the image processing to image analysis.

Image segmentation techniques or methods are classified into two main categories **Layer-Based Segmentation Methods** and **Block-Based Segmentation Methods**.<sup>[1]</sup> See Fig. 2. Layer based methods are used for object detection and image segmentation that composites the output of a bank of object detectors in order to define shape masks and explain the appearance, depth ordering, and that evaluates both class and instance segmentation[1]. The block based methods are based on various features found in the image such as color information that is used to create histograms, the information about the pixels that indicate edge, boundaries or texture information.<sup>[1]</sup> The Block Based Methods are classified into Region based Methods and Edge or Boundary based methods. The region based methods have further various techniques such as clustering, split and merge, normalized cuts, region growing and threshold. Edge or boundary based methods are based on Roberts, Prewitt and Sobel methods. The soft computing approaches are based on Fuzzy logic based, Genetic Algorithm and Neural Networks.

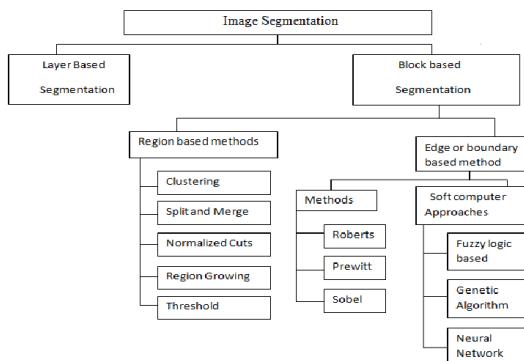


Figure 2. Methods for Image Segmentation

Sasi et all[2] in the study concluded that it provides better acceptable performances in terms of accuracy and Mathew's correlation coefficient for detecting shadows. Vadher et al[3] concluded that Discrimination is much clear in normalized cut method compared to Graph cut method. From the results, Graph cut method is not able to differentiate between the components similar in intensity even if they are spatially well separated. S Kansal et al[4] proposed an algorithm for automatic seed selection for seeded region growing. This Paper shows that this algorithm can be successfully used to achieve automatic image segmentation and this method can be used in object recognition algorithms. Aja-Fernandez et al[5] overcome the common drawbacks of thresholding methods when images are corrupted with artifacts and noise. It is based on relating each pixel in the image to different output centroids via a fuzzy membership function, avoiding any initial hard decision. N. Senthilkumaran et al [6] concluded that soft computing Approaches are applied on a real life example image of nature scene, and the results show the efficiency of image segmentation. O.J. Tobias et al[7] proposed an approach to threshold the histogram according to the similarity between gray levels. Such a similarity is assessed through a fuzzy measure. Paulinas et al[8] proved that genetic algorithms are the most powerful unbiased optimization techniques for sampling a large solution. Al Mohair et al[9] concluded that The proposed method is a hybrid method that combines the advantages of two clustering techniques: neural networks and k-means. In addition, different combinations of color-texture descriptors were investigated to determine the optimal descriptor among the possible combinations. A powerful optimization method, DE, was used in order to determine the optimal combination that can be used for accurate skin detection.

## II. RELATED WORK

Image Segmentation is a process of dividing an image into its constituent parts to extract data from the attributes of the image. As a result, a good segmentation should result in regions in which the image elements should have uniform properties in terms of brightness, color or texture etc. The Segmentation process can be divided into various categories. Each approach has its own advantages and disadvantages.

Nida M. Zaitoun and Musbah J. Aqel [1] proposed that there is no universal segmentation method for all kinds of images and also an image can be segmented by using different segmentation methods. Satish Kumar et al [10], in the survey explained the various applications that uses the concept of the image segmentation which includes computer vision, medical, scanning, recognition etc. M Akhtaruzzaman et al [11] explained the strategy of object segmentation from color image through automated threshold selection procedure. The procedure could be used in characterizing human walking pattern and recognition of gait. H. P. Narkhede [12], in the review of image segmentation study, has described various methodologies and issues regarding to digital image processing used in various recognition patterns. KK Verma et al [13] observed that the different techniques of image segmentation may be used in many advanced mission for identification of regions images or object. Vishal B. Langote [14] explained the image segmentation target which is to split an image into individual regions which are unvarying in computable property such as brightness, color, or texture. Use of adaptive thresholding reduces the difficulty of segmentation. Gupta Mehul et al[15] brings a conclusion that the algorithms that should be used for printed or handwritten text document image differs greatly. The pixel counting algorithm is simple to implement and concluded that it excels only for the printed text document. This algorithm can be used for a handwritten document if it has some kind of guidelines provided or when the document has even text size and uniform interline spacing, but it fails to provide satisfactory results while working with handwritten text images. Also, additional overhead like skew correction module is required. Nameirakpam Dhanachandra et al[16] proposed that an image be segmented using  $k$ -mean clustering algorithm and using subtractive cluster to produce the initial centroid. At the same time partial contrast stretching is used to improve the quality of original image and median filter is used to improve segmented image. After that the final segmented result is compared with  $k$ -means clustering algorithm and concluded that the proposed clustering algorithm has better segmentation. The output images are also tune by varying the hyper sphere cluster radius and concluded from that result that by varying the hype sphere cluster radius we can get different output.

## III. K-MEANS CLUSTERING

$k$ -Means is a popular unsupervised learning algorithm that is used in a wide range of applications, such as data mining, because of its simplicity [17]. K-Means Clustering separates the ' $n$ ' observations into ' $k$ ' clusters so that each observation is assigned to the cluster with the nearest mean. Each cluster has a centroid; and the centroids should be defined and distinct from each other. After defining the  $k$  centroids, the next step is to take each point belonging to a given data set and associate it with the nearest centroid. The groupage is completed when no point is pending. Then,  $k$  new centroids are recalculated as barycenters of the clusters resulting from the previous step. After that, a loop is used to bind the same data set points and the nearest new centroid. During the loop, the  $k$  centroids change their location step by step, and the loop stops when no more changes can be made. Because of its simplicity,  $k$  means clustering works well with large databases. If  $k$  is small,  $k$ -means maybe computationally faster than other techniques such as hierarchical clustering [18].  $K$ -means may produce tighter

clusters than hierarchical clustering [19]. It can be employed in many fields, such as medical image segmentation, brain tumor detection, and content-based image retrieval.

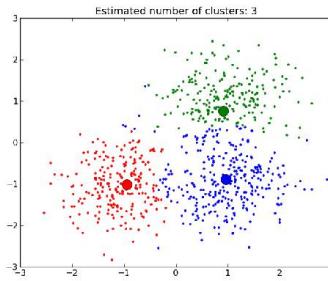


Figure 3. Clustering

Consider there is an image and resolution of image is ' $x \times y$ '. The image has to be cluster into ' $k$ ' clusters. Suppose ' $p(x, y)$ ' be an input pixel to be cluster and ' $c_k$ ' be the cluster centers. The algorithm for  $k$ -means [20] clustering is following as:

1. Initialize number of cluster  $k$  and centre.
2. For each pixel of an image, calculate the Euclidean distance  $d$ , between the center and each pixel of an image using the formula given below:

$$d = \|p(x, y) - c_k\| \quad (1)$$

3. Assign all the pixels to the nearest centre based on distance  $d$ .
4. Once all pixels have been assigned, new position of the centre is recalculated using the formula given below.

$$c_k = \frac{1}{k} \sum_{y \in c_k} \sum_{x \in c_k} p(x, y) \quad (2)$$

5. The process is repeated until it satisfies the tolerance or error value.
6. Reshape the cluster pixels into image.

$K$ -means clustering is easy to implement but it suffers from disadvantages. The main problem with this technique is that the quality of the final clustering results depends on the arbitrary selection of initial centroid. Therefore if the initial centroid is randomly chosen, it will get different result for different initial centers. So the initial center will be carefully chosen so that we get our desire segmentation. Another demerit is computational complexity which we need to consider while designing the  $K$ -means clustering. It depends upon the number of data elements, number of clusters and number of iteration [16].

Van Huy Pham and Byung Ryong Lee [21] introduced that a split and merge approach for Image segmentation. In this paper, the aim is to detect defects on fruit peel. In the method, the original image is over-segmented with  $k$ -means clustering method in  $L^*a^*b^*$  color space. After that by merging to the neighbourhood regions, small regions are filtered out. RAG is built based on regions obtained from previous stage to serve the merging process. Regions are iteratively merged based on minimum spanning tree technique. The results showed that this method is faster than graph-based algorithm because of using regions resulted from split procedure as the initial universe instead of pixels and higher quality than  $k$ -means only method. J. Macqueen [22] proposed that  $K$ -means clustering algorithm is an iterative technique used to classify the data points into  $K$  groups according to the similarity between them. It is one of the simplest unsupervised learning algorithms that solve the well known clustering. Oliver et. al. [23] proposed a new strategy for clustering segmentation which uses  $K$ -means clustering integrating region and boundary information. Jumb et. al. [24] integrated  $K$ -means and thresholding techniques to implement the segmentation of color image. Kochra [25] used the Hill-climbing with  $K$ -means algorithm for color image segmentation.

#### IV. COMPARISON

S No.	Paper	Year	Segmentation Type	Feature	Application Domain
1	Bora et al[26]	2015	Clustering	shape-based image segmentation	Brain MRI Images
2	R.K. Sasi et al[2]	2015	Split and Merge	IQM reduces lengthy neighbour	Neural Network Edge Pattern

				problems during merging	
3	J Vadher et al[3]	2015	Normalized Cuts	Discrimination much more clear	Medical Images
4	S Kansal et al[4]	2015	Region Growing	multiple criterions at the same time and give very good results with less noisy	Segment the parts of Human Body
5	S.Aja Fernandez et al[5]	2015	Thresholding	Overcome the common drawbacks of thresholding methods when images are corrupted with artifacts and noise	Medical imaging, Locate tumours
6	N. Senthilkumaran et al[6]	2009	Edge based	human perceives objects and works well for images having Good contrast between regions.	Medical imaging, face detection
7	Orlando J. Tobias et al[7]	2002	Fuzzy logic based	Pixels are divided into fuzzy sets i.e. each pixel may belong partly to many sets and regions of image	Gray scale Image can be easily transformed into a fuzzy Image by using Fuzzification Function
8	M Paulinas et al[8]	2015	Genetic Algorithm	Help to solve various complex image processing tasks in the future	Used on pattern's recognition applications
9	Hani K. Al-Mohair et al[9]	2015	Neural Networks	different combinations of color-texture descriptors were investigated to determine the optimal descriptor among the possible combinations	Artificial neural networks (ANN) are applied for Pattern Recognition.

## V. CONCLUSION

In this paper, different image segmentation techniques are studied. There are two main image segmentation techniques: Layer based segmentation and Block based Segmentation. Block based image segmentation is further divided into two main categories: Region based and Edge or Boundary based segmentation which are further divided into several categories. There is no universal segmentation method for all kinds of images and also an image can be segmented by using different segmentation methods.

## VI. REFERENCES

- [1] Nida M. Zaitoun and Musbah J. Aqel "Survey on Image Segmentation Techniques" International Conference of Communication, Management and Information Technology (ICCMIE 2015) Procedia Computer Science 65 (2015) 797 – 806.
- [2] Sasi, Remya K., and V.K Govindon, "Fuzzy split and merge for shadow detection" Egyptian Informatics Journal 16, No. 1 (2015): 29-35.
- [3] Vadher, Jagruti, "Normalized cut based image segmentation" (2015)
- [4] Kansal, Shweta and Pradeep Jain, "Automatic seed selection algorithm for image segmentation using region growing", International Journal of advances in engineering and technology 8, No. 3, (2015):362
- [5] Aja Fernandez, Santiago, Ariel Hernen Curiale, and Gonzalo Vegas-Sanchez-Ferraro. "A Local Fuzzy thresholding methodology for multiregion image segmentation." Knowledge- based Systems 83(2015): 1-12
- [6] N. Senthilkumaran and R. Rajesh (2009, May). "Edge detection techniques for image segmentation- A survey of soft computing approaches". International journal of recent trends in engineering, Information paper volume 1 (issue 2)
- [7] Orlando J. Tobias, Rui Scara (2002, December), "Image segmentation by histogram thresholding using fuzzy sets." IEEE TRANSACTIONS ON IMAGE PROCESSING, Volume 11(issue 12)
- [8] Paulinas, Mantas, and Andrius Usinskas. "A Survey of genetic algorithms applications for image enhancement and segmentation", Information Technology and control 36, No. 3 (2015)
- [9] Al-Mohair, Hani K., Junita Mohamad Saleh and Shahrel Azmin Saundi. "Hybrid human skin detection using neural network and K-means clustering technique" Applied soft computing 33(2015): 337-347
- [10] Satish Kumar, Raghavendra Srinivas, "A Study on Image Segmentation and Its Methods", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013.
- [11] Akhtaruzzaman, Md, Amir A. Shafie, and Md Raisuddin Khan. "Automated Threshold Detection for object segmentation in Colour image." ARPN Journal of Engineering and Applied Sciences, Asian Research Publishing Network (ARPN) 11, no. 6 (2016): 4100-4104.s

- [12]H.P. Narkhede, "Review of Image Segmentation Techniques", International Journal of Advanced Research in Computer Science and Modern Engineering (IJSCME) ISSN: 2319-6386, Volume-1, Issue-8, July 2013.
- [13] Verma, Kamal Kant, Pradeep Kumar, Ankit Tomar, and Mayur Srivastava. "A COMPARATIVE STUDY OF IMAGE SEGMENTATION TECHNIQUES IN DIGITAL IMAGE PROCESSING."(2015).
- [14]Vishal B. Langote, Dr. D.S. Chaudhari. "Segmentation Techniques for Image Analysis." International Journal of Advanced Engineering Research and Studies, Vol 1, Issue 2, Jan-March(2012), 252-255
- [15]Gupta Mehl, Patel Ankita, Dave Namrate, Gordia Rahul, and Savrin Seth, "Text-based Image Segmentation Methodology", 2<sup>nd</sup> International Conference on Innovations in Automation and Mechatronics Engineering, ICIAME 2014.
- [16]Nameirakpam Dhanachandra, Khumanthem Manglem and Xambem Jina Chanu, "Image Segmentation Using K-Means Clustering Algorithm", Eleventh International Multi-Conference on Information Processing-2015(IMCIP-2015).
- [17]K.Wagstaff, C. Cardie, Constrained K-Means Clustering with Background Knowledge, in: The Eighteenth International Conference on Machine Learning, 2001, pp. 577-584.
- [18]A. Halder, Color Image Segmentation using Rough set based K-Means Algorithm, Int.j.Comput. Appl.57 (12) (2012)32-38.
- [19]D. Sonagara, S. Badheka, "Comparison of Basic Clustering Algorithms, Int. J. Comput. Sci. Mob. Comput. 3(10)(2014)58-61.
- [20]Aimi Salihai, Abdul Yusuff Masor, Zeehaida Mohamed, "Color Image Segmentation Approach for detection of malaria parasites using various color models and K-means Clustering", in WSEAS Transaction on Biology and Biomedicine, vol 10, January(2013).
- [21]Van Huy Pham and Byung Ryong Lee, "An Image Segmentation Approach for Fruit Defect Detection Using K-Means Clustering and Graph Based Algorithm, Vietnam J Comput Sci (2015) 2:25-33.
- [22]J. Macqueen et. al., Some Methods for Classification and analysis of multivariate observations, in: Providings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol.1, Oakland, CA, USA, 1967, pp.281-297.
- [23] A. Oliver, X. Munoz, J. Batlle, L. Pacheco, J. Freixenet, Improving clustering algorithms for image segmentation using contour and region information, in: Automation, Quality and Testing, Robotics, 2006 IEEE International Conference on, Vol. 2, IEEE, 2006, pp.315-320.
- [24] V. Jumb, M. Sohani, A. Shrivats, Color image segmentation using k-means clustering and otsus adaptive thresholding, Int. J. Innov. Technol. Explor. Eng 3 (9) (2014) 72–76.
- [25]S. Kochra, S. Joshi, Study on Hill Climbing Algorithm for Image Segmentation.
- [26] Bora, Dibya Jyoti, and Anil Kumar Gupta, "A novel approach towards clustering based image segmentation" arXiv Preprint arXiv: 1506.01710(2015).

# Review Paper on Big Data And Implementation Using Hadoop And Grid Computing

Simranjot Kaur\*

Er. Sikander Singh Cheema\*\*

\*Research Student, M. Tech. (CE), Department of Computer Engineering, Punjabi University Patiala

\*\*Assistant Professor, Department of Computer Engineering, Punjabi University Patiala

**Abstract:** The rapid growth of Internet and WWW has led to vast amount of data to be captured, processed and analyzed to get correct information. Such type of data can be in structured, unstructured and semi structured form. The following part of the paper is devoted to the characterization of tools, techniques and implementation of Big Data using these techniques. For implementation, this paper focus on some methods to use Grid Computing along with Hadoop. Hadoop is used to analyzed the data and Grid Computing is used to provide large storage capacity and computation power. This paper also listed some open source toolkit used to implement solution such as Hadoop, Globus toolkit.

**Keywords:** big data, grid computing, hadoop distributed file system, map reduce, processing capacity, Storage capacity.

## I. INTRODUCTION

Growth of technologies and services produces large amount of data which can be in structured, semi-structured and unstructured form. Such type of data contains billions records of millions people information is very difficult to process includes audios, images, social media etc. Big Companies like Google, Yahoo, Facebook etc. generates massive datasets in unstructured form. So for the processing of this vast amount of data, we need of Big Data Analytics. Big Data Analytics refers to the process of collecting, analyzing and organizing the large amount of datasets.

This vast amount of data is very difficult to be capatured, managed, processed and analyzed by relational databases or desktop statistics within time limit. So some other techniques and tools are there which need to be understood, because Big Data Analyzing and Implementation of Big Data are the challenges for researchers system that need special techniques for this purpose.

HDFS, the Hadoop Distributed File System, is a distributed file system designed to run on commodity hardware and is inspired by the Google File System. It is designed to hold large amounts of data (terabytes or petabytes or even zetabytes), and provide high-throughput access to this information.

Secondly, Hadoop Map Reduce is a technique which analyze big data. The term MapReduce refers to two separate tasks map and reduce that Hadoop programs perform. Hadoop provides reliability by replicating data over multiple nodes.

Another one is Grid Computing which is used to satisfy the computing power. Globus toolkit, a computing tool is available for this.

As the Big data is the latest technology that can be beneficial for the business organizations and social networking etc, so it is necessary that various issues and challenges associated with this should bring out into light and to be understood. The main problems regarding big data are capture, storage, search, sharing, analytics the storage capacity and the processing of the data.

#### A. *Big Data: Definition*

Big data refers to the data sets or combination of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed and analyzed by relational databases and desktop statistics within time limit. It works with petabytes, exabytes or even zetabytes of data. As Big Data is too big in different forms so characterized by the 3 V's.

These are Volume, Variety, Velocity.

##### *Volume of data*

Volume of data refers to the amount of data. It stores data in megabytes and gigabytes to petabytes.

##### *Variety of data*

Variety of data means different types of data and sources of data which includes semi-structured, unstructured, audio, video, XML etc.

##### *Velocity of data*

Velocity of data refers to the speed of data processing. It is used to analyzed the increase in profit of business before the information lost.

## II. TOOLS AND TECHNIQUES

#### A. *HADOOP*

Hadoop is a programming framework which was developed by Google's MapReduce. It helps in processing the large amount of datasets. It uses divide and conquer method to breakdown complex data into smaller units. The functionality of Hadoop includes two stages:

##### *Map()*

In this, master node takes the information and then divide it into subparts. After that, subnodes are distributed into worker nodes. The function of the worker node is same as the master node, it further divides the subparts which leads to the multi-level tree structure. So at the end, worker node send answer to the master node.

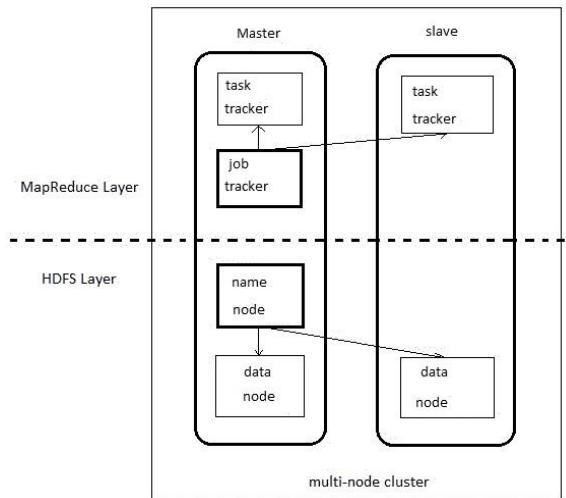


Fig.- Architecture of Map Reduce Function

### *Reduce()*

In this, answers generated by all the subparts are being collected by master node and then master node combine all the answers to make one output.

### *B. HDFS*

Hadoop Distributed File System is designed to run on commodity hardware and is inspired by the Google file system. It is used to hold large amount of data. Data is stored in blocks also known as ‘Chunks’. It is a client-server architecture which contains one Name Node and multiple Data Nodes. Data about the Data Node is stored in the Name Node.

Some other components of Hadoop are:

### *HBase*

It is a open source system which is written in Java and runs on the top of the HDFS. It can be treated as the input as well as output of the MapReduce.

### *Pig*

Pig is a high level program. MapReduce program are created here.

*Hive*

Hive provides SQL interfaces and relational model. It build on the top of the Hadoop and provides summarization, query, analysis.

*Sqoop*

Sqoop is used to transfor data between relational database and Hadoop.

*Chukwa*

It is used to process and analyze the large amount of data. It is also built on the top of the HDFS and MapReduce framework.

*Flume*

It focused on streaming of data from different sources.

*C. HPCC*

It is an open source computing platform which is defined by user. This system is used to manage complex problems. HPCC is not only single platform system but single architecture and a single programming language which is used to process the data.

The main components of HPCC are :

- a) HPCC data refinery
- b) HPCC data delivery
- c) Enterprise Control Language

### III. IMPLEMENTATION USING HADOOP

Hadoop provides two services, Map Reduce and Hadoop Distributed File System(HDFS). HDFS file system has one Name Node and multiple Data Nodes. On the other side, in Map Reduce engine job tracter present on Name Node and task trackers on Data Node.

*A. Name node*

In this node, files are categories into number of Data blocks and Data blocks are further stored on Data Node. These Data blocks are replicated on different Data Nodes for the purpose of security and reliability. All the records are maintained by Name Node. It contains metadata about all the Data blocks. So if any client requesting data, contacts to Name Node which gives the location of data on Data Nodes. Secondary Name Node is used for the purpose of maintenance to avoid problems occurs due to the failure of primary Name Node.

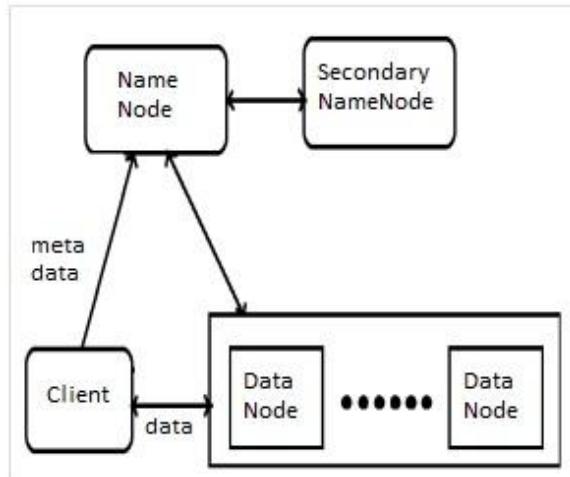


Fig. Architecture of Name Node

#### B. Data Node

DataNodes store application data in the format of data block. When DataNode starts, it does handshake with NameNode. Handshake process registers DataNode to NameNode. DataNode is registered with unique storage id. It identifies replicas present on it and send block report to NameNode . NameNode stores metadata of data blocks present on that DataNode from block reports. Task trackers on Data Node executes jobs on DataNode.

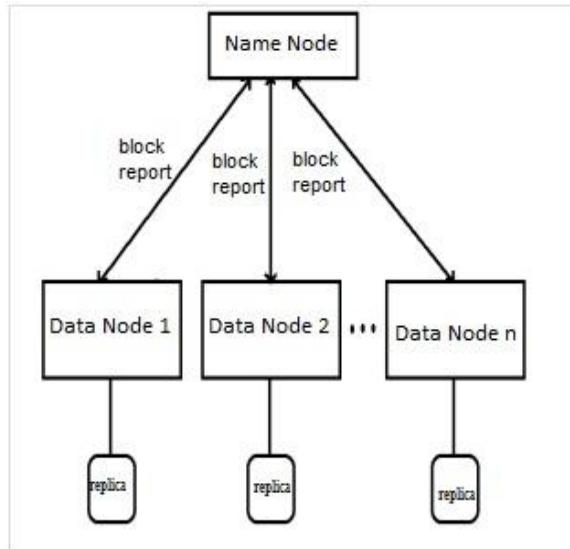


Fig. Architecture of Data Node

#### C. Job tracker and Task Tracker

The Main purpose of job tracker and task tracker is to execute map and reduce functionalities on different nodes. Job is submitted to the node on which job tracker runs. Job tracker has information of the node on which data blocks are present and it send jobs on respective task trackers for execution purpose. Task trackers execute jobs sent by job tracker and Job tracker checks status of task tracker after every few minutes.

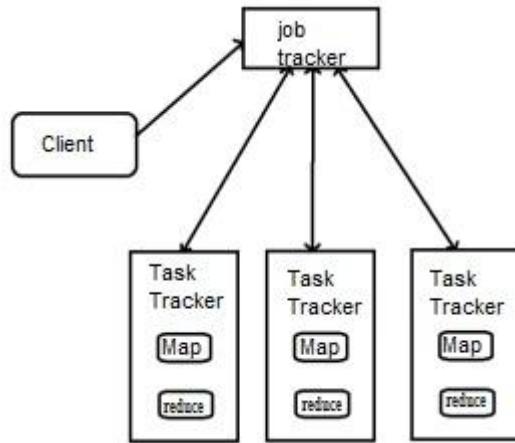


Fig. Architecture of Job tracker and Task tracker

#### D. Problems in Hadoop

1. NameNode failure due to limited capacity of storage.
2. Searching the huge metadata at NameNode may go inefficient due to its size.
3. There is no inbuilt facility for parallelization of jobs.

### IV. IMPLEMENTATION USING GRID COMPUTING

Grid computing is the combination of computers connected together using network and sharing resources. It increases computational power with the use of underutilized resources on network and decrease time requirement for job execution. Jobs executing on grid are non interactive and this is the main difference between Grid and distributed computing. Jobs are executing in parallel way. Grid computing creates a virtual computer whose hardware and software resources are scattered. It has one control node which controls all other nodes. Control node is responsible for resource sharing and job execute on different number of nodes.

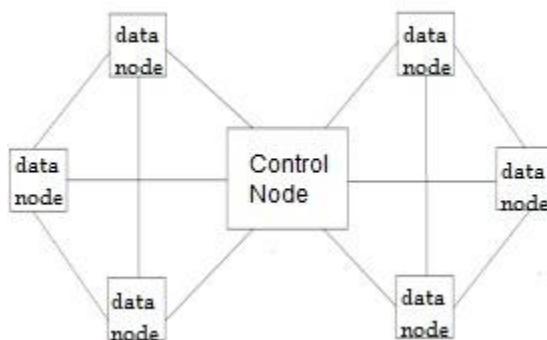


Fig. Grid Computing Architecture

As given in figure, Grid computing has one control node and all others are grid nodes connected to each other. User can send job from any node and control node decides node on which job is going to execute. It seems like a single

system to the user when user access grid network. Open source tools such as Globus Toolkit can be used to create computer grid network. Globus toolkit has GRAM, GSI, GridFTP components which provides all services required for grid computing. GRAM service controls all job execution tasks. GRAM service uses job manager such as condor to locate job executing node GSI handles security and GridFTP is used to transfer data between different nodes.

## V. COMBINE ARCHITECTURE OF HADOOP AND GRID COMPUTING

During Big Data analysis, Hadoop shows some problems in storage and computation power. But Grid computing is used to increase the computation power. So to reduce the problem faced in Hadoop, we combine these two architecture.

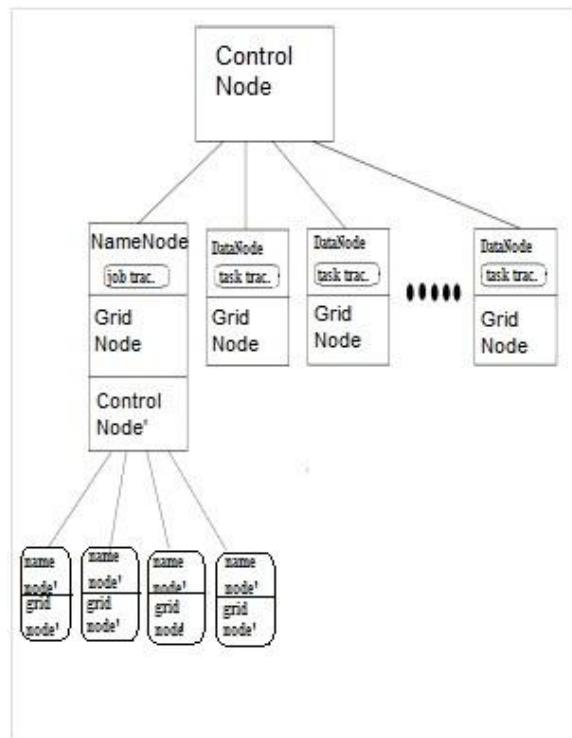


Fig. Combine architecture of Grid and Hadoop

Very First grid node act as NameNode of Hadoop and all other grid nodes are act as DataNodes. One more sub grid is formed with NameNode act as control node and different number of grid nodes are attached to it.

Solution for first problem is given by this sub grid. Each grid node of sub grid shares storage resources. Therefore NameNode can use this storage memory to store metadata. All storage memory treated as only one system. User has no knowledge about grid storage. Whole sub grid is treated as one Name Node. So problem of storage is solved by combining Hadoop and grid together. Second problem of computation power is also solved by sub grid. It can parallelized the job of searching metadata of particular data block on number of grid nodes which minimizes time require to searching in metadata. Sub grid increases computation power of NameNode.

Third problem is parallelizing map reduce jobs on DataNodes. Hadoop cannot parallelize map reduce problem but grid can. If one job has to process more than one data block on one DataNode, then it find any idle system and

parallelized by transferring next data blocks to that idle systems and execute same job on that system. So that number of data blocks processed parallel according to the number of idle systems. The process of transfer and execution of job cannot be done by only Hadoop. Grid helps hadoop to parallelize jobs. Grid handles transfer of data between DataNodes and execution of job. So third problem can also solved by integration of grid and hadoop system.

## VI. CONCLUSION

In today's era of technology, data is used in a big amount. We daily use various forms of data as mails, transactions, videos and pictures etc. and increasing day by day. So as increased use of data, the problem of its storage is also increasing. To store such a big data, a single database is not a good idea. So when we deal with data storage, Big Data is an important term. It is the leading technology used by most of the organizations to store data. Big Data is managed by the use of Hadoop, which consists of huge volume of clusters of data. Hadoop includes many technologies, a few of which are discussed in this paper. This paper gives information about tools of Hadoop such as flume, sqoop, HBase, Hive, Pig, MapReduce, HDFS and HPCC. The implementation of Hadoop and related technologies like Grid Computing is discussed here. Various types of data is managed with Hadoop but if we face any exception such as the data which is not easy to manage through Hadoop that can be managed by using a combination of technologies Hadoop and Grid Computing. Grid Computing is used to overcome the drawbacks of Hadoop. It gives a good start up information about managing Big Data. Such technologies are very important in future and almost all the organizations are going to implement these technologies.

## VII. REFERENCES

- [1] Kaur, M. (2013). BIG Data and Methodology-A review. International Journal of Advanced Research in Computer Science and Software Engineering, 5.
- [2] Sabia, S. K. (2014). Applications of Big Data. International Journal on Advanced Computer Theory and Engineering (IJACTE), 5.
- [3]"ApacheHadoop" [http://en.wikipedia.org/wiki/Apache\\_Hadoop](http://en.wikipedia.org/wiki/Apache_Hadoop)
- [4]"GridComuting" [http://en.wikipedia.org/wiki/Grid\\_computing](http://en.wikipedia.org/wiki/Grid_computing)
- [5] Yuvraj S. Sase, P. A. (2014). Big Data Implementation Using Hadoop and Grid Computing. International Journal of Innovative Research in Science, Engineering and Technology , 6.
- [6] Harshawardhan S. Bhosale, P. D. (2014). A review paper on Big data and Hadoop. International Journal of Scientific and Research Publications , 7.
- [7] C.Chandhini, MeganaL.P , P. A (2013). Grid Computing-A Next Level Challenge with Big Data. International Journal of Scientific & Engineering Research , 7.

# A Review Paper on Crop Models DSSAT and CropSyst

Parmeet Kaur\*

Er. Sikander Singh Cheema\*\*

\*Research Student, M. Tech. (CE), Department of Computer Engineering, Punjabi University Patiala

\*\*Assistant Professor, Department of Computer Engineering, Punjabi University Patiala

**Abstract:** Crop Simulation Models helps to estimate crop yield on the basis of some parameters like temperature, soil, weather and crop etc. DSSAT-CSM (Decision Support System for Agro-technology Transfer – Cropping System Model) is an extension of DSSAT crop model. DSSAT-CSM contains a modular structure which has various modules that work along various parameters and give the desired output. CropSyst is a user friendly, multi-year, multi crop, daily time step cropping system simulation models which is used to study the effect of cropping system management on productivity and the environment. It takes the factors such as soil-water budget, crop canopy and root growth, soil-plant nitrogen budget, residue production and decomposition and management options like crop rotations, irrigation, nitrogen fertilization and tillage operations into consideration. DSSAT and CropSyst both are used to analyse growth production but based on different input parameters.

**Keywords:** cropping system; crop simulation; DSSAT; ClimGen; CROPGRO module; CropSyst; nitrogen budget; decision system

## I. INTRODUCTION TO DSSAT-CSM

DSSAT was developed by an international network of scientists, cooperating in the International Benchmark Sites Network for Agro-technology Transfer Project (IBSNAT) to get the estimated forecasting of any crop on the given field area. The main difference between DSSAT and DSSAT-CSM is that DSSAT contains soil model components for each individual crop model whereas DSSAT-CSM contains a single soil model for all crop models.

The main aims of DSSAT-CSM can be defined as (1) to estimate crop yield on the basis of weather, genetics, soil water, soil carbon, nitrogen and other management components, (2) it provides information about mono-crop production and crop rotations at any location where minimum inputs are provided, (3) helps to give idea of management of crop such as harvesting, cropping, irrigation dates estimates and it helps to easily compare forecasted data and actual yield.

## II. DESCRIPTION OF COMPONENTS

The DSSAT-CSM is based on a modular structure where modules are independent of each other on the basis of working and linked to each other. It helps to plugged in or plugged out different modules without affecting the main program.

It consists of three main modules which consist of a main driver program, a Land Unit module and a Primary module. The description of all these modules can be defined as follows:

#### *A. Main Driver Module*

In this each module consists of six operational steps that are- run initialization, season initialization, rate calculations, integration, daily output and summary output. This module read inputs, initializes itself, compute rates and integrate its state variables. The working of this module is completely independent of the working of other modules. This mode provides interactive sensitivity analysis and comparison of simulated versus observed field data. This mode checks for weather conditions and according to that estimate irrigation time and estimate yield of crop. This module gets the input from Land Unit module. Further each sub module also perform same functionality that is each module perform six operational steps. This module decides which module is to be called based on the input given by the user.

#### *B. Land Unit Module*

This module acts as an interface between Main Driver module and other modules. In the starting of main program, it calls the land unit module to initialize variables. Land unit module gives the information about soil. The main program calls land unit module at start and end of the season to calculate variables and to summary output. The primary module also passes its interface variables that are the required inputs to the land unit module. When a new crop season starts, it gets management information from DSSAT input file that is from the database.

#### *C. The Primary Modules*

The primary module consists of various other module which passes information to the land unit module. The sub modules in primary modules are named as- weather module, soil module, soil-plant-atmosphere module(SPAM), CROPGRO Template module, plant module and management module.

##### *1. Weather Module*

Weather module is used to generate or read daily weather data which includes the parameters like daily weather values (maximum and minimum air temperature, solar radiation and precipitation, relative humidity and wind speed) and also rainfall, day length and atmospheric carbon dioxide concentrations.

##### *2. Soil Module*

Soil module gives information about soil, soil consists of vertical layers. Soil module can read its inputs from a file. Soil module has various sub-modules, which are explained as below:

###### *1.1 Soil water sub module*

This module is used to compute daily changes in soil water content by soil layer due to infiltration of rainfall and irrigation, vertical drainage etc. Soil water infiltration is calculated by subtracting surface runoff

from rainfall that falls on that particular day. The irrigation water content is added to rainfall to calculate soil infiltration.

#### *1.2 Soil carbon and nitrogen balance sub module*

This module calculate carbon and nitrogen ratio. This module gets information of water flux values from soil-water sub module which tells about the transportation of N through deeper layers of soil. This also computes values for organic and inorganic fertilizers and residue placements. It also updates the values of soil nitrate and ammonium concentrations on daily basis for each layer.

#### *1.3 Soil temperature sub module*

The temperature of soil is calculated in this module, which is computed from air temperature and deep soil temperature boundary condition which is calculated from average annual air temperature and amplitude of monthly mean temperatures.

#### *1.4 Soil-Plant-Atmosphere Module*

This module computes soil evaporation, plant transpiration, and root water uptake processes. This module takes soil, plant and atmosphere inputs together and compute the values for all these processes. This module also needs leaf area index (LAI) and root length density of each layer as inputs.

### *3. CROPGRO Template Crop Module*

This module has one common source code and it is used to predict growth of many crops. CROPGRO module accepts various parameters such as Photosynthesis, Respiration, Carbon and nitrogen mining parameters, Nitrogen fixation parameters, Phenology parameters and Canopy height and width growth parameters etc. Phenology is an important component of CROPGRO template module. Phenology is defined as the study of periodic plant and animal life cycle events and how these are influenced by seasonal and inter-annual variations in climate, as well as habitat factors. It uses cardinal temperature values and information from cultivar and ecotype files. Life cycle progress of any plant depends on physiological day accumulator as a function of temperature and day length, because some crops like Soybean are sensitive to day length.

### *4. Plant Module*

This module does same work as CROPGRO module but it defines same parameters for individual crop. It is based on CERES –maize, wheat and barley models. In CERES model, plant growth is divided into various stages that depend on the crop for which we are defining it. For example- the stages for maize are germination, emergence, end of juvenile, floral induction, beginning grain fill, maturity and harvest. The stages for wheat are germination, emergence, terminal spikelet, end ear growth, beginning grain fill, maturity and harvest. The rate of development or growth of plants is calculated by thermal time or growing degree days (GDD), which are based on maximum and minimum day temperatures.

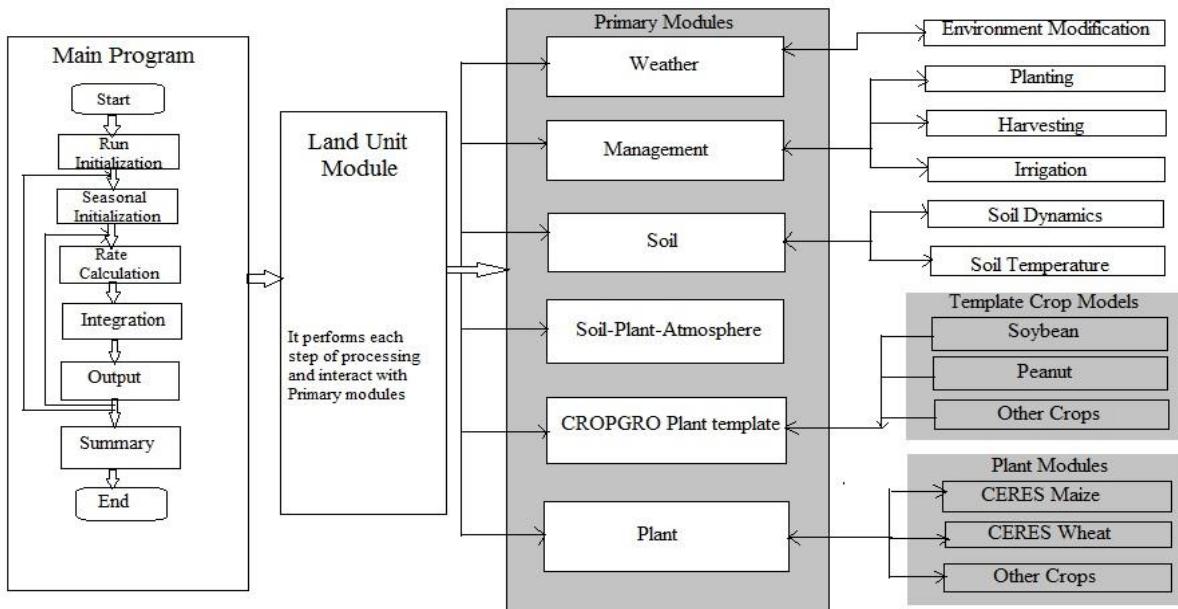


Fig.- Components and modular structure of DSSAT-CSM

### 5. Management Module

This module defines when to perform field operation and also gives options to manage the crops. These field operations contain planting, harvesting, applying inorganic fertilizer, irrigation and applying crop residue and organic material. This module gives the dates for all these operations. For example harvesting dates are given on the basis of maturity of crop or according to soil water conditions of the field. Irrigation dates are fixed on the basis of rainfall and required amount of water by the crop and soil.

### III. INTRODUCTION TO CropSyst

CropSyst is a multi-year, multi crop simulation model which consider daily time step. It is developed by a team at Washington State University's Department of Biological Systems Engineering. The CropSyst is linked to a GIS software, a weather generator and other utility programs to calculate daily data and to make it user friendly. CropSyst simulates crop growth and various other factors like crop canopy(the above ground portion of plant community or crop, formed by the collection of individual plant crowns), soil water budget, soil-plant nitrogen budget, crop phenology(study of periodic plant and animal life cycle events), biomass production, residue production and decomposition etc. The main feature of this model is that it enables crop rotation.

### IV. DESCRIPTION OF COMPONENTS

CropSyst is a suite of programs that work together co-operatively. CropSyst is mainly divided into four modules or parts that are: CropSyst Parameter Editor, CropSyst model(cropping system simulator), ClimGen(a weather generator), ArcCS(a GIS-CropSyst simulation co-operator), CropSyst Watershed(a watershed analysis tool) and other utility programs. The description of these modules is as under:

#### A. *CropSyst Parameter Editor*

It is the main user interface which interacts with the user and asks for input. It helps to modify and set CropSyst parameters, run the model and display the output. It includes separate component input files for location, soil, crop and management data. Weather files are simple text files which can be created by any editor of user's choice. This editor runs simulation files and these simulation files consist of the period of simulation (starting and ending dates) and specification of soil, location, crop rotation and management practices. This editor also contains parameter files, which are: Simulation control, soil, crop, management, location and batch run.

#### B. *CropSyst Model*

This simulator is the core of the suite of programs. It has various processes or procedure to simulate productivity of crop and crop rotations. This model simulates crop growth over a single land unit fragment with uniform soil, weather and management conditions. This model calculates following values:

##### 1. *Water Budget*

It includes precipitation, irrigation, runoff, interception, water infiltration, crop transpiration and evaporation. Actual transpiration and soil evaporation depend on the water availability in the soil profile explored by roots and soil surface respectively. Precipitation is taken directly from daily weather files. Irrigation amounts and timing are either user selected or automatically assigned if the automatic irrigation option is selected.

##### 2. *Nitrogen Budget*

It calculate separate budget for nitrate and ammonium and is calculated with the help of processes such as N transformation, ammonium sorption, crop N demand and crop N uptake. Crop nitrogen demand is the amount of nitrogen that is needed by a crop to grow properly and without any deficiency. The Nitrogen concentration of water remaining within and leaving a layer is calculated to find out nitrogen budget.

##### 3. *Crop Phenology*

Crop development is based on thermal time, which is accumulation of average air temperature above a base temperature and below a cut-off temperature to reach given growth stages. The thermal time or Degree Days are calculated from planting and it should be specified for each stage of crop. The base and cut-off temperatures may vary depending upon location and cultivar. It also depends upon water stress, which tends to increase crop canopy temperature, which may accelerate accumulation of degree-days.

##### 4. *Leaf Area Development*

Leaf area development depends on daily biomass production. It is represented as Leaf Area Index (LAI). Leaf area duration is expressed as thermal time units that are degree-days, it is assigned to each unit of daily LAI produced. When daily LAI duration completes its function it is removed from current LAI and water stress corrections are applied to both leaf area duration and leaf area development. The leaf area expansion-

related biomass accumulation is used to determine green area index (GAI) during the active growth stage. The value of LAI and GAI is same during active growth stage.

#### 5. Root Growth

It is defined in terms of root depth and root density. It comes under Crop Morphology part of simulator. The maximum root depth is measured in meters from soil surface when plant reaches the end of vegetative growth. Root density is assumed to be zero at soil depth equal to current root depth, and it increases linearly to a maximum at the depth near the soil surface. Its information comes from empirical measurements, observed water extraction patterns or from crop morphology parameter table which has values for some given crops.

#### 6. Crop Rotations:

It performs simulation by using a daily time step within a period specified by start date and end date. State variables like evapotranspiration, soil water content, soil N content are updated daily and rotation of crop is decided according to a sequence given by crop rotation template. CropSyst has crop rotation table, which tells about simulation of multi-year crop rotations. The user can add or delete any crop in rotation using management mode of CropSyst model.

#### C. ClimGen

Weather generators are computer programs that are used to analyze weather conditions by using the existing weather data. The properties of generated data are expected to be same as that of actual weather conditions. ClimGen is a weather generator which is used to generate precipitation, daily maximum and minimum temperature, solar radiations, wind speed and air humidity. It uses Weibull distribution to generate precipitations amount. It is easy to parameterize and can be applied to any world location when we are provided with enough information to parameterize the program.

#### D. ArcCS

ArcCS is a simulator environment extension which facilitates GIS-based CropSyst simulation projects. It deals with the database files generated by ARCVIEW or Arc/Info GIS. The common file used is Polygon Attribute files. Each polygon represents a land block fragment. This table is used to identify, generate and run a simulation scenario for each unique land unit block. A new polygon table is generated which can be used by ARCVIEW to produce maps of CropSyst output. The input and output variables can be differentiate by using an option “Separate by crops”. In this normal CropSyst parameter files are used.

#### D. CropSyst Watershed

CropSyst watershed use ARCVIEW for Windows and its Spatial Analyst extension as geographical base. It controls watershed simulation projects and watershed wizard in CropSyst consist of a set of pages describing each step in preparing and running a watershed simulation. In this land block fragment are known as raster cells in a grid and all these are hydrologically connected. The Spatial Analyst will allow user to define watershed

boundaries and drainage network from digital elevations model data. These simulations are run from upper to lower elevations until the entire watershed is covered.

#### E. Utility Programs

CropSyst provides various utilities to handle output and to estimate soil hydraulic parameters and global solar radiations. It includes a dynamic link library that can be used in other models to calculate reference crop evapotranspiration. Other utilities in CropSyst simulator are report viewer which is used to display output files from last simulation run, graphics viewer is used to graph daily variables from simulation run and standalone validation program is used for detailed validation of the simulation and included parameter files.

### V. DATA REQUIREMENTS OF CROPSYST MODEL

CropSyst work on data files and create the required output. The main five types of files used in CropSyst are : Simulation Control, Location, Soil, Crop, Management and Batch run files. Simulation control files combine various input files and specify start and end date of simulation and crop rotation. It also controls simulation of soil erosion, soil salinity and nitrogen and carbon dioxide effect on crop growth. Location parameters give the information about site of crop. It identifies daily weather data files and latitude to generate solar radiation. Soil file consists of the information about general characteristics of soil such as soil type, soil pH, soil texture profile and soil hydraulic properties. Crop file contains a set of parameters that are description of crop, planting, growth, morphology (Maximum LAI, root depth, specific leaf area, leaf area duration), harvest, nitrogen, residue files etc. It allows user to select classification and simulation options for crop. The management file includes scheduled and automatic management events. The main management file sections are harvest, irrigation, nitrogen, conservation and tillage. Managed events are planned according to actual date and relative date of planting, scheduled events includes irrigation and nitrogen fertilization and automatic event manager checks soil water and nitrogen content. Batch run editor is an optional utility which runs a list of simulations. It is used to run multiple simulations.

### VI. CONCLUSION

Agriculture is the most important occupation as it feeds the whole world. Most of the people get their earnings from agriculture. Agriculture is dependent in various factors such as type of crop, soil, weather, manpower, tools and management of all these factors. It is very beneficial to know about crop growth before the actual planting and harvesting of crop, just by analyzing these factors to the past estimates. It gives estimation of crop yield and required efforts to maintain a crop and tells that which crop and weather conditions will be effective for a given crop. It reduces the effect of loss, if any type of problem will occur. This crop estimation is done by using various Crop Simulation Models, two of these models are discussed in this paper. DSSAT and CropSyst both are used to check crop simulation. DSSAT is a simple framework which has various modules under it and all the modules work independently. It provides information about various crop that are stored in its database and all the weather information from files stored in it. CropSyst has various parts which perform its functions and it is also linked with a GIS software which is used to update daily weather data. Due to GIS, it can be applied to any world location because it can analyze each

site and can get soil and weather properties. It also provides crop rotation between multiple crops. Both these models are equally beneficial and are made advanced day by day.

## VII. REFERENCES

- [1] Chunlei, R. W. (2013). Application of DSSAT model to simulate weather growth in China. *Canadian Center of Science and Education* , 10.
- [2] Claudio O. Stockle, M. D. (2003). CropSyst, a cropping systems simulation model. *European Journal of Agronomy* , 19.
- [3] Claudio O. Stockle, R. N. ( ). Cropping Systems Simulation Model : User's Manual. *Biological Systems Engineering Department* , 235.
- [4] Eric K. Forkuo, A. K. (2011). Digital Soil Mapping in GIS Environment for Crop-Land Suitability Analysis. *International Journal of Geomatics and Geosciences* , 14.
- [5] J.H.M. Wosten, A. L. (1999). Development and use of a database of hydraulic properties of European soils. *Geoderma* , 17.
- [6] J.W. Jones, G. H. (2003). The DSSAT cropping system model. *European Journal of Agronomy* , 31.
- [7] Kazeem O. Rauff, R. B. (2015). A Review of Crop Growth Simulation Models As Tools for Agricultural Meterology. *Scientific Research Publishing* , 8.
- [8] Kelly R. Thorp, K. C. (2008). Methodology for the use of DSSAT models for precision agriculture decision support. *Computers and Electronics in Agriculture* , 11.
- [9] Moritz Reckling, J.-M. H. (2015). A cropping system assessment framework- Evaluating effects of introducing legumes into crop rotations . *European Journal of Agronomy* , 12.
- [10] Murthy, V. R. (2000). Crop Growth Modeling and Its Applications in Agricultural Meterology. *Satellite Remote Sensing and GIS Applications in Agricultural Meterology* , 27.
- [11] P. Banterng, G. H. (2010). Application of the Cropping System Model (CSM)-CROPGRO-Soyabeanfordetermining Optimum Management Strategies. *Journal of Agronomy and Crop Science* , 12.
- [12] S.K. Jalota, G. S. (2006). Performance of CropSyst Model in Rice-Wheat Cropping System. *Journal of Agricultural Physics* , 7.
- [13] Samida Ouda, M. M. (2015). Parameterization of CropSyst Model for four Wheat Cultivars in Egypt. *Global Journal of Advanced Research* , 11.
- [14] Wu, R. C. (2013). Applications of DSSAT model to simulate weather growth in China. *Canadian Center of Science and Education* , 10.

# Penetration Testing-

## Assessing all the vulnerabilities before an intruder can do

Gagandeep Kaur<sup>#1</sup>, Dr. Jaswinder Singh<sup>\*2</sup>

#Department of Computer Engineering, Punjabi University, Patiala  
Patiala, Punjab, India

\*Assistant professor, Department of Computer Engineering, Punjabi University, Patiala  
Patiala, Punjab, India

**Abstract**—Penetration testing is the testing to verify the security of a Website, server and network by safely trying to exploit vulnerabilities. Security is the most important issue in today's world .This paper outlined various features of penetration testing such as importance of penetration testing, steps of performance and identification of vulnerabilities. This paper reviewed the various online Penetration testing tools and comparison between them. Kali linux has wonderful set of inbuilt tools of kali linux for the penetration testing. Some of these inbuilt tools have been discussed in this paper.

**Keywords**:- Penetration testing; Vulnerabilities; kali linux; Metasploit;

### I. INTRODUCTION

Penetration testing aimed to find the vulnerabilities in computer or network system that simulates attack by the hacker. The main difference between the penetration tester and hacker is pen-tester has license and permission granted by the organization. The pen-tester identify the weaknesses of system and report back to the organization who have grant the permission, with their findings. These vulnerabilities may exist in service of operating systems, unwarranted configurations, or risky quit customers. A Pen Testing is performed by testers, Network specialist, security consultants. There are numerous steps to perform the penetration. Inside the first step, it defines set of Vulnerabilities/potential problem areas that would result in a security breach for the systems. In step second, this set of Vulnerabilities must be ranked in the order of priority. In the third step consider the high-risk vulnerabilities as in comparison to low-risk vulnerabilities. Inside the fourth step Apply penetration tests that would work (attack your system) from both within the network and out of doors (externally) to determine if you can gain access to data/network/server/websites unauthorized. In step fifth, If the illegal access is achievable, the system must be corrected. The series of steps need to be repeated until the vulnerabilities get eliminated.

The main aim of Penetration testing is on the security to different computer resources such as confidential documents, databases and other information that should be kept secret. The questions arise in the mind of researchers are:-

- Why is penetration testing required
- How vulnerabilities are identified
- What are the various methodology and techniques involved in penetration testing
- Which tools are required to perform penetration testing

This paper attempts to answer above questions by examining the papers of time period from 2008 to 2012(0). We review the articles briefly that enlightened the penetration testing and tools of penetration testing.

#### *A. Need of Penetration Testing*

Penetration testing is important to determine how an intruder attacks the system. It protects the confidential information that is needed to be kept secret. It helps to find weak areas where the attacker can gain access to the computer's private data and can destroy it. Penetration testing can be applied when system discover new danger by intruder. This estimated magnitude of invade of potential business. It provides facts to suggest, why it is important to increase investments in security part of technology. There are various situations when we can use penetration testing such as when we add new network to our system, install new software or update our system.

#### *B. Procedure to identify Vulnerabilities:-*

Weaknesses need to be determined by both the penetration specialist and the vulnerability scanning. The steps are very similar for the security tester and an illegal attacker. The attacker may choose to proceed more slowly to avoid recognition, but some penetration testers will also start gradually so the target company can learn where their recognition threshold is and make improvement. This kind of is where the specialist attempts to learn as much as possible about the target network as possible. This normally is dependent on identifying publicly accessible software program as email and web servers from their service banners. A large number of server will report the Operating System they are running on, the version society they are operating, patches a modules which may have been enabled, the current time, and perhaps even some internal information as an internal server name or IP address. Once the tester comes with an idea what software might be working on the prospective computers, that information must be tested. The tester really doesn't know what is operating but he may have a pretty good idea. The information that the tester has can be combined and then as opposed with known vulnerabilities, an then those vulnerabilities can be tested to see if the results support or contradict the information. In penetration test, these first steps may be regular for some time before the tester decides to launch a specific harm. In the case of a strict vulnerability analysis, the attack may never be launched hence the owners of the target computer would never really know if this was an exploitable vulnerability or not.

## II. ONLINE PENETRATION TESTING TOOLS

#### *A.Ping sweep:-*

A ping sweep is a central system checking procedure used to find out which of a scope of IP locations decide the live has. It clears all the live has inside system range which is extremely vital for infiltration testing to decide the entire damage surface of customer. In this testing is just done to the live has. Most extreme 256 IP locations can be swepted in a row(1). Titled ping breadth will attempt to discover the DNS name connected with each live hosts(2). Ping range can send ICMP ECHO interest to

numerous hosts. On the off chance that the host is alive, it will in the long run sent back the ICMP ECHO REPLY. To cripple titled ping clears on a system, executives can square ICMP ECHO asks for from outside source.

*Port Detective:-*

A Port Detective is the online penetration testing tool which determines that which ports are opened and which ports are blocked. Port Detective tool is developed by Tzolkin corporation. It is the most simple tool to find out the status of ports (open, blocked or in use). Port Detective is the open source tool that will help the user to scan his/her computer at current location. Then the user will grant the permission by clicking on OK button. After scanning user's computer port detective will keep the information confidential. The interface used by port detective is very straightforward. Then the user is presented with the list of port from where the user can check the ports which are in use or which ports are blocked.

*Advanced port scanner:-*

Advanced port scanner that quickly discovers the open ports and get the versions of program running on detected port. It can check every one of user's ports or only those in extents user characterize, and on user's primary PC as well as on all user's arranged PCs. The system has an easy to understand interface and rich functionality.

*PORt SCANNER ACTIVE X CONTROL:-*

Port Scanner Active X Control to integrate port-scanning capabilities into applications. Port Scanner can be used for network exploration or security auditing. Port Scanner can determine what ports are open by remote hosts. It is capable of scanning multiple hosts simultaneously.

- Port Scanner can be configured to scan the protocols as follows:
  - TCP ports only
  - UDP ports only
  - TCP and UDP ports simultaneously
- It can operate in 3 different scan modes:
  - Scan range of ports
  - Scan authorized ports

*Port Scanner :-*

Proport or iportscan is the port scanner instrument that is utilized for the IPHones or IPodTouch. This tool is valuable for framework administrators who assess what administrations are listening on a known framework. This is exceptionally helpful for the framework administrator who can utilize this instrument to rapidly port sweep the majority of his frameworks to ensure nothing is open that shouldn't be. This is an unquestionable requirement have system device for any security expert, programmer, or system/framework administrator. You can portscan from the 3G or EDGE system to test whattcp ports are open from the web.

*Retina Scanner:-*

Retina scanner find vulnerabilities across Network, Web, Virtual and Database Environments . Retina Network Security Scanner is the most sophisticated vulnerability assessment solution on the market . Retina Security Scanner enables you to

efficiently identify IT exposures and prioritize remediation enterprise-wide. Continually monitor and improve enterprise security posture.

Table 1. Comparison between penetration tools

Properties	Ping sweep	Port detective scanner	Advanced Port scanner	Port scanner active X control	Port scanner active X control	Port scanner active X control
Definition	Determines live host in range of IP address	Determines the status of the ports(in use or blocked)	Discovers open ports quickly and gets the version of program running on that port	Determines which ports are opened by remote hosts	Determines which ports are opened by remote hosts	Determines which ports are opened by remote hosts
Developer	DIGIT BLAST	TZOLKIN CORPORATION	FAMATECH	MAGNETO SOFTWARE	MAGNETO SOFTWARE	MAGNETO SOFTWARE
Version	Version is 1.2.0	Version is 2.0	Version is 2.4	Version is 5.0.0.3	Version is 5.0.0.3	Version is 5.0.0.3
File Size	506 KB	615KB	675 KB	3.08 MB	105.69 KB	3.93 MB
Operating system	All window based	NMAP for Unix system, SOLARWINDS for windows	Window 32-bit and 64-bit	WINDOW XP/VISTA/ SERVER 2008/7/8	IOS,IPHONE ,IPAD TOUCH	NT/2K/XP/95
License model	Free of charge	Free of charge	Free of charge	PURCHASE \$300	PURCHASE	PURCHASE \$ 1.00
Pros	Easy to setup and maintain, Cheaper in rate	Small ,fast ,robust ,easy to use	User friendly interface, Rich functionality	Lightweight, Intelligent, Powerful port scanner control	User friendly interface, Rich functionality	Fast, accurate ,non-intrusive scanning, Risk management

### III. KALI LINUX

Kali Linux is the best platform to perform kali linux. Kali linux is open source downloadable linux distribution .Kali linux is debian based linux distribution and can run on many platforms. We have examined primarily inbuilt tools of kali linux which are used for penetration testing. The main focus of kali linux is advanced penetration testing and security auditing. Kali Linux is available in 32 bit, 64 bit, ARM, Live USB, and VMware versions, is maintained and funded by Offensive Security Ltd.(0) maintained by offensive security.Kali linux supports a wide variety of wireless devices. It is free of charge known as open source and can be downloaded easily. It contains more than 600 inbuilt tools and it is possible to add numerous tools to it. Kali linux has multi-language support.Kali is actually installed in the virtual machine as environment to find vulnerabilities in the system that one wants to analyze.

### IV. INBUILT TOOLS OF KALI LINUX FOR PEN-TESTING

Kali Linux suite has numerous inbuilt tools for Penetration testing. Kali Linux is a suite of tools worked to accumulate data what's more, endeavor shortcomings, however the consistent basic leadership and investigation is yours. Outside of the specialized parts of assaulting, being quiet and composed will help you more than anything. Further, dependably ensure you have direct consent or responsibility for destinations required in your infiltration testing. When you have constrained your danger to undue outside impacts, it is time to start stage one of the infiltration test.

The penetration testing tools are divided into five phases such as :-

Table 2. Tools of penetration testing

Information gathering phase	Vulnerability detection	Penetration tests	Exploitation	Reporting
Nmap	Nessus	OWASP-ZAP	Armitage	CaseFile
Maltego	BBQSQL	Bluelog	Backdoor Factory	CutyCapt
Dmitry	CISCO OCS	Blueranger	BeEF	dos2unix
Dnsenum	CISCO TORCH	Aircrack-ng	cisco-auditing-tool	Dradis
Dnstracer	GSD	bluesnarfer	cisco-global-exploiter	KeepNote
Acccheck	jSQL	Crackle	Commix	MagicTree
Intrace	Lynis	Gr-scan	Crackle	Metagoofil
Goofile	Openvas-manager	Mdk3	SET	Nipper-ng

Some of these tools are discussed below:

*Phase1: Information Gathering:-*

Information gathering is first step of penetration testing. In this step we gather all the Information, when we target the victim for pen-testing. Information Gathering starts with surfing of internet using search engines such as Google, on-line news sources, business posjtions, and numerous other on-line assets. The end goal of phase one is to have a logical map of the target's network.

- Nmap:

Nmap is the Network map which defines the way to discover the live hosts. In this, firstly by using the scanner we gathers the IP of victim or the IP address of machine which is connected to same Wi-Fi. In Nmap we paste the IP address of victim machine then we go for intense scan. It gives a lot of information quickly in one scan. The output of Nmap is which ports are open. The ports which are open are very much vulnerable.

Nmap also determines the operating system of target machine. Nmap maintains a type of database and will match the responses to make a guess at what type of operating system the target computer is running. This OS detection isn't perfectly accurate but it can help the attacker tailor his attack strategy, especially when coupled with other pieces of information. Nmap gather information such as:-

- which ports are open
- what services those ports are offering
- which operating system of target machine
- what type of packets filters/firewalls in use

- Maltego:-

Maltego is an incredible built-in tool developed by Paterva company which is located in south Africa. Maltego is free intelligent tool and is used for forensic applications. Maltego is GUI based tools so its appearance is different. This tool gathers information and uses GUI as a interface. This tools represents the gathered information in easy and understandable form. Maltego is an information gathering tool that allows you to visually see relationships.

Maltego allows the user to gather information like:-

- Domain Names
- Net blocks
- IP Addresses
- URLs, Websites
- Personal information like Email address
- Phone numbers

*Phase 2 :-Vulnerability detection:-*

This phase helps in finding the Vulnerabilities that lie in the target system. In this phase local network is checked for its vulnerabilities.

- Nessus:-

Nessus is the Vulnerability scanner that finds vulnerabilities in target structure and endeavor to modify them. Neesus makes summon of the accompanying step, finding vulnerabilities in the area structure, in the close-by framework, and in both Linux and Windows site.

While checking a framework for vulnerabilities, Neesus is as comprehensive as gadgets come. Regardless of the way that Neesus wears down Kali Linux, it is not bundled with the download, and ought to be downloaded and unpackaged on the Kali Linux OS. Enlistment through the Neesus site is moreover required to run this gadget. Nessus consolidates port looking at and OS ID, so every so often a feebleness evaluation will simply utilize. Nessus and let Nessus call nmap or distinctive scanners for these sections of the test. For a stealthy yield, a security capable or an aggressor may run these mechanical assemblies freely to keep up a key separation from acknowledgment.

*Phase 3:-Penetration Tests*

At this phase, penetration testers will take the information gathered and list of vulnerabilities collected from phase one and two. In a team of assaultants, this is the perfect time for a brief pause and gathering of the troops. Upto this point most of the tools used were relatively quiet and noninvasive. If the attacking team is properly prepared, choosing which attack vector to hit is the next key step.

- OWASP-ZAP:-

The Open Web Application Security Project is well known open source tool. ZAP is the “Zed Attack Proxy Project”. ZAP 2.4.3 has been released, this is a bug fix and enhancement release.(0)The tool is simple enough for new penetration testers, and robust enough for professional environments.(0)This tool is ideal for the new testers and are specially for webapp pentesting. It also cross platform so can be used with windows, linux and macs. It is an intercepting proxy so the user can typically configure the browser to proxy through ZAP so ZAP can see all the request and responses. It also contain both active and passive scanner. Passive scanner run all the time and has report of all the request and response. Passive Scanner is safe to use on any website, it do perform any attack. Active scanner can attack so it can only be used with only that applications for which we have permission to test. It has support of Smart card and Client Digital Certificates.

- Wifi Attacking :- Aircrack-ng

Aircrack-ng is an important tool for infusing remote packets into a dynamic system. This instrument depends on the assailants learning of remote cards, both on the assaulting machine what's more, on the objective machine, so before conveying. Aircrack-ng in your hostile environment, make sure you have the essential data assembled from stage one. When dynamic, Aircrack-ng can likewise recuperate 802.11 WEP and WPA-PSK keys by gathering parcels sniffed remotely. WEP assaults have been surely understood and well reported in the security group subsequent to no less than 2007, but since of the nature of arranged correspondence, infusion assaults are still an exceptionally well known strategy for accessing a network.

*Phase 4:-Exploitation*

- Metasploit:-

The Metasploit Framework gone through the Metasploit Framework Console is among the most progressive devices in the Kali Linux arms stockpile. The Metasploit group is fabulous, also, their work in the hostile info-sec field is without parallel. Kali Linux itself was in light of building up an OS that joined every one of the apparatuses of Metasploit and Backtrack together. Metasploit itself could be viewed as an all-in-one infiltration testing device, also, for some regardless it is. Of the considerable number of apparatuses in this rundown, just Burp Suite approaches in power and clean that Metasploit offers, and the Burp Suite instruments are a far off second when contrasted with the profundity of Metasploit's toolbox. Metasploit offers tools can be utilized as a part of each period of an infiltration test, from uninvolved data gathering devices to weakness filters.

*Phase5:-Reporting*

- RecordMyDesktop

While working with all the above tools, we leap over the line from safe to illegal and work directly with tools that could easily break a business. The point of a penetration test is to attack an environment in a controlled way so the defenders can have accurate and honest information on their weaknesses. Offensive security is a defensive tool. As flashy as exploits may be, everything in your offensive arsenal comes down to a simulated attack. Wargaming is only as good as the lessons learned at the end. RecordMyDesktop is the least technical tool on this list, but in my opinion, the most important. Showing exactly how an exploit worked, and having a clear and objective record of the attack taking place will be essential for the analysis and cleanup stages after the penetration test has completed.

#### CONCLUSION

Penetration Testing is an important topic that an IT companies should be aware of it. As the need of internet is growing very quickly so the user should realize that the security is very essential for them. Today cybercrime has increased a lot so security is the crucial part. Any attacker or intruder can hack the user's machine or passwords any time, in that situation penetration testing can be helpful. The wide range of penetration testing tools helps in so many fields, as they are free and unlimited production. In the future penetration testing tools can help in hacking the satellites and change the prediction about weather patterns. There are several tools in Kali Linux suite, but in this paper the researcher has worked on some of the tools of each phase. In the future work can be done on other tools like Blue Range for hacking Bluetooth, WPA and WPA2 for hacking wi-fi passwords and can achieve further improvements in security issues.

#### REFERENCE

- [1] Bingchang Liu, Liang Shi, Zhuhua Cai, Min Li, "Software Vulnerability Discovery Techniques: A Survey", 2012 Fourth International Conference on Multimedia Information Networking and Security, 978-0-7695-4852-4/12
- [2] Omar H. Alhazmi, Yashwant K. Malaiya. "Modeling the vulnerability discovery process", Proceedings of the 16th IEEE International Symposium on Software Reliability Engineering (ISSRE'05). Washington, DC, USA: IEEE Computer Society, 2005

- [3] Sumit Goswami, Sudip Misra, Mukesh Mukesh, "A replay attack resilient system for PKI based authentication in challenge-response mode for online application" 3rd International Conference on Eco-friendly Computing and Communication Systems, 978-1-4799-7002-5/14
- [4] A.Ojha, V.Belwel,R.C.Agarwal,G.,"Password based authentication: Philosophical survey," Intelligent Computing and Intelligent Systems(ICIS), 2010 IEEE International Conference on, vol.3, no., pp.619,622, 29-31 Oct. 2010.
- [5] Jia-Lun Tsai; Nai-Wei Lo; Tzong-Chen Wu, "Novel Anonymous Authentication Scheme Using Smart Cards," Industrial Informatics, IEEE Transactions on , vol.9, no.4, pp.2004,2013, Nov. 2013.
- [6] Sumit Goswami, Nabanita R Krishnan, Mukesh,, Saurabh Swarnkar, Pallavi Mahajan, Reducing attack surface of a Web Application by OWASP Compliance, Defence Science Journal Vol. 62, No. 5, pp 324-330, September 2012
- [7] R. Shirey. Internet Security Glossary. RETF RFC2828, 2002
- [8] Viet Hung Nguyen, Fabio Massacci. An Idea of an Independent Validation of Vulnerability Discovery Models, Engineering Secure Software and Systems 4th International Symposium, ESSoS 2012, Eindhoven
- [9] Sung Whan Woo, Hyun Chul Joh, Omar H. Alhazmi, Yashwant K.Malaiya. Modeling vulnerability discovery process in Apache and IIS HTTP servers. Computers & Security, Volume 30, Issue 1, 2011
- [10] HERBERT H. THOMPSON. Application Penetration Testing, IEEE SECURITY & PRIVACY, 2005
- [11] Hayajneh, T.; Krishnamurthy, P.; Tipper, D.; Le, A. Secure neighborhood creation in wireless ad hoc networks using hop count discrepancies. Mobile Netw. Appl. 2012, 17, 415–430.
- [11] Elizabeth Fong Romain Gaucher Vadim Okun Paul E. Black :National Institute of Standards and Technology “Building a Test Suite for Web Application Scanners”. Proceedings of the 41st Hawaii International Conference on System Sciences – 2008

# DIGITAL INDIA: THE ROADMAP AHEAD

**Gugneet Kaur**

M.Tech (Student)

Punjabi University, Patiala

Punjab, India

**Dr. Jaswinder Singh**

Assistant professor

Punjabi University, Patiala

Punjab, India

**Abstract:** This paper focuses on the initiative taken by Indian Government that is Digital India and it deals with the technologies involved in the project and concentrates on looking into the design and key innovation of Internet of Things. Besides, the uses of Internet of Things are additionally portrayed in this paper. A great part of the IoT activity is bolstered by the abilities of assembling minimal effort and vitality productive equipment for gadgets with correspondence limits, the development of remote sensor system innovations, and the interests in incorporating the physical and digital universes.

**Keywords:** IoT, Bluemix , DevOps , IDE,API, SaaS, PaaS.

## INTRODUCTION

*DIGITAL INDIA:* This is an activity taken by Indian government to convey administrations to natives electronically by enhancing online foundation and by expanding Internet network. It was propelled on 1 September 2015 by Prime Minister Narendra Modi. This activity incorporates arrangements to associate country ranges with fast web systems [1].

Parts:-

1. The formation of advanced framework.
2. Conveying administrations digitally.
3. Advanced education.

The plan will be checked and controlled by the Digital India Advisory gathering which will be taken care of by the Ministry of Communications and IT.

## ABOUT THE PROJECT

Broadband in 2.5 lakh towns, widespread telephone availability, Net Zero Imports by 2020, 400,000 Public Internet Access Points . Wi-fi in 2.5 lakh schools, all universities ,Public wi-fi hotspots for citizens. Digital Inclusion: 1.7 Cr prepared for IT, Telecom and Electronics Jobs Job creation.

## NINE PILLARS OF DIGITAL INDIA PROGRAMME

1. Broadband Highways.
2. All inclusive Access to Mobile Connectivity.
3. Open Internet Access Program.
4. e-Governance – Reforming Government through Technology.
5. eKranti-Electronic conveyance of administrations
6. Data for All.
7. Hardware Manufacturing.
8. IT for Jobs.
9. Early Harvest Programs.

## TECHNOLOGIES:

The accompanying rundown of 12 enabling advancements for computerized India fall into three sorts:

- Technologies that "digitize" life and work,
- Smart physical systems ,and
- Technologies for reevaluating vitality.

## DIGITISING LIFE AND WORK:

Digitising life and work includes:

1. Mobile Internet.
2. Cloud technology.
3. Automation of knowledge work.
4. Digital payments.

## SMART PHYSICAL SYSTEMS:

- Smart physical systems include following technologies:
- Internet of things.
- Intelligent transportation and distribution.
- Advanced geographic information systems (GIS).
- Next-generation genomics.

## RETHINKING ENERGY

Rethinking energy includes following technologies:

1. Advanced oil and gas exploration and recovery.
2. Renewable energy.
3. Advanced energy storage.

## INTERNET OF THINGS

Broadening the present Internet and giving association, correspondence, and between systems administration amongst gadgets and physical articles, or "Things," is regularly alluded to as the Internet of Things [2]. "The innovations that empower mix of genuine information and administrations into the present data organizing advancements are regularly portrayed under the term of the Internet of Things (IoT)".

## APPLICATIONS OF CLOUD COMPUTING

*Test and improvement:* Probably the best situation for the utilization of a cloud is a test and advancement environment. This involves securing a financial plan, setting up your surroundings through physical resources, huge labor and time. At that point comes the establishment and arrangement of your stage. This can frequently amplify the time it takes for a venture to be finished and extend your milestones. With distributed computing, there are presently promptly accessible situations custom fitted for your necessities readily available. This frequently consolidates, yet is not constrained to, mechanized provisioning of physical and virtualized assets.

*File storage:* Cloud can offer you the likelihood of putting away your records and getting to, putting away and recovering them from any web-empowered interface. In this situation, associations are paying for the measure of capacity they are really expending, and do as such without the stresses of directing the everyday upkeep of the capacity framework. There is additionally the likelihood to store the information either on or off premises relying upon the administrative consistence prerequisites. Information is put away in virtualized pools of capacity facilitated by a third get-together taking into account the client particular prerequisites.

*Disaster recovery:* This is yet another advantage got from utilizing cloud taking into account the cost adequacy of a debacle recuperation (DR) arrangement that accommodates a quicker recuperation from a cross section of various physical areas at a much lower cost than the conventional DR site with altered resources, unbending systems and a much higher expense.

*Backup :* Backing up information has dependably been a complex and tedious operation. This included keeping up an arrangement of tapes or drives, physically gathering them and dispatching them to a reinforcement office with all the inalienable issues that may happen in the middle of the starting and the reinforcement site. Along these lines of guaranteeing a reinforcement is performed is not safe to issues, for example, coming up short on reinforcement media , and there is likewise time to stack the reinforcement gadgets for a reestablish operation, which requires some serious energy and is inclined to breakdowns and human blunders.

## Applications Of RFID:-

*Race timing :* Timing marathons and races are a standout amongst the most prominent employments of RFID, yet frequently race members never acknowledge they're being planned utilizing RFID innovation, and that is a demonstration of RFID's capacity to give a consistent purchaser experience.

*Attendee Tracking* : If you've ever dealt with a substantial gathering some time recently, you'll realize that it's vital to keep the stream of activity moving at a consistent pace, particularly all through workshops. With a RFID participant arrangement, kill the requirement for enrollment lines at doors.

*Materials management*: In development and other related businesses, materials are frequently the biggest venture consumption. On huge occupation locales, essentially discovering materials can be dangerous. RFID arrangements like Jovix remove the mystery from the condition.

*Library Systems*: An RFID library arrangement enhances the productivity of flow operations. While scanner tags require observable pathway, RFID labels can be perused from numerous points which implies the checkout and registration procedure is fundamentally speedier.

*IT Asset Tracking*: IT resources, for example, server cutting edges, portable PCs, tablets, and different peripherals are excessive speculations for any organization, also that data put away on those things could demonstrate negative in the wrong hands. IT resource labels give your IT group the capacity to rapidly do a stock number and ensure everything is set up.

## RENEWABLE ENERGY IN INDIA

Renewable vitality in India: goes under the domain of the Ministry of New and Renewable Energy. India was the principal nation on the planet to set up a service of non-routine vitality assets, in mid 1980s[3]. India's aggregate network intelligent or framework tied renewable vitality limit (barring extensive hydro) has achieved 33.8 GW, of which 66% originates from wind, while sun powered PV contributed about 4.59% alongside biomass and little hydro force of the renewable vitality introduced limit in India.

## WIND POWER

Wind power represents 6% of India's aggregate introduced power limit, and it produces 1.6% of the nation's power. Indian Wind Energy Alliance (IWEA) is the zenith body for the wind vitality industry in India. It was dispatched in December 2014 and Mr. Sumant Sinha is the director of IWEA.

## SUN BASED POWER

One of the main utilizations of sun powered force has been for water pumping, to start supplanting India's four to five million diesel fueled water pumps, each devouring around 3.5 kilowatts, and off-network lighting. Some substantial tasks have been proposed, and a 35,000 km<sup>2</sup> range of the Thar Desert has been put aside for sunlight based force ventures, adequate to produce 700 to 2,100 gigawatts. Indian Electrical and Electronics Manufacturers Association (IEEMA) Plays a noteworthy part in Renewable Energy.

## WASTE TO ENERGY

Consistently, around 55 million tons of civil strong waste (MSW) and 38 billion liters of sewage are produced in the urban zones of India. What's more, huge amounts of strong and fluid squanders are created by industries.

Prominent organizations in the Waste to Energy area:

1. A2Z Group of organizations
2. Hanjer Biotech Energies
3. Ramky Enviro Engineers Ltd
4. Hitachi Zosen India Pvt Limited
5. Clarke Energy

## GIS APPLICATIONS

GIS are PC programming and equipment frameworks that empower clients to catch ,store, dissect and oversee spatially referenced information.

Examples include:

- Crime mapping.
- Remote detecting applications.
- Road organizing.
- Waste administration.
- Wastewater and storm water frameworks.

## INTERNET OF THINGS

The Internet of Things (IoT) is the system of physical articles—gadgets, vehicles, structures and different things inserted with hardware , programming , sensors, and system availability that empowers these items to gather and trade data. The IoT permits articles to be detected and controlled remotely crosswise over existing system infrastructure."Things," in the IoT sense, can allude to a wide assortment of gadgets, for example, heart checking inserts, biochip transponders on homestead creatures, electric shellfishes in waterfront waters, automobiles with implicit sensors [4]. Integration with the Internet suggests that gadgets will utilize an IP address as a one of a kind identifier. Objects in the IoT will need to utilize IPv6 to suit the greatly extensive location space required. Objects in the IoT won't just be gadgets with tangible capacities, additionally give activation abilities (e.g., globules or locks controlled over the Internet). IoT frameworks could likewise be in charge of performing activities, not simply detecting things. Canny shopping frameworks, for instance, could screen particular clients' buying propensities in a store by following their particular cellular telephones. "web of living things" has been proposed to portray systems of natural sensors that could utilize cloud-based investigations to permit clients to study DNA or different particles [15].

## APPLICATIONS

### ENVIRONMENTAL MONITORING

IoT regularly utilize sensors to help with ecological insurance by observing air or water quality , air or soil conditions. Different applications like quake or torrent early-cautioning frameworks can likewise be utilized by crisis administrations to give more powerful guide.

#### ENERGY MANAGEMENT

Reconciliation of detecting and incitation frameworks, associated with the Internet, is prone to enhance vitality utilization as a whole. It is normal that IoT gadgets will be coordinated into all types of vitality devouring gadgets (switches, electrical plugs, globules, TVs, and so forth.) and have the capacity to speak with the utility supply organization so as to successfully adjust power era and vitality usage [14]. Such gadgets would likewise offer the open door for clients to remotely control their gadgets, or midway oversee them through a cloud based interface, and empower propelled capacities like booking (e.g., remotely driving on or off warming frameworks, controlling broilers, changing lighting conditions and so forth.).

#### MEDICAL AND HEALTHCARE SYSTEM

IoT gadgets can be utilized to empower remote wellbeing observing and crisis notice frameworks. These wellbeing observing gadgets can go from circulatory strain and heart rate screens to cutting edge gadgets fit for checking specific inserts, for example, pacemakers, Fitbit electronic wristbands or propelled listening to aids. Specialized sensors can likewise be prepared inside living spaces to screen the wellbeing and general prosperity of senior subjects [12].

#### BUILDING AND HOME AUTOMATION

IoT gadgets can be utilized to screen and control the mechanical, electrical and electronic frameworks utilized as a part of different sorts of structures (e.g., open and private, modern, organizations, or private) in home computerization and building robotization frameworks.

#### TRANSPORTATION

The IoT can help with reconciliation of interchanges, control, and data handling crosswise over different transportation frameworks. Utilization of the IoT reaches out to all parts of transportation frameworks (i.e. the vehicle, the base, and the driver or client) [11]. Dynamic association between these parts of a vehicle framework empowers bury and intra vehicular correspondence, savvy activity control, brilliant stopping, electronic toll gathering frameworks, logistic and armada administration, vehicle control, and wellbeing and street help.

#### ARCHITECTURE

Internet of Things might be a non-deterministic and open system in which auto-composed or canny elements (Web administrations), virtual articles (symbols) will be interoperable and ready to act autonomously contingent upon the connection, circumstances or environments [5]. The framework will probably be a case of occasion driven engineering, base up made (taking into account the setting of procedures and operations, progressively) and will consider any backup level.

## NETWORK ARCHITECTURE

Internet of Things requires gigantic adaptability in the system space to handle the surge of gadgets. IETF 6LoWPAN would be utilized to interface gadgets to IP systems. With billions of gadgets being added to the web space, IPv6 will assume a noteworthy part in taking care of the system layer adaptability. IETF's Constrained Application Protocol, MQTT and ZeroMQ would give lightweight information transport [6]. Fog processing is a reasonable contrasting option to anticipate such extensive burst of information course through Internet. The edge gadgets calculation force can be utilized to break down and handle information, in this way giving simple continuous adaptability.

## ENABLING TECHNOLOGIES FOR IoT

There are many technologies that enable IOT :

1. RFID and close field correspondence – In the 2000s, RFID was the prevailing innovation. Later, NFC got to be predominant (NFC). NFC have ended up regular in cell phones amid the mid 2010s, with utilizations, for example, perusing NFC labels or for access to open transportation[10].
2. Optical labels and brisk reaction codes – This is utilized for minimal effort labeling. Telephone cameras translate QR code utilizing picture handling procedures. In actuality QR notice crusades gives less turnout as clients need another application to peruse QR codes.
3. Bluetooth low vitality – This is one of the most recent tech. All recently discharging cell phones have BLE equipment in them. Labels in view of BLE can flag their nearness at a force spending that empowers them to work for up to one year on a lithium coin cell battery[9].
4. Low vitality remote IP systems – inserted radio in framework on-a-chip plans, lower power WiFi, sub-GHz radio in an ISM band, regularly utilizing a packed variant of IPv6 called 6LowPAN.
5. ZigBee – This correspondence innovation depends on the IEEE 802.15.4 convention to execute physical and MAC layer for low-rate remote Private Area Networks[7]. Some of its principle attributes like low power utilization, low information rate, minimal effort, and high message throughput make it a fascinating IoT empowering influence innovation.
6. Z-Wave – is a correspondence convention that is for the most part utilized as a part of savvy home applications.
7. LTE-Advanced – LTE-A will be a rapid correspondence detail for portable systems. Contrasted with its unique LTE, LTE-A has been enhanced to have expanded scope, higher throughput and lower inertness. One vital utilization of this innovation is Vehicle-to-Vehicle (V2V) interchanges [13].
8. WiFi-Direct – It is essentially WiFi for peer-to-peer communication without needing to have an access point. This feature attracts IoT applications to be built on top of WiFi-Direct to get benefit from the speed of WiFi while they experience lower latency.

## SIMULATION

IOT displaying and reenactment (and copying) is regularly done at the configuration stage before sending of the system. System test systems like OPNET, NetSim and NS2 can be utilized to recreate IOT systems[8].

## FEEDBACK AND DEBATES

While numerous technologists tout the Internet of Things as a stage towards a superior world, researchers and social spectators have questions about the guarantees of the pervasive processing transformation.

## CONCLUSION

The Internet of Things will change the services of general public, and will bring consistence 'at whatever time, anywhere' business, diversion and informal communication over quick solid and secure systems. This implies the end of the separation between advanced, virtual and physical universes.

In this new connection specialized design that locations purpose of control issues for following objects crosswise over systems, organizations and business procedures will get to be fundamental. Straightforward, circulated, models to empower numerous identifier powers supporting the interests of neighbourhood, local and national business, policymakers, and people and addresses concerns over security and protection will be the answer.

Today, the web administrations are standard and generally embraced innovation for the Internet. The future Internet will bring new difficulties where the remote identifiable installed frameworks at the edge of the system need and use comparative functionalities. Remote identifiable gadgets and inserted dispersed frameworks will execute administration arranged structures to explain the multifaceted nature of conveyed implanted applications and spread web administrations as a cross area innovation.

Remote sensor systems and universal systems, where the sensors will be associated with also, controlled by installed frameworks will utilize this methodology, where administrations epitomize the usefulness and give brought together access to the usefulness of the framework.

## REFERENCES

- [1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. ACM Trans. Program. Lang. Syst. 15, 5 (Nov. 1993), 795-825.
- [2] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.
- [3] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (The Hague, The Netherlands, April 01 - 06, 2000). CHI '00. ACM Press, New York, NY, 526-531.
- [4] Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.
- [5] Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis . UMI Order Number: UMI Order No. GAX95-09398., University of Washington .
- [6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification . J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.
- [7] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In Proceedings of the 16th Annual ACM Symposium on User interface Software and Technology (Vancouver, Canada, November 02 - 05, 2003). UIST '03. ACM Press, New York, NY, 1-10.

- [8] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press .
- [9] Spector, A. Z. 1989. Achieving application requirements . In Distributed Systems, S. Mullender, Ed. Acm Press Frontier Series. ACM Press, New York, NY, 19-33.
- [10] IBM "Smart China" report, 2009
- [11] ITU Internet report, 2005
- [12] INFSO 0.4 NETWORKED ENTERPRISE&RFID, INFSO G.2 MICRO&NANOSYSTEMS, WORKING GROUP RFID OF THE ETP EPOSS, "Internet of Things in 2020 -- Roadmap for the future", 2008.
- [13] Yuh-Jzer Joung , "RFID and the Internet of Things" ,Taiwan University, 2007
- [14] Christine Legner, Frederic Thiesse, "RFID-Based Maintenance at Frankfurt Airport" ,IEEE Distributed Systems
- [15] Margery Conner," Sensors empower the ' Internet of Things'", 2010

# A REVIEW ON CONCEPTUAL DATA MINING

**Ramandeep Kaur**

Student, Department of Computer Engineering  
Punjabi University, Patiala  
ramandeep.rapal@gmail.com

**Gaurav Gupta**

Department of Computer Engineering  
Punjabi University, Patiala  
gaurav.shakti@gmail.com

**Abstract-**In the field of research area, industry and media both methodologies that are data mining from knowledge discovery from database get more attention and attracted significantly. This paper provides an overview that how these two methodologies that data mining and knowledge discovery from database related to each other in the field of machine learning, statistics and databases.

**Keywords-***data mining, applications of data mining, knowledge discovery of data, components of data mining, tasks of data mining and association rule mining.*

## I. INTRODUCTION

In Today's scenario human beings are used different technologies to adequate in the society. In each and every field human being are working with huge amount of data and these data are in different fields that may be in the form of audio, video, graphical format documents records etc. [1] In day to day life, we see that there is huge amount of data available in the field of education medical, industry and many other areas. This type of data may provide information for making decisions. For example, with the help of this kind of data you can discover drop out student in any university, sales data in any shopping database. Now here, you need to analyse summarised, understand existing data to meet the required challenges. [2] For this purpose we need the concept of data mining. The process of data mining can be defined as to extract previously unknown or hidden information from existing data. The process of data mining is the extraction of hidden predictive information from large databases. It is very high performance technology that can help organization to focus on the most important information in their data warehouses. It is a knowledge driven decision process. [3] Data mining is a subfield of computer science with large data patterns. Data mining is analysis of data into meaningful rules. The aim of analysis process is to extract data from data set. Data mining is used for analysing data and solves the problem in present database. It is an arranging technique which is helpful to solve the hidden patterns from database. [4]

## II. APPLICATIONS OF DATA MINING

*1. Financial Data Analysis:-*In banking and financial industries generally data is of high quality that required systematic data analysis by following data mining process.

2. *Retail Industry*:-In retail industry data mining process is very helpful to identify buying patterns and trends of the customer that leave to improve customer service quality and also there satisfaction.

3. *Telecommunication*:-In area of telecommunicate industry data mining provide help to identify telecommunication patterns to detect fraudulent activities, make better use of sources and also improve service quality.

4. *Scientific Applications*:-In this area huge amount of data get generated in various fields such as echo system modelling, satellite etc. Data get fastest numerical computation that is done with the help of data mining software. [5]

### III. KNOWLEDGE DISCOVERY FROM DATABASE

Sometimes data mining treated as knowledge discovery from database that is KDD. Data mining is a concept of extracted large amount of data that is available, and then convert it into useful information or knowledge that spot decision making process. We organise new information in order to get decision making process. Fig.1 shows knowledge discovery from database process consists of various emerged step:

1. *Selection*:-In this step selection of data from multiple resources such as spread sheets, browser and graphic document etc.

2. *Data Cleaning*:-Data cleaning is used to remove redundant data such as noise and bugs.

3. *Transformation*:-Transformation means change the existing data into new formats in which they are compatible to use by user.

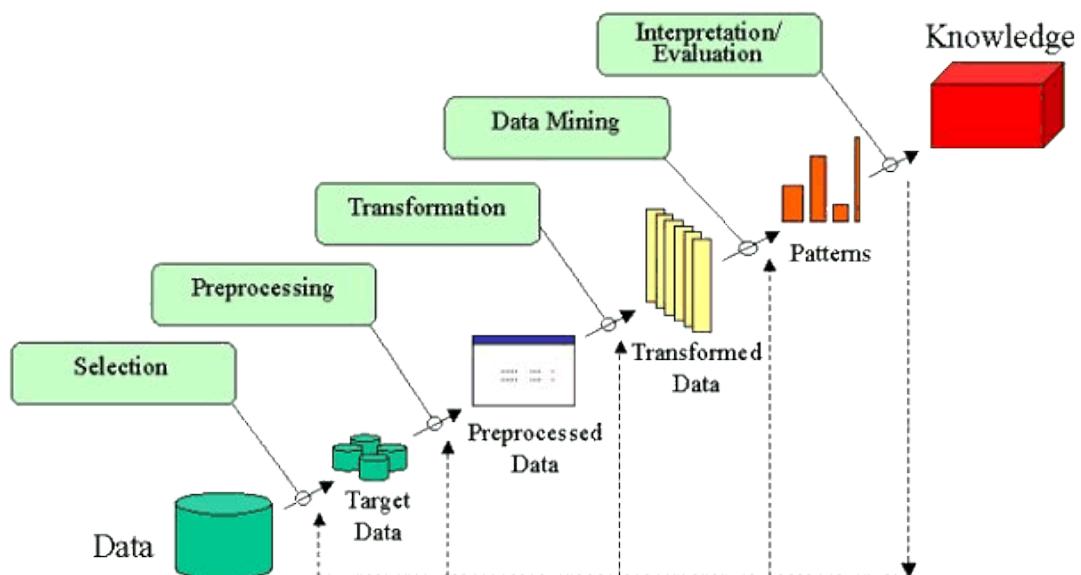


Fig.1 Knowledge Discovery from Database

4. *Data Mining*:-Data mining is used for achieving results related to data.

5. *Interpretation*:-Interpretation is also called evaluation. It is used for predicting the results for the meaningful information and data. [2]

#### IV. COMPONENTS OF DATA MINING SYSTEM

Data mining system consists of various components that fig.2 follow the sequence one after another given as follow:

1. *Database*:-Here data from multiple sources get combined such as from web sites, data warehouses and spread sheets etc.
2. *Integration and Selection*:-Here user required data get aggregated into a single warehouse and extract only those data items that come under user area of interest.

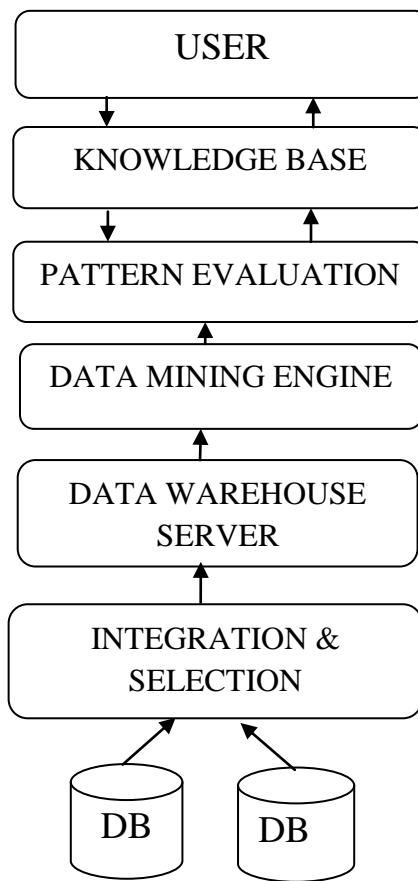


Fig.2 Components of Data Mining System

3. *Data Warehouse Server*:-Data warehouse server is responsible for fetching the required data as per the user requirement request.

4. *Data Mining Engine*:-Here set of operations of different techniques such as association, classification, clustering and segmentation are applied on the fetched data according to the requirement to the output for final user.
5. *Pattern Evaluation*:-In the pattern evaluation module those patterns are analysed that have similar type of data can be combined under one group.
6. *Knowledge Base*:-Knowledge base module is used to address the search to evaluate the interestingness of resulting patterns.
7. *Users*:-This module provides user interface between the user and data mining system software that helps user to interact with data mining tools and tasks by using data mining queries. [6]

#### V. DATA MINING MODELS

Data mining is extraction of compatible data from unarranged data. So it provides help to achieve the specific goal. Data mining mainly performs two basic tasks that are:

1. *Descriptive Modelling*:-It is used to describe all the features of database at current level that means all the operations performed by that database system.
2. *Predictive Modelling*:-It is used to predict future actions or objectives of any known or unknown class. Basically this type of modelling uses historical data. [2]

#### VI. BASIC DATA MINING TASKS

This section gives the outline of various tasks performed by data mining system that are given as follows:

1. *Classification*:-It is classic data mining technique that is based on machine learning. Basically classification is used for classify the data item into the data set into pre-defined set of groups. Classification technique is using mathematical equations such as decision tree, neural network and statistics. In classification software can be developed how to learn classified data items into groups.
2. *Prediction*:-Prediction is basically used to predict future states any database depending upon the past and current data of any database system. It extracts complete valuable knowledge from data warehouse.
3. *Association*:-Association is also referred to as link analysis because it is used to find out link between two or more data items. Association is used to find out the relationship between items in the same transaction that is why association technique is called relation technique.
4. *Clustering*:-In clustering data alone is predefined but data groups are not predefined. Only those data items are defined in a cluster that has similar behaviour. It is an unsupervised technique because only after analysing data user can made a cluster of those data items that have similar type of behaviour. [7]

## VII. ASSOCIATION RULE MINING

In many application areas it is required to discover the frequent patterns, frequent patterns are those data items that frequently occur in the data set. Classical problem of association data mining is super market analysis. Super market contains huge amount of data of their customers and it is required to maintain the data in the well-mannered form. Association is a criterion for analysing patterns that frequently occur are defined by two parameters that are support and confidence to identify the relationship.

- Support indicates those data items that frequently occur in the database.
- Confidence indicates that how many times the given statement is found to be true.[8]

Very common approach to discover association rule is to break the problem into two parts:-

1. Find large item sets.
2. Generate rules for frequent item sets, an item set is a subset of set of all items says I. [7]

### *Example*

Association can be implemented in banking system. As so many customers have their accounts in different branches. Now, say one customer (he/she) wants to open their account in any branch than first of all he is required to give his identity proof in the bank. Then customer have to follow the complete procedure to open his account. Now, here customer also required to mention his nominee. Obviously, nominee is that person in his relationship that may have authorization to access his account in his absence.

## VIII. BASIC TOOLS USED IN DATA MINING

1. *WEKA*:-WEKA means Waikato Environment for Knowledge Analysis. It is a learning of algorithm through machine for achieving data mining tasks. Algorithm is directly applied to the dataset or your own java code. User can done classification, clustering, visualization, association with the help of this tool. WEKA is used in today's world for achieving new developments in various fields. [9]

2. *R-Programming*:-R-programming was developed by Ross Ihaka and Robert Gentleman that is why it is called as R-programming. R is a programming language that is graphic based on the R foundation that depends upon statistical computing. R programming is a package that is written in C, FORTRAN and R. [10]

3. *Rapid Miner*:-rapid miner is an open source system. It is a standalone application analysis of data and two products are integrated by data mining engine. This tool is written in java programming language. This tool is used for template related like frameworks. User can write any code that is offered as a service. [11]

4. *NLTK*:-NLTK means Natural Language Toolkit. It is a collection of dataset, libraries, and modules for natural language processing. It is written in python programming language. In advanced projects in data mining this tool provide flexible framework. [12]

#### IX. FUTURE SCOPE AND CONCLUSION

In today's scenario, so many problems has occurred in the field of databases and get solved by using data mining, mostly the problems of collecting the data from multiple source for an organization. As data mining provides various techniques such as association, clustering, classification and time series etc. to extract the knowledge from the data stored at data warehouse. In future work, we will review the association rule mining algorithms and their significance to follow the approach of designing more efficient association rule mining algorithms.

#### REFERENCES

- [1] Neelamadhab Padhy, Dr. Pragnyaban Mishra, and Rasmita Panigrahi. (2012). The Survey of Data Mining Applications and Feature Scope. *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*.
- [2] Smita,Priti Sharma. (2014). Use of Data Mining in Various Field:A Survey Paper. *IOSR-Journal of Computer Engineering(IOSR-JCE)*.
- [3] Aarti Sharma,Rahul Sharma,Vivek Kr.Sharma,Vishal Shrivastava. (2014). Application of Data Mining:A Survey Paper. *(IJCSIT) International Journal of Computer Science and Information Technologies*.
- [4] Nikita Jain, Vishal Srivastava. (2013). DATA MINING TECHNIQUES:A SURVEY PAPER. *IJRET:International Journal of Research in Engineering and Technology* .
- [5] (<http://www.tutorialspoint.com/>)
- [6] Jiawei Han and Micheline Kamber. (2006). *Data Mining: Concepts and Techniques*. Diane Cerra.
- [7] Dunham, M. H. (2003). *Data Mining:Introductory and Advanced Topics*. Pearson Education.
- [8] (2010-2016). Retrieved from searchbusinessanalytics.techtarget.com.
- [9] (<http://www.siliconafrica.com/the-best-data-minning-tools-you-can-use-for-free-in-your-company/>)
- [10] ([https://en.wikipedia.org/wiki/R\\_\(programming\\_language\)\)](https://en.wikipedia.org/wiki/R_(programming_language))
- [11] (<http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/>)
- [12] ([https://en.wikipedia.org/wiki/Natural\\_Language\\_Toolkit](https://en.wikipedia.org/wiki/Natural_Language_Toolkit))

# Big Data Analysis Using RHadoop

Authors

Sarpreet Kaur<sup>1</sup>, Sheena<sup>2</sup>, Dr. Williamjeet Singh<sup>3</sup>

<sup>1</sup>Student (M.Tech), Department of Computer Engineering, Punjabi University, Patiala

<sup>2</sup>Student (M.Tech), Department of Computer Engineering, Punjabi University, Patiala

<sup>2</sup>Assistant Professor, Department of Computer Engineering, Punjabi University, Patiala

Email:sarusandhu100@gmail.com, sheenarewari@gmail.com, williamjeet@gmail.com

## *Abstract*

**Big Data** is hastily exploding in all domains, including physical and biological. **Big Data Mining** is the real world data mining. Traditional databases systems are not able to store and process big data. With increase in the size of data, amount of irrelevant data also increases and the process of mining becomes inaccurate. **Big Data** technologies are important in providing precise analysis, which lead to firm decision-making. Various **Big Data** technologies are available in market from different vendors including Amazon, IBM, Microsoft and Apache etc. The huge size of analytics requires large computation which can be done with the help of distributed processing **HADOOP**. This paper describes big data and processing of big data using **RHADOOP**.

**Keywords**— Hadoop,, Hadoop Distributed File System (HDFS), HBASE, rhdfs, rrmr, rhbase, RHadoop

## 1. INTRODUCTION

Big Data describes tremendous volume of structured, unstructured and semi-structured data that is so large that it is difficult to process using traditional database and existing software techniques. Every digital process, social media exchange, sensors and mobile devices etc. generate Big Data. Mining of big data is termed as Big Data Mining. Big Data sets are those that surpass the simple type of data handling architectures and databases that were used in prior times. For example, sets of data that are too immense to be assuredly handled in Microsoft Excel spreadsheet could be referred to as big data sets. Big Data Mining refers to the task of proceeding through big sets for appropriate information. Big Data Mining is proficient example of the old axiom “looking for a needle in a haystack”[1].

IBM states that 80% of data collected today is unstructured. It contains posts of social media sites, pictures and videos, purchase and sale transaction records, stock exchange records, records from sensors used to gather climate information and cell phone GPS signals. All of this unstructured data is big data. Definition of Big Data is articulated as 3V's: Volume, Velocity and Variety [2]. The 3V's are defined as:

Volume: hundreds and thousands of terabytes and petabytes of information.

Velocity: Data is streaming at unprecedented speed.

Variety: Different types of data including text, audio, video, log files etc. which can be structured, semi-structured or unstructured.

Initially 3V's were given, now the number of V's is increased to 7. Other V's are:

Veracity: Analysis performed is useless if the data is not accurate.

Variability: Meaning of the same data could be different based on the context.

Visualization: Visual representation of analyzed data in a comprehensible way.

Value: Data itself is not valuable at all. The value is in the analysis done on that data and what information data provides. Big Data Mining requires the use of different kinds of software packages such as analytics tools [3]. Data

Mining implies to activities that include search operations and return intended results. Big Data is the resource and data mining is the executor that is used to provide profitable results [4]. Hadoop is the Big Data tool used to store and process large data sets. Hadoop is an open source software. Hadoop framework can store and process data on clusters of commodity hardware.

The objective of this paper is to focus on Hadoop, RTool and collaboration of R and Hadoop for Big Data analysis. The paper is organized as follows. Section 2 presents Hadoop and its architecture. Section 3 presents Big Data Analysis using R. Section 4 presents the need of Integration of R and Hadoop. Section 5 concludes the paper with summary.

## 2. HADOOP AND ITS ARCHITECTURE

Hadoop is a framework which deals with big data, it has its own family for processing big data, which is tied up in one umbrella called as Hadoop Ecosystem. Doug Cutting, Mike Cafarella and team took the solution provided by Google and started an open project called Hadoop in 2005 [5]. Hadoop is an open source, java based programming framework required for storing and processing enormous data in distributed computing environment. Hadoop provides substantial storage for any type of data (structured, semi-structured, unstructured), enormous processing power and the capability to handle limitless concurrent tasks or jobs. Figure 1 represents the architecture of Hadoop.

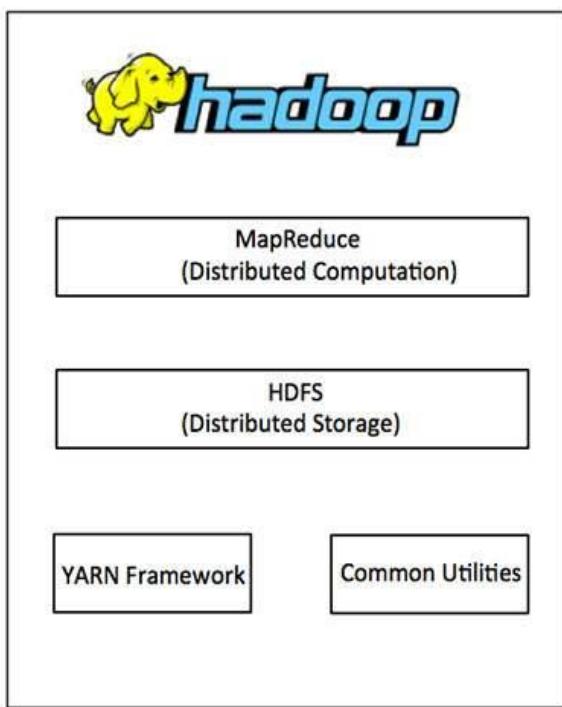


Figure1 Hadoop Architecture (Source: Apache Hadoop)

### 2.1 Hadoop MapReduce

It is the main component of Hadoop and method of programming in a distributed data stored in Hadoop Distributed File System(HDFS). MapReduce give its functionality Map that do mapping of logic into data and once computation is over reducer collects the result of Map to generate final output result of MapReduce[6]. MapReduce program can

be applied to any type of data whether structured, unstructured or semistructured stored in HDFS. The MapReduce refers to following two different tasks:

Map Task: This is the first task, which takes input data and converts it into a set of data, where individual elements are divided into tuples that is defined as (key/value pairs).

Reduce Task:

This task takes the output from map task as input. The reduce task is always accomplished after the map task. Figure 2 represents the MapReduce model. The benefit of MapReduce is that it becomes simple to extent data processing over multiple computing nodes. In the MapReduce model, data processing natives are known as mappers and reducers.

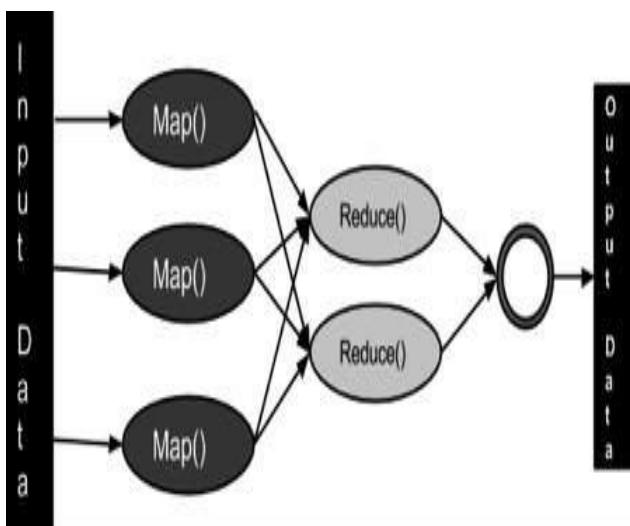


Figure 2. MapReduce Model (Source: Apache Hadoop)

## 2.2 Hadoop Distributed File System(HDFS)

HDFS is fundamental component of Hadoop and a technique to store data in distributed manner in order to process fast. HDFS saves data in block of 64MB (default) or 128MB in size which is logical division of data in a datanode. Datanode is the physical storage of data in Hadoop cluster. Metadata implies all information about data that are actually stored in datanodes and is encapsulated in namenode[7].

HDFS uses a master/slave architecture. Master has a single NameNode that undertakes the file system metadata and one or more slave DataNodes that store the actual data. A file stored in HDFS is divided into number of blocks and these blocks are saved in the set of DataNodes. The NameNode decides the mapping of blocks to the DataNodes. HDFS can retain enormous amount of data and provides assuredly access. Files are stored over multiple machines in order to handle immense data. The files are stored in redundant fashion to bail out the system from possible data losses in case of failure.

## 2.3 Hadoop YARN

Hadoop YARN is a framework responsible for job scheduling and cluster resource management. YARN is the architectural center of Hadoop. It allows number of data processing engines such as SQL, data science, real-time

streaming and batch processing to deal with the data stored on one platform. It unbolts a new approach to analytics.

#### **2.4 Hadoop Common Utilities**

These are java libraries and utilities required by other Hadoop modules. These libraries provides filesystem and Operating System(OS) level abstractions. Libraries contain the necessary java files and scripts required to start Hadoop.

### **3. BIG DATA ANALYSIS USING R**

R is an open source language and software environment that is used for statistical analysis, visualization of data, forecasting and time series analysis. R has 25 standard packages and many more packages are available for download from the CRAN family [2]. The source code for the R software environment is written in C, Fortran and R. R and its libraries implement a wide variety of statistical and graphical techniques, statistical tests, time series analysis, classification, clustering and others. R is easily extensible through functions. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made.

R language uses the RAM of our machine, so large the RAM of our machine, larger the data we can hold to work upon using R. Initially, R was not used as Big Data Analysis language due to its limitation of memory. Gradually, R got some libraries like ff, Rodbc, fffbase, rmr2 and rhdfs to handle big data. Rmr2 and rhdfs together use the power of Hadoop in order to handle big data effectively [9].

### **4. R AND HADOOP INTEGRATION**

#### **4.1 Need of R and Hadoop integration**

R has statistical capabilities and Hadoop has high processing capabilities. To use statistical capabilities of R and processing capabilities of Hadoop, we use integration of R and Hadoop.

R is powerful for statistical analysis, however, all the calculations are performed by loading entire data in RAM. RAM can be scaled up to some limit. Hadoop framework allows parallel processing of enormous amount of data. Using R with Hadoop facilitates scalability of statistical calculations [10].

#### **4.2 Ways to integrate R and Hadoop**

There are different ways of using R and Hadoop together:

1. Hadoop streaming
2. Rhipe
3. Using RHadoop packages

There are also other ways to integrate R and Hadoop. For example, RODBC/RJDBC could be used to retrieve data from R. The surveys on Internet presents that the most used methods for integrating R and Hadoop are Streaming, Rhipe and RHadoop [6].

##### **4.2.1 Hadoop streaming**

Hadoop streaming is an applicability that comes with the hadoop distribution. Hadoop Streaming utility allows to develop and run map/reduce tasks or jobs with any executable script as the mapper and/or reducer [11].

##### **4.2.2 Rhipe**

Rhipe is an acronym for “R and Hadoop Integrated Programming Environment”[12]. Rhipe is an open source project that provides assimilation between R and Hadoop. Rhipe software package is freely available for download.

It allows the user to carry out big data analysis directly in R. Rhipe is R library which allows running a MapReduce job within R.

#### 4.2.3 RHadoop packages

Revolution Analytics has developed an open source project called RHadoop. RHadoop provides client-side integration of R and Hadoop. RHadoop consists of three packages: rmr, rhdfs and rhbase. RHadoop has dependencies on other R packages and setting up of RHadoop is a tedious task. Working with RHadoop requires to install R and RHadoop packages. Rmr require Rcpp, RJSONIO, bitops, functional, plyr, digest, reshape2, stringr. Rhdfs require rJava package. RHadoop is a collection of these packages:

Rmr2- This package allows R developers to perform statistical analysis in R via Hadoop MapReduce functionality on Hadoop cluster. This package needs to be installed at every node in the cluster.

Rhdfs- This package provides basic connectivity to the Hadoop distributed file system. R programmers can retrieve, read, write and modify the files stored in HDFS from within R. This package is to be installed only on the node that runs the R client.

Rhbase- This package provides the basic connectivity to the HBASE distributed database using thrift server. R programmers can read, write and modify tables stored in HBASE from within R. This package is to be installed only on the node that runs the R client.

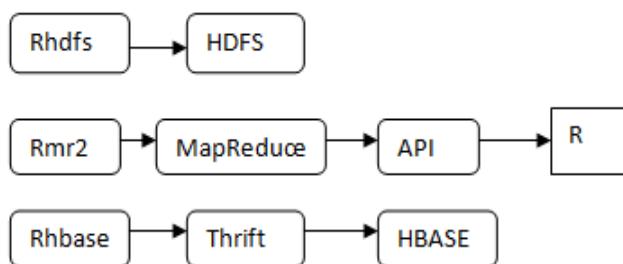


Figure 3. RHadoop Architecture

## 5. CONCLUSION

RHadoop is the complete platform where we can store and process our data efficiently and can perform some meaningful analysis. RHadoop is an open source project developed by Revolution Analytics.

We have examined the design, framework and architecture of Hadoop's MapReduce framework in detail. We would conclude by saying that big data is the new buzzword and Hadoop is the best tool available for processing data. With Revolution Analytics's RHadoop packages, data scientists can utilize the full potential of Hadoop from the R environment. If we need to combine processing capabilities of Hadoop and Statistical capabilities of R, RHadoop is the complete set for it. RHadoop has packages to integrate R with MapReduce, HDFS and HBASE. RHadoop uses package rmr to combine R with MapReduce, rhdfs to integrate with HDFS and rhbase to integrate with HBASE.

## REFERENCES

- [1] "Apache Hadoop," [Online]. Available: <http://hadoop.apache.org>.
- [2] "Udacity," [Online]. Available: <http://classroom.udacity.com>.
- [3] A. Bifet, "Mining Big Data in Real Time," *Informatics*, December 2012.
- [4] A. Gahlawat, "Big Data Analysis using R and Hadoop," *International Journal of Computational Engineering & Management(IJCEM)*, vol. 17, no. 5, pp. 9-14, September 2014.
- [5] D. Saumya Salian, "Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients," *International Journal of Science and Research(IJSR)*, vol. 4, no. 4, pp. 534-538, April 2015.
- [6] K. S. Vilas, "Big Data Mining," *International Journal of Computer Science and Management Research*, pp. 12-17, October 2013.
- [7] A. B. Wei Fan, "Mining Big Data: Current Status, and Forecast to the Future," *SIGKDD Explorations*, vol. 14, no. 2, pp. 1-5, 2014.
- [8] A. K. S. T. S.D. Ghewar, "Data Mining: Task, Tools, Techniques and Applications," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 10, pp. 8095-8098, 2014.
- [9] "Wisegeek," 2015. [Online].
- [10] A. M. D. D. S. K. Harish D, "Big Data Analysis using RHadoop," *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, vol. 2, no. 4, pp. 180-185, April 2015.
- [11] R. M. D. Bogdan Oancea, "Integrating R and Hadoop for Big Data Analysis," *Revista Română de Statistică* , 2014.
- [12] S. P. T. B. Akshay Mittal, "RHadoop: An Improved Execution Environment for Restricted MapReduce Programs," New Jersey, United States.
- [13] R. H. J. R. J. X. W. S. Saptarshi Guha, "Large Complex Data: Divide and Recombine (D&R) with RHipe," *stat*, 26 August 2012.
- [14] X. Z. G.-Q. W. W. D. Xindong Wu, "Data Mining with Big Data," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 26, no. 1, January 2014.
- [15] H. J. Alexandros Labrinidis, "Challenges and Opportunities with Big Data," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032-2033, August 2012.
- [16] C.-Y. Z. C.L. Philip Chen, "Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data," *Information Sciences*, pp. 314-347, January 2014.
- [17] S. M. Y. L. Min Che, "Big Data: A Survey," *Springer Science+Business Media*, vol. 19, pp. 171-209, 22 January 2014.
- [18] D. S. Seref Sagiroglu, "Big Data: A review," in *International Conference on Collaboration Technologies and Systems (CTS)*, San Diego, CA, 2013.

# Polarizing Sentiments in Movie Reviews

## Using Improved kNN Classifier

<sup>1</sup>Sanjeev Dhawan, <sup>2</sup>Kulvinder Singh, <sup>3</sup>Tanya Arora

<sup>1,2</sup>Faculty of Computer Science & Engineering, <sup>3</sup> PG Student of M.Tech. (Computer Engineering),

<sup>1,2,3</sup>Department of Computer Science & Engineering,

University Institute of Engineering and Technology (U.I.E.T),

Kurukshetra University, Kurukshetra (K.U.K)-136119, Haryana, INDIA

E-mail(s): <sup>1</sup>rsdhawan@rediffmail.com, <sup>2</sup>kshanda@rediffmail.com, <sup>3</sup>tanya.arora121@gmail.com

**Abstract:** With the flourish of the Web, online review is becoming a more and more useful and important information resource for people. As a result, automatic review mining and summarization has become a hot research topic recently. Different from traditional text summarization, review mining and summarization aims at extracting the features on which the reviewers express their opinions and determining whether the opinions are positive or negative. In this paper, we focus on a specific domain – movie reviews. Sentiment analysis of movie reviews has been done to detect the sentiment behind any review- positive or negative and an improved k- Nearest Neighbor (ImpkNN) classifier has been proposed which deploys the notion of attribute weighted-kNN and the weights associated are trained using 10-fold cross validation. In the end, the outputs of both Basic kNN and ImpkNN are evaluated using graphs.

**Keywords:** *Sentiment Analysis, Movie Reviews, ImpkNN classifier, Feature extraction, Cross validation*

### I. INTRODUCTION

In this internet period, online data is multiplying every second. However, the colossal information existing on web in a way has become a forceful issue for the internet users. Due to the diversification of online information, it has become a bit arduous for the users to purchase a particular choice of product or service. To subdue this problem, many online websites including e-commerce and social networks have aided users with online discussion environments or online reviews segment. Reviews voiced by electronic means encompass opinion or sentiment of the people through areas such as ordering products or rating movies or events. Hereafter, e-opinions given by numerous users on shopping websites and social media have currently been the hotspots which can be analyzed to judge sentiments. Sentiment analysis is a rational process of learning peoples' opinion towards objects, events and their elements. It extends the probability to understand the users' comments and explain how a certain product or brand is comprehended- positively or negatively. Several researches have been done to detect the polarity of the sentiment using text mining techniques-text classification and clustering. In this research, text classification approach is adopted to characterize the polarity of an opinion in a movie dataset which consists of movie reviews given by the users. The Java-ML (Java Machine Learning Library) tool is used for the purpose of data classification and an improved version of kNN classifier is recommended for sentiment analysis. Besides introduction in section I, section II surveys the previous researches done in this area, section

III describes the proposed work, section IV depicts the implementation and results and at last section V concludes the discussion and throws a light on the future scope.

## II. LITERATURE SURVEY

According to Bang Po & Lillian Lee [1], Synonyms like opinion, view, belief, conviction, persuasion, and sentiment mean a verdict one grasps as right. In similar concern, Sheng Yu & Subhash Kak [2] have found that if extracted and analysed properly, the data on social media can lead to useful predictions of certain human related events. Mike Thelwall *et al.* [3] expressed that texts frequently comprise of a combination of positive and negative sentiments and sometimes it is necessary to investigate both side by side. In text classification, k-nearest neighbour (KNN) is a simplest and efficient classifier (proposed by Cover P. & Hart T. [4]). However, due to its lazy learning approach (further surveyed by Sebastiani F. [5]) without pre-modelling, KNN is not cost efficient to categorize new documents when training set is large. Duoqian Miao *et al.* [6] developed a hybrid system grounded on inconsistent precision rough set to aggregate the efficiency of both KNN and Rocchio techniques and overcome their weaknesses. Also, a reassessment of the KNN model was being approached by L.Cucala *et al.* [7] as a statistical technique. Yuki Murahmatsu *et al.* [8] built a program capable to perform entailment judgment on the account of word overlap. To get to know how users are making choices from given information and what are the reasons that persuaded them, Lionel M. [9] presented different methods to detect influence in social media by emotion recognition. Jansen & Zhang [10] investigated the entire framework of these twitter posts, expressions, and the mobility in positive or negative sentiment. Narayanan *et al.* [11] observed sentiment analysis of conditional sentences. Thelwall *et al.* [3] also attempted to identify sentiments' polarity in text along with the strength of those sentiments. Consequently, from this background analysis, it can be inferred that different researchers have introduced different techniques by using distinct classifiers in their attempt to ameliorate accuracy while extracting emotional behaviour. Thus, in the present paper, an attempt has been made to design an improved version of the basic kNN classifier for sentiment analysis of movie reviews.

## III. PROPOSED WORK

A proposal has been made to distinguish human sentiments/emotions based on their polarity i.e. whether positive or negative using text classification. The proposed research deals with the improvements made in the basic kNN classifier so that to provide more accurate and precise results while polarizing emotions.

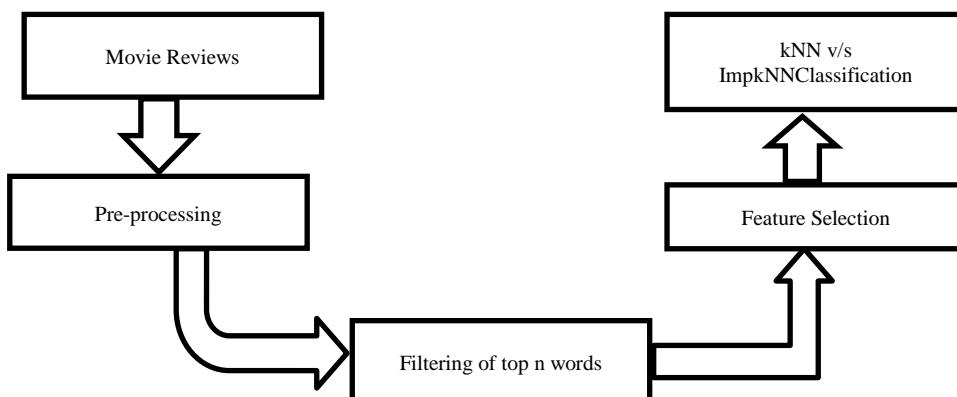


Figure 3.1: Outline of Proposed System

#### A. Movie Reviews Dataset

At first, movie reviews dataset file is read thoroughly for understanding and manually classifying the positive and negative words that express sentiments or emotions as per our knowledge.

#### B. Pre-processing

The dataset collected may be loosely organised and can consist of irrelevant or irrational text. Text pre-processing is defined as the task of cleaning the unwanted data from the whole and preparing the relevant data for classification. Various pre-processing methods are described below:

(a) *Convert to lowercase*: This is done to avoid make out between same words simply on case. Thus, all the words of the text documents are converted to lowercase for ease.

(b) *Removing URLs*: As URLs do not account for any kind of sentiment expression, so should be removed for any further confusion. For example, “I hate this website: [www.Happiness.com/](http://www.Happiness.com/)”. Here, negative sentiment is expressed by a person but due to the presence of the word happiness in the sentence, it is considered as neutral.

(c) *Symbols, Numbers and Punctuation removal*: Special symbols and numbers are not pertinent while observing sentiments. Although Punctuation can provide grammatical context which supports understanding, yet it is irrelevant in determining sentiments.

(d) *Apply Stemming*: Stemming is applied to transform any word to its root word or dictionary based form. Porter algorithm is used for stemming in this paper. For example, for words like ‘amazing’, ‘amazed’, ‘amazes’ in different sentences, the root word is ‘amaze’.

#### C. Top n words selection

After pre-processing, words are selected from data according to their recurrence or count frequency. A threshold cut of  $(n > i)$ , where “ $i$ ” is a natural number; is applied to select top  $n$ -words that are occurring frequently. These frequent words amount to the interesting features that need to be analysed and further leads to more accurate results while polarizing sentiments.

#### D. Feature Extraction

Feature extraction is used to decrease the dimensional reduction of the feature space. In the basic approach of statistical feature extraction, recurring words in the corpus are treated as feature values. A vector space model (VSM) represents the words (feature values) and their frequency score in a form of matrix (rows contain words and columns contain their corresponding terms (weights); any greater than zero entry indicates the presence of the word. TF-IDF (Term Frequency and Inverse Document Frequency) weighting is simply used for calculating weight for each word which reflects how important a word is to the document in a corpus of data.

(a) *Classification*: Text classification is defined as the process of assigning documents to suitable pre-defined classes/categories. The oldest and the simplest classifier used for text classifying is kNN, which is an illustration-based method that delays the decision to extrapolate outside the training instances until another query comes across. Whenever there is a new instance to classify, its  $k$ -nearest neighbours are explored by selecting a reasonable distance measure such as Euclidean distance. However, due to curse of dimensionality in basic kNN, as the similarity metrics do not take into account the relation of attributes. Thus, to subdue this problem, ImpkNN is purported that is much more efficient than the basic kNN.

(b) *ImpkNN*: Classification using ImpkNN uses the concept of attribute (word) selection and weighting. As it has been usually seen that some attributes are more relevant than others while classification. It has been frequently observed that the classification of data depends more upon some attributes than others. Finding those

relevant attributes is called attribute selection and making a method in which those relevant attributes are given more value is called attribute weighting. Thus, at very first, a filtering approach is deployed for choosing relevant attributes by fixing a threshold value. Then, a method in which those relevant attributes are given more importance than the rest, known as attribute weighting is used. Normalized weighting, a type of attribute weighting, normalizes weights using standard deviation. Using this measure, random weights are assigned to each attribute which are then trained by cross validation.

(c) *Cross validation:* In the cross-validation each record is used for the same number of times for training but only once for testing. As an example two partitions of the data set are made and one partition is used for training and the other for testing and next time vice versa is done. This is called two-fold cross-validation. This can be generalized by partitioning the dataset into  $k$  equal sized subsets. During each run one subset is used for testing while others are used for training. This process is repeated until each subset is tested once. This method is called  $k$ -fold cross validation.

#### IV. PROBLEM DESIGN AND IMPLEMENTATION

The dataset used for analysing sentiments is Cornell movie-review corpora, containing users' reviews about movies which are further polarized using proposed sentiment analysis method. There are 100000 entries of users' reviews about movies. This dataset is obtained from <http://www.cs.cornell.edu/people/pabo/movie-review-data/>

##### A. Algorithm of ImpkNN

- 1) *Read the training data from a file  $\langle x, f(x) \rangle$*
- 2) *Read the testing data from a file  $\langle x, f(x) \rangle$*
- 3) *Set K to some value.*
- 4) *Set the learning rate  $\alpha$*
- 5) *Set the value of N for number of folds in the cross validation*
- 6) *Normalize the attribute values by standard deviation*
- 7) *Assign random weight  $w_i$  to each attribute  $A_i$*
- 8) *Divide the number of training examples into N sets*
- 9) *Train the weights by cross validation*
  1. *For every set  $N_k$  in N, do*
    - a. *Set  $N_k = Validation\ Set$*
    - b. *For every example  $x_i$  in N such that  $x_i$  does not belong to  $N_k$  do*
      - (i) *Find the K nearest neighbours based on the Euclidean distance*
      - (ii) *Return the class that represents the maximum of the k instances*
      - (iii) *If actual class  $\neq$  predicted class then apply gradient descent*
        - *Error = Actual Class – Predicted Class*
        - *For every  $W_k$* 
          - *$W_k = W_k + \alpha * Error * V_k$  (where  $V_k$  is the query attribute value)*
    - c. *Calculate the accuracy as*
      - *Accuracy = (# of correctly classified examples / # of examples in  $N_k$ ) X100*
      - *Repeat the process till desired accuracy is reached*
  - 10) *Train the weights on the whole training data set*

1. For every training example  $xi$ 
    - a. Find the  $K$  nearest neighbors based on the Euclidean distance
    - b. Return the class that represents the maximum of the  $k$  instances
    - c. If actual class  $\neq$  predicted class then apply gradient descent
      - Error = Actual Class – Predicted Class
      - For every  $Wk$
      - $Wk = Wk + \alpha * Error * Vk$  (where  $Vk$  is the query  $n$  attribute value)
  2. Calculate the accuracy as Accuracy = (# of correctly classified examples / # of training examples)\*100
  3. Repeat the process till desired accuracy is reached
- 11) For each testing example in the testing set
- a. Find the  $K$  nearest neighbours based on the Euclidean distance
  - b. Return the class that represents the maximum of the  $k$  instances
  - c. Calculate the accuracy as
- Accuracy = (# of correctly classified examples / # of testing examples)\*100

B. Evaluation( Basic kNN versus ImpkNN)

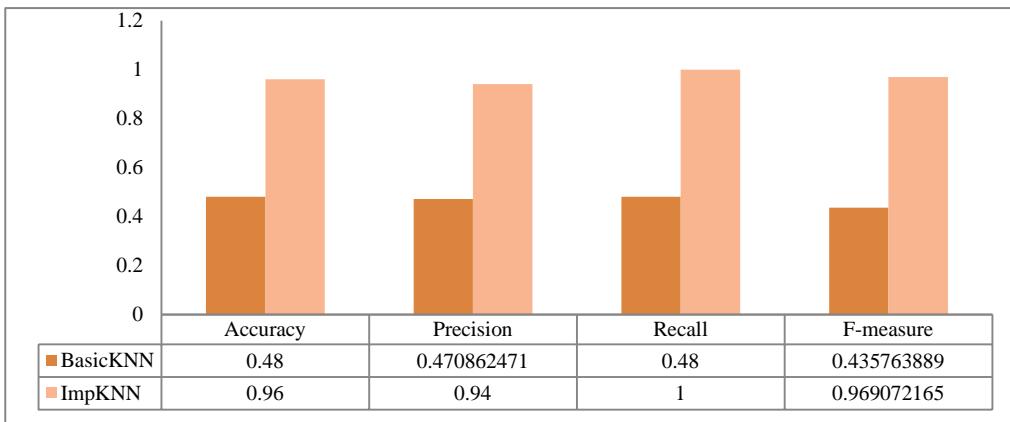


Figure 2: Basic kNN versus ImpkNN

- a. Accuracy: In sentiment analysis, it is the ratio of rightly classified words to the total count of words.
- b. Precision: It is the count of words rightly labelled as positive or negative multiplied by the inverse of total number of times that words are classified as positive or negative.
- c. Recall: It is described as the count of words aptly labelled as positive divided by the total count of words that are truly positive.
- d. F-Measure: It is given by the harmonic mean of precision and recall. Figure 2 compares F-measure of Basic kNN and ImpkNN.

## V. CONCLUSION

In the present research, ImpkNN classifier is used for sentiment analysis of movie dataset which outdoes the Basic kNN as observed from Figure 2. Sentiment analysis of movie reviews studies the polarization of opinions articulated by the users which can be used for business and commercial commitments. In future, classifiers like SVM, Naïve Bayes, Genetic algorithm and fuzzy classification can also be applied on movie reviews dataset to

find out the most competent classifier. The classification can be done not only on movie reviews but also on many other products' or services' reviews available on social networks and e-commerce websites.

#### REFERENCES

- [1] Pang B. & Lillian L., 'Opinion Mining and Sentiment Analysis', Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2, pp. 1–135, 2008.
- [2] Yu S. & Kak S., 'A Survey of Prediction Using Social Media', Springer, pp. 353-374, 2009.
- [3] Thelwall M., Buckley K., Paltoglou G. & Cai D., 'Sentiment Strength Detection in Short Informal Text', Journal of the American Society for Information Science and Technology, 61(12), 2544–2558, 2010.
- [4] Cover, T., & Hart, P., 'Nearest neighbour pattern classification', IEEE Transaction on Information Theory, 13(1), 21–27, 1967.
- [5] Sebastiani, F., 'Machine learning in automated text categorization', ACM Computing Surveys, 34(1), 1–47, 2002.
- [6] Miao D., Duan Q., Zhang H. & Jiao N., 'Rough set based hybrid algorithm for text classification', Journal of Expert Systems with Applications, 36(5), pp. 9168-9174, 2009.
- [7] Cucala L., Marin J., Robert C. & Titterington D., 'A Bayesian reassessment of nearest neighbour classification', arXiv:0802.1357v1 [stat.CO] 10 Feb 2008.
- [8] Muramatsu Y., Ueda K. & Yamamoto K., 'Textual Entailment Recognition using Word Overlap,Mutual Information and Subpath Set', Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 18-27, Beijing, August 2010.
- [9] Lionel Martin, 'User Behavior under the Influence of Groups in Social Media', EDIC Research Proposal, 2009.
- [10] Jansen B. & Zhang M., 'Twitter Power: Tweets as Electronic Word of Mouth', Journal Of The American Society For Information Science And Technology, 60(11):2169–2188, 2009.
- [11] Narayanan R., Liu B. & Choudhary A., 'Sentiment Analysis of Conditional Sentences', Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 180–189, Singapore, 6-7 August 2009.
- [12] Tanya Arora, Sanjeev Dhawan, Kulvinder Singh, "Sentiment Analysis of Movies' Reviews using Improved k-Nearest Neighbour Classifier", pp. 241-245, Advances in Computer and Information Technology(ACSIT), p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 4; April-June,2016.
- [13] Sanjeev Dhawan, Kulvinder Singh, Tanya Arora, "Examining Polarization of Emotions in Online Shopping Websites", accepted in Proceedings of National Conference on Advances in Computer Science & Engineering (ACSE-2016) held on April 29-30, 2016.

# Iris Recognition using Optimization Technique

Rajvir kaur<sup>1</sup>

M.tech(Student)

Baba Banda Singh Bahadur Engineering  
College Fatehgarh sahib, Punjab, India  
[kaurrajvir190@gmail.com](mailto:kaurrajvir190@gmail.com)

Ishpreet singh<sup>2</sup>

Assistant Professor

Baba Banda Singh Bahadur Engineering  
College Fatehgarh sahib, Punjab, India  
[ishpreet.virk@bbsbec.ac.in](mailto:ishpreet.virk@bbsbec.ac.in)

---

## ABSTRACT

**Iris biometrics research is an exciting, broad, and rapidly expanding field. At the same time there are successful practical applications that illustrate the power of iris biometrics, there are also many fundamental research issues to be solved on the way to larger scale and more complex applications.**

In this research work, iris recognition has been done using Hough Man Circular Transform (HCT), Scale Invariant Feature Transform (SIFT) and Genetic Algorithm (GA) method. Hough Man Circular Transform (HCT) localize the retina, Scale Invariant Feature Transform (SIFT) extract the features of the iris templates and then in the end Genetic Algorithm (GA) reduce the obtained features. The whole simulation is being implement in MATLAB 2010 environment. The performance of the system is evaluated by using with False Acceptance Rate (FAR), False Rejection Rate (FRR) and Recall Rate (RR) parameters.

## KEYWORDS

**Iris Recognition, Security, Biometrics, Scale Invariant Feature (SIFT), Hough man Circular Transform (HCT).**

---

## I. INTRODUCTION

An extensive variety of systems require dependable individual recognition schemes to either confirm or decide the identity of an entity requesting their services. The reason of such schemes is to make sure that the render services are access only by a rightful user, and not anyone else. Example of such applications includes secure access to buildings, computer systems, laptops, cellular phones and ATMs. In the nonexistence of strong personal recognition schemes, these systems are susceptible to the wiles of an impostor [1, 2]. Biometric recognition, or just biometrics, refers to the mechanical recognition of persons based on their physiological and behavioral individuality. By using biometrics it is probable to confirm or establish an individual's identity based on "who she is", rather than by "what she possesses" (e.g., an Identity Card ) or "what she remembers" (e.g., a password). In this document, give a concise impression of the field of biometrics and summarize some of its compensation, disadvantage, strengths, limitations, and linked isolation concerns [3, 4].

Substitute representations of identity such as passwords and ID cards are not sufficient for reliable identity determination because they can be easily misplaced, shared, or stolen. Biometric recognition is the science of establishing the identity of a person using his/her anatomical and behavioral traits. Commonly used biometric traits include fingerprint, face, iris, hand geometry, voice, palm print, handwritten signatures [5, 6, 7, 8]. Three features that influenced the increased interest in the biometric are as follows: 1) public acceptance; 2) new user-friendly capture devices with broad improved capabilities; and 3) a broadened range of applications. But the main issue of all authentication systems is that it must have, high accuracy rate and low error rate. So it can only be achieved by using SIFT feature extraction for Iris, then for feature reduction genetic algorithm is applied and features are reduced. That's why proposed algorithm is based on this algorithm along with some preprocessing on CASIA dataset. During feature extraction process, we employ feature extraction method to get feature vectors. In addition to this, we will propose a local feature extraction method to reduce the amount of data using GA and improve performance. [9, 10].

## II. HOUGHMAN CIRCULAR TRANSFORMATION (HCT)

The Hough transform can be applied to detect the presence of a circular shape in a given image. It is being utilized to discover any figure or else to find the iris in the human being's face [15]. The characteristic equation of a circle of radius  $r$  and centre  $(a, b)$  is given by:

$$(x-a)^2 + (y-b)^2 = r^2 \quad (1)$$

This particular circle could possibly be illustrated through the two of subsequent equations:

$$\begin{aligned} x &= a + r \cos(\theta) \\ y &= b + r \sin(\theta) \end{aligned} \quad (2)$$

Thus, the role of the Hough transform is to search for the triplet of parameters  $(a, b, r)$  which determines the points. Two cases may be presented as described:

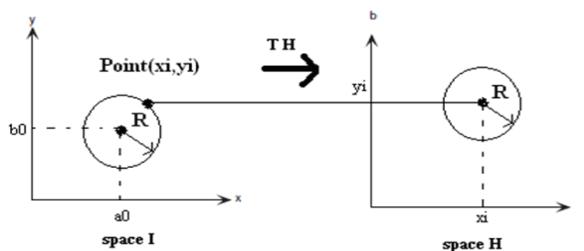


Fig. 1 Hough Transformation

### A. Case of known radius

If we know the radius of the circle to be detected in the actual picture, the real parameter so that to search is reduced in the direction of a pair  $(a, b)$  and the H space is two-D. We deliberate a circle of radius  $R_1$  as well as centre  $(a, b)$ , the transformation for each point  $(x, y)$  in space I yields a circle in space H having a centre  $x, y$  and radius  $R$ .

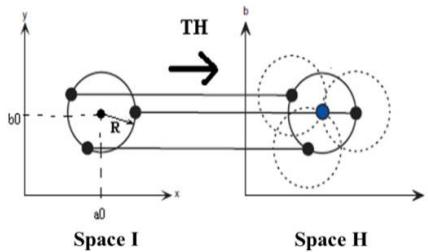


Fig. 2 Detection of Circle

Similarly, we transform all points of the circle in picture. The end result will be more circles where their intersection is the point  $(a, b)$ . This point is obtained by searching the maximum of the accumulator.

#### B. Case of unknown radius

In this case, the work consists to find the triplet parameters  $(a, b, r)$  which define the points of the circle to discover. The space will be available in 3D.

For each point  $(x, y)$  of the space I will match a cone in space H, as the radius  $r$  vary from 0 to an actual given value. After transforming of all points of contour in the equivalent way, the intersection will able to provide a spherical surface corresponding to the maxim of accumulator. The area is characterized by a center  $(a, b)$  and the radius  $r$  searched.

### III. SCALE INVARIANT FEATURE TRANSFORM (SIFT)

Scale Invariant Feature Transform (SIFT) was proposed by David Lowe [11] that has the capacity to distinguish and depict neighborhood picture elements positively. SIFT extract the so many features but some features are :

- Edges
- Intersect point
- Blobs
- Ridges

#### A. Edges

The points at which image brightness changes sharply are typically organized into a set of curved line segments termed edges.

### B. Intersect point

An interest point is a point in the image which is general can be characterized. Intersect point is a well-defined position in image space.

### C. Blobs

A blob is a region of an image in which some properties are constant or approximately constant. All the points in a blob can be considered in some sense to be similar to each other.

### D. Ridges

A ridge is a curved line in a finger image. Some ridges are continuous curves and some ridges terminate at specific points called ridge endings. When two ridges come together at a point called a bifurcation. Ridge endings and bifurcations are known as minutiae.

The necessary SIFT appraisal comprises of five noteworthy stages:

- A. Scale-space local extreme detection
- B. Key-point localization
- C. Orientation assignment
- D. Key-point descriptor
- E. Trimming of false matches

The accompanying sub-segment will portray every stage.

#### A. Scale-space Local Extreme Detection

The first step of key point detection involves identification of locations that can be assigned with a change in view as well as change in scale. In such locations, which are invariant towards scale changes, are found by searching stable features across all the possible scales using scale space that is a continuous function of scale [20]. Gaussian function is the only possible scale space function. For that reason the scale space of picture is well-defined as, 2D Gaussian operator  $G(a_1, b_1, \sigma)$  with the i/p picture  $J(a_1, b_1)$ :

$$L(a_1, b_1, \sigma) = G(a_1, b_1, \sigma) * J(a_1, b_1) \quad (3)$$

Where  $J(a_1, b_1)$  is the input image and  $*$  is the convolution operation in  $a_1$  and  $b_1$ .  $G(a_1, b_1, \sigma)$  is the variable scale Gaussian defined as

$$G(a_1, b_1, \sigma) = (1/2\pi\sigma^2) e^{(-a_1^2 + b_1^2)/2\sigma^2} \quad (4)$$

Difference of Gaussian (DoG) function is convolved with the image to detect stable key-point locations. For two nearby scales of an iris image  $J$ , the Difference of Gaussian (DOG) is computed as of images are:

$$\begin{aligned} D(a_1, b_1, \sigma) &= (G(a_1, b_1, k\sigma) - G(a_1, b_1, \sigma)) * J(a_1, b_1) \\ &= L(a_1, b_1, k\sigma) - L(a_1, b_1, \sigma) \end{aligned} \quad (5)$$

Where  $k$  is a constant multiplicative factor in scale space that is used for changing the scale and  $a_1, b_1$  are the coordinates of a pixel in image  $J$ . Nearby extremes are then recognized by watching each image point in  $J(a_1; b_1; \sigma)$ . A point is decided as a local minimum or maximum when its value is smaller or larger than all its surrounding neighboring points. This technique is scale invariant, hence is appropriate for annular iris images as the dimension of iris varies due to dilation and contraction of the pupil [42].

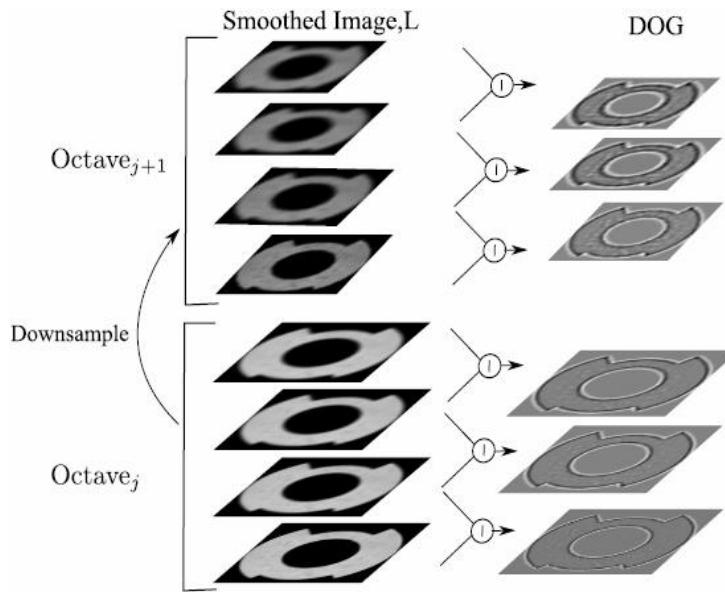


Fig. 3 Detection of scale space extreme

#### B. Accurate Key-point Localization

To detect the importance points, DOG pictures are utilized also local maxima as well as local minima are computed across different scales. Each pixel of a DOG image is compared to 8 neighbors in the same scale and 9 neighbors in the neighboring scales.

After key point detection, the next step is performing the detailed fit to the adjoining data intended for location, the proportion of principal curve as well as the scale. The basic idea behind this is to reject all those key points which

are low in contrast. These low contrast key points are not considered because as stated in, such key point are sensitive to noise or badly limited to a small area. 3D quadratic operator is fixed in the direction of local key point in order to determine the interpolated location of maximum. The authors have used Taylor expansion of the scale space function,  $D_m(a_1, b_1, \sigma)$  shifted so that the origin lies at the sample point

$$D_m(z) = D + (\partial^2 D_m^{-1} / \partial z^2) * (\partial D_m / \partial z) \quad (6)$$

Where  $D_m$  and its derivatives are calculated at the sample point and  $z=(a_1, b_1, \sigma)^T$  is an offset from this point. The location of extremum ( $z$ ) is obtained by taking the derivative of this function with respect to  $x$  and set the situation towards zero, accordingly giving

$$z = (-\partial^2 D^{-1} / \partial z^2) * (\partial D / \partial z) \quad (7)$$

The offset is compared to a predefined threshold and if it is larger, then it implies that  $z$  is close to some different sample point. In this case sample point is changed and interpolation is performed about that point. The final offset is then added to the sample point to get the interpolated location of extremum.

### C. Orientation Assignment

To attain invariance to picture rotations, an orientation is allocated towards each and every one of the key-point localities. The descriptor could possibly be represented comparative to this orientation. For determination of the key point orientation, a gradient orientation histogram is worked out in the neighborhood of the key point. A Gaussian smoothed image  $L_1$  is selected using the scale of a particular key-point. On behalf of a Gaussian smoothed picture  $L_1(a, b)$ , magnitude ( $m(a, b)$ ) and orientation ( $\theta(a, b)$ ) are calculated as

$$m(x, y) = \sqrt{L_1(x+1, y) - L_1(x-1, y)}^2 + \sqrt{L_1(x, y+1) - L_1(x, y-1)}^2 \quad (8)$$

$$\theta(x, y) = \tan^{-1}\left(\frac{L_1(x, y+1) - L_1(x, y-1)}{L_1(x+1, y) - L_1(x-1, y)}\right) \quad (9)$$

This is followed by formation of the orientation histogram for gradient orientation around each of the particular key-points. The actual histogram encompasses 36 bins designed for 360 orientations and before adding it to the actual histogram, each and every example is weighted by means of gradient magnitude and Gaussian weighted circular frame, by means of  $\sigma$  of 1.5 times the scale of actual key-point. Peaks in histogram correspond to the orientations.

### D. Key-Point Descriptor

In this stage, a particular descriptor is registered at every key-point. The picture gradient magnitudes and introductions, with respect to the significant introduction of the key point, are inspected inside a 16X16 locale around every key-point. These specimens are then amassed into orientation histograms summarizing the contents over 4X4 sub regions.

#### *E. Trimming of False Matches*

The key-point matching procedure described may generate some erroneous coordinating focuses. We have evacuated spurious coordinating focuses using geometric limitations. We constrain ordinary geometric varieties to small rotations and displacements. Therefore, if we place two iris images side by side and draw matching lines true matches must appear as parallel lines with similar lengths. According to this observation, we compute the predominant orientation  $QP$  and length  $lp$  of the matching, and keep the matching pairs whose orientation  $\mu$  and length ` are within predefined tolerances  $cQ$  and  $cP$  so that  $|Q - QP| < cQ$  and  $|l - lp| < cl$ .

## IV. GENETIC ALGORITHM

The genetic algorithm is a replica of machine learning which follows its actions as of metaphor process of development in nature. This is completed by the formation inside a machine of a population of individuals shown by chromosomes by a set of character strings which are similar to the base-4 chromosomes. The person in the inhabitants departs during an evolution process . Evolution is not considered or directed process. Mean, there is no confirmation to hold up the declaration that the objective of development is to create Mankind[16].

Genetic algorithm has mainly three operators;

1. Selection, I which selection of chromosome is done.
2. Mutation, in which two chromosomes gets mutated to generate child.
3. Crossover, to apply new changes.

Genetic Algorithm Various Phases:

Step 1: At random, produce an initial population  $M(0)$ .

Step 2: Compute as well as help save the actual fitness  $f(m)$  for every specific individual  $m$  in the current population  $M(t)$ ;

Step 3: Specify selection probabilities  $p(m)$  for every specific individual  $m$  throughout  $M(t)$  making sure that  $p(m)$  is actually proportional to  $f(m)$ .

Step 4: Crank out  $M(t+1)$  by simply probabilistically choosing individuals from  $M(t)$  to produce offspring via genetic operators.

Step 5: Repeat step 2 until satisfying solution is actually attained.

## V. METHODOLOGY

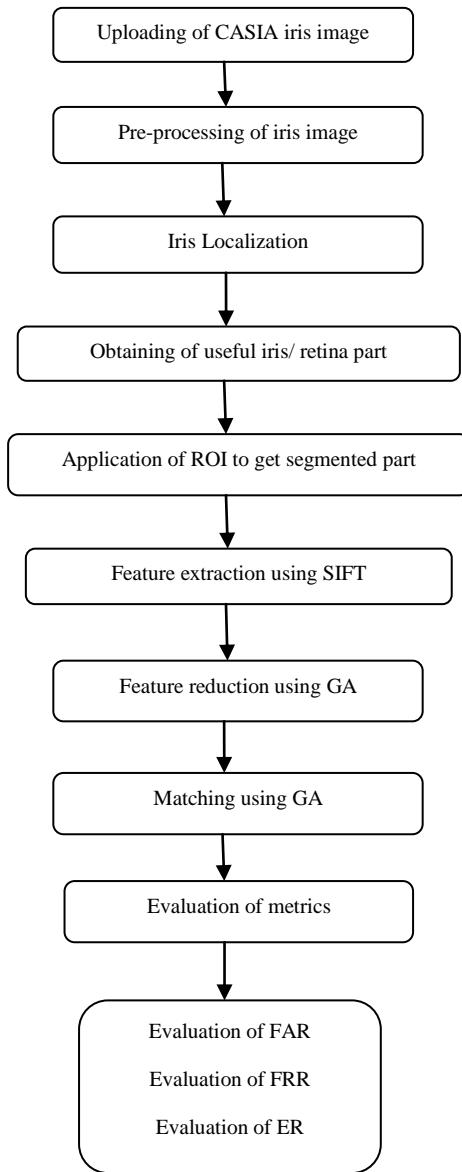


Fig. 5.1 Proposed work Flow chart

Step 1. Pre processing of iris image

- Pre-processing can convert original image into gray image. In pre-processing we use edge detection for finding the edge image to extract the iris part from the eye image.
- For preprocessing is binarization is used. Conversion of unassigned 8 bit data into 0 and 1 form in Binarization. Black and white image appears after that. After the binarization , the image is in the form of 0

& 1. Then the edges are discovered of the image using “edge” command by canny technique. We use in this format edge (I,'canny',thresh) specifies sensitivity thresholds for the Canny method.

#### Step 2. Iris localization

The acquired iris image has to be preprocessed to detect the iris, which is an annular portion between the pupil (inner boundary) and the sclera (outer boundary). The first step in iris localization is to detect pupil which is the black circular part surrounded by iris tissues. The center of pupil can be used to detect the outer radius of iris patterns. The important steps involved are:

- A. Pupil detection
  - B. Outer iris localization
- A. *Pupil Detection*

The iris image is converted into grayscale to remove the effect of illumination. As pupil is the largest black area in the intensity image, its edges can be detected easily from the binarized image by using suitable threshold on the intensity image. But the problem of binarization arises in case of persons having dark iris. Thus the localization of pupil fails in such cases. In order to overcome these problems Circular Hough Transformation for pupil detection can be used. The basic idea of this technique is to find curves that can be parameterized like straight lines, polynomials, circles, etc., in a suitable parameter space. The transformation is able to overcome artifacts such as shadows and noise.

#### B. *Outer Iris Localization*

External noise is removed by blurring the intensity image. But too much blurring may dilate the boundaries of the edge or may make it difficult to detect the outer iris boundary, separating the eyeball and sclera. Thus a special smoothing filter such as the median filter [8] is used on the original intensity image. This type of filtering eliminates sparse noise while preserving image boundaries. After filtering, the contrast of image is enhanced to have sharp variation at image boundaries using histogram equalization of different radii from the pupil center and the intensities lying over the perimeter of the circle are summed up. Among the candidate iris circles, the circle having a maximum change in intensity with respect to the previous drawn circle is the iris outer boundary. Figure shows an example of localized iris image.

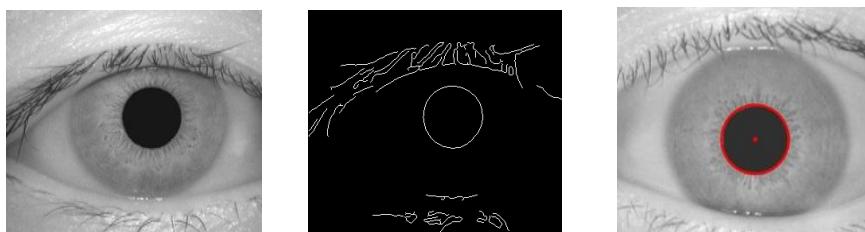


Fig. 5.2 Steps involved in detection of inner pupil boundary

### Step 3. Obtaining useful iris

The iris image should be rich in iris texture as the feature extraction stage depends upon the image quality and accurate extraction of iris area from image. High resolution and good sharpness: It is necessary for the accurate detection of outer and inner circle boundaries. Good lighting condition: The system of diffused light is used to prevent spotlight effect.

### Step 4. Application of ROI using segmentation techniques

For improvement in the recognition rate we need to extract the feature of iris but there are problem during feature extraction of iris. When segmentation of iris is not well then we cannot extract more features. So for feature extraction and improvement in recognition rate of proposed system we need to segment proper area of iris.

### Step 5. Feature Extraction using SIFT

After the ROI to find the feature of segmented iris. So for feature extraction use Scale-invariant feature transform (SIFT). SIFT is an image descriptor for image-based matching and recognition developed by David Lowe. This descriptor as well as related image descriptors is used for a large number of purposes in computer vision related to point matching between different views of a 3-D scene and view-based object recognition. The SIFT descriptor is invariant to translations, rotations and scaling transformations in the image domain and robust to moderate perspective transformations and illumination variations. Experimentally, the SIFT descriptor has been proven to be very useful in practice for image matching and object recognition under real-world conditions. The SIFT descriptor was computed from the image intensities around interesting locations in the image domain which can be referred to as interest points, alternatively key points.

### Step 6 Feature Reduction and Matching using GA

Firstly load all extracted feature point for the feature reduction process. Before applying Genetic Algorithm to initialize the GA basic function using ‘optimset’ function like population size, mutation function, crossover etc. After that we use ‘ga’ function for feature reduction according to the fitness function. Fitness function set according to our requirement like which type of feature we can use for classification purpose. Genetic algorithm (GA) is a method for solving both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution. The algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm randomly selects individuals from the feature vector.

## VI. RESULTS AND DISCUSSION

The whole simulation is being done in MATLAB environment to build the iris recognition system. In proposed work various techniques has been compared with recent one using FAR as well as FRR.

**False Accept Rate (FAR):** FAR is the type of error in the pattern recognition system which is measured by:

$$FAR = \frac{\text{Total Number of Features} - \text{Total Number of Falsely Accepted Features}}{\text{Total Number of Features}}$$

False rejection rate (FRR): The percentage of times a valid user is rejected by the system. Its formula is given as:

$$FRR = \frac{\text{Total Number of Features} - \text{Total Number of Falsely Rejected Features}}{\text{Total Number of Features}}$$

Accuracy: Accuracy is a general term used to describe how accurate a biometric system performs. Its formula is given as:

$$\text{Accuracy} = 100 - (FAR + FRR)$$

*Database:* CASIA Iris Image Database (CASIA-Iris) developed by our research group has been released to the international biometrics community and updated from CASIA-IrisV1 to CASIA-IrisV3 since 2002. More than 3,000 users from 70 countries or regions have downloaded CASIA-Iris and much excellent work on iris recognition has been done based on these iris image databases. The most important part of any test of a biometric system is the data collection effort. One of our aims is to create a multi-database, where each component database is created by using iris collected with a different sensor. Database contains total 10 images from 10 people. The image format is JPEG, 256X256 gray-level, uncompressed. The image resolution, which could slightly change depending on the database, is about 500 dpi. The image size varies depending on the image. The orientation of iris image is approximately in the range [-30°, +30°] with respect to the vertical orientation. The databases used in this contest have not been necessarily acquired in a real-world application environment and are not collected according to a formal protocol.

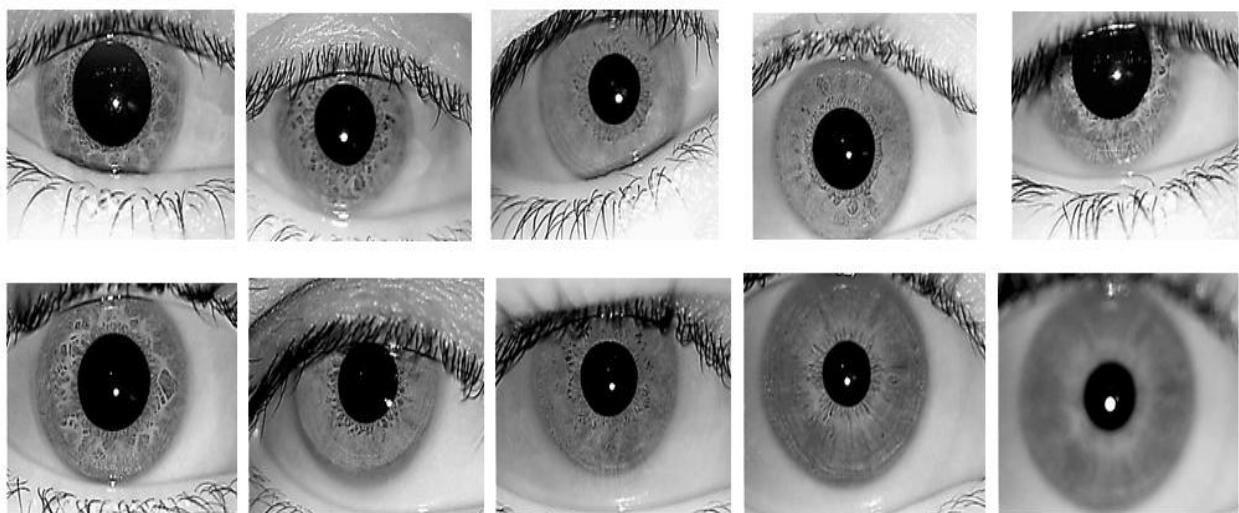
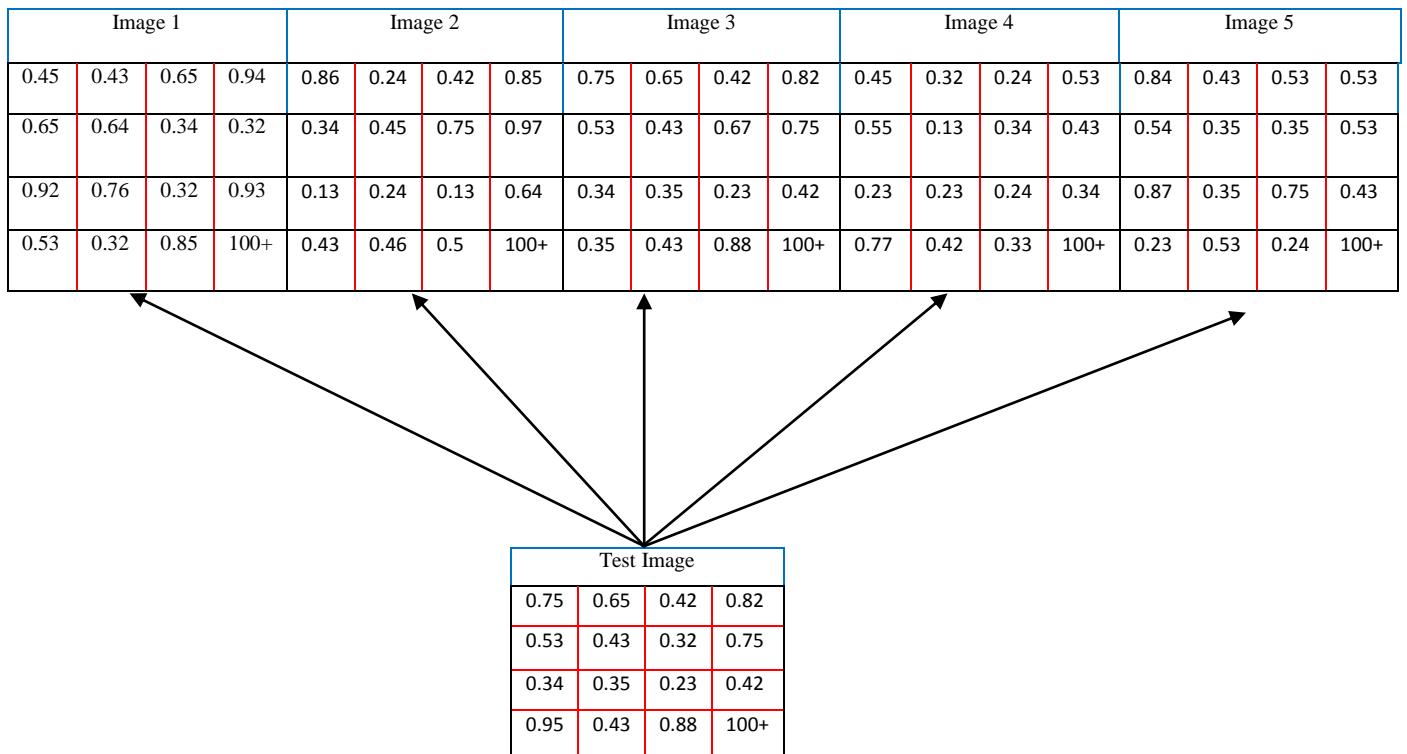


Fig. 6 Database Images Samples

Feature Datasets of Training



Proposed matching Process

In above table show the matching process in our proposed work. Using Euclidean Distance we check the test feature data matching ratio with dataset features. After that we observed that test image features matched with 3<sup>rd</sup> number of dataset features. But some feature point different from dataset feature so we calculate FAR and FRR.

In proposed work we obtained FAR value:

$$FAR = \frac{129 - 2}{129} = 0.985$$

In proposed work we obtained FRR value:

$$FRR = \frac{129 - 0.985}{129} = 0.993$$

In proposed work we obtained Accuracy value:

$$Accuracy = 100 - 0.985 + 0.993 = 98.022$$

TABLE : 1 Comparison with Previous Methods

Author	FAR	FRR	ACCURACY
Wildes [12]	2.4	2.9	87.43
Avila [13]	0.03	2.08	76.98
Tisse [14]	1.94	8.79	85.32
Proposed	0.985	0.993	98.022

## VII. CONCLUSION AND FUTURE SCOPE

In this work, we have focused on iris biometrics. The reasons for selecting this is that iris is acceptable personally, socially and legally as an identification procedure. In this work, iris biometric system is developed using GA and SIFT feature extraction method. GA was used to optimize SIFT result. The results show that the proposed biometric system leads to better security than the existing unimodal biometric identification systems as the FAR and FRR values have decreased and accuracy has increased. The efficiency of authentication system can be further increased using more modalities for fusion. In future, work can be done to remove the complexity of multimodal biometrics and to remove the errors. Other classification and optimization techniques can be opted to get the same. There is scope for further improvement in feature extraction also. Improved or advance algorithms can be adopted to get better and clear features.

## REFERENCES

- [1] N. Khatoon, M. K. Ghose, "Multimodal Biometrics: A Review", International Journal of Computer Science and Information Technology & Security, Vol. 3, No.3, 2013 12. A. A.
- [2] Libor Masek. Recognition of Human Iris Patterns for Biometric Identification. School of Computer Science and Soft Engineering, the University of Western Australia, 2003.
- [3] J.Daugman, "How Iris Recognition Works", University of Cambridge, 2001.
- [4] F.Adler, Physiology of the Eye: Clinical Application, fourth ed. London: The C.V. Mosby Company, 1965.
- [5] J.Daugman, Biometric Personal Identification System Based on Iris Analysis, United States Patent, no.5291560, 1994
- [6] R. Roizenblatt, P. Schor et al. Iris recognition as a biometric method after cataract surgery. Biomed Eng Online. 2004; 3: 2
- [7] . G. B. Iwasokun, O. C. Akinyokun, B. K. Alese and O. Olabode, "Fingerprint Image Enhancement: Segmentation to Thinning", International Journal of Advanced Computer Science and Applications (IJACSA), Vol 3, No. 1, Pages 15 – 24, 2012
- [8] H. Rhody Chester F. "Hough Circle Transform", Carlson Center for Imaging Science Rochester Institute of Technology October 11, 2005.
- [9] A. Bouridane, "Recent Advances in Iris Recognition: A Multiscale Approach", Imaging for Forensics and Security, Signals and Communication Technology, DOI 10.1007/978-0-387-09532-5 4, Springer Science+Business Media, LLC 2009.
- [10] Flitton, G.; Breckon, T. (2010). "Object Recognition using 3D SIFT in Complex CT Volumes" (PDF). Proceedings of the British Machine Vision Conference. pp. 11.1–12.
- [11] Scovanner, Paul; Ali, S; Shah, M (2007). "A 3-dimensional sift descriptor and its application to action recognition". Proceedings of the 15th International Conference on Multimedia. pp. 357–360.

- [12] R. P. Wildes, "Iris recognition: an emerging biometric technology," in Proceedings of the IEEE, vol. 85, no. 9, pp. 1348-1363, Sep 1997.
- [13] R. Sanchez-Reillo, C. Sanchez-Avila, Iris recognition with low template size, in: Proceedings of International Conference on Audio-and VideoBased Biometric Person Authentication, pp. 324–329, 2001.
- [14] C. Tisse, L. Martin, L. Torres and M. Robert, " Person identification technique using human iris recognition", International Conference on Vision Interface, Canada, 2002.
- [15] H. Rhody Chester F. "Hough Circle Transform", Carlson Center for Imaging Science Rochester Institute of Technology October 11, 2005.
- [16] Tomoiagă B, Chindriş M, Sumper A, Sudria-Andreu A, Villafafila-Robles R. Pareto Optimal Reconfiguration of Power Distribution Systems Using a Genetic Algorithm Based on NSGA-II. Energies. 2013; 6(3):1439-1455.

# ANALYZING VARIOUS CORNER BASED METHODS

Sarabjeet Kaur,  
Student Department of Computer Engineering  
Punjabi University, Patiala  
sarabmaan1786@gmail.com

Er. Sumandeep Kaur  
Assistant Professor, Department of Computer Engineering  
Punjabi University, Patiala  
sumandhanjal@gmail.com

**Abstract--** This paper is based on the different corner detectors used for the purpose of corner detection. This gives the brief description of various corner detectors. Each corner detector has a different way of execution. Various corner detectors discussed are Moravec corner detector, Harris Detector, SUSAN corner detector, FAST corner detector. Based on different papers a comparison is made showing the differences among the studied detectors.

**Keywords--** Corners, Corner detection, Corner detectors: Moravec, Harris, SUSAN, FAST.

## I. INTRODUCTION

Corner in a simple language can be defined as the intersection point of two edges. These are the points having two different edge directions in the surrounding area of point. Corners have high curvature and lie at the intersection of different brightness regions of images. Corners play an important role in image segmentation. Corners are the points having high intensity changes in all directions. Hence corners play an essential role in detecting features and edges in an image and also help in matching the patterns of different images.

Corner detection plays an important role in research related to image processing. Different corner detectors are there to detect the corners. Extraction of corners helps in minimization of processing without any loss of important information.

## II. CORNER DETECTOR REQUIREMENTS

Any corner detector need to possess number of characteristics such as:

- 1) This should detect all the true corners and avoid false corner detection.
- 2) This should have high detection rate for providing bigger starting set.
- 3) This should provide efficient computation.
- 4) This should provide the least possible execution time with best results.

- 5) The repeatability rate should be high.
- 6) This should be robust against noise.

### III. TYPES OF CORNER DETECTORS

There are many different corner detectors used for corner detection. Some of the corner detectors are listed below:

#### A. *Moravec corner detector*

It is one of the earliest corner detection algorithms developed by Hans P. Moravec [13] in 1977. Moravec took into account the idea of “interest points” that are the points which help in the detection of similar regions. Moravec operator works on the basis of local window in the image which helps in determining the change in the intensity upon shifting the window by small size in different directions. The calculations for intensity variation for a given shift are made by taking the sum of squares of intensity differences of corresponding pixels in these two windows. The similar operation is carried out for different pixels and on the basis of this the pixels showing large intensity variation in every direction are considered as corners. Pixels of the nearby regions have uniform intensity and therefore look similar.

According to Moravec, the main problem with this operator is that it is isotropic as sometimes the edge gets incorrectly chosen as an interest point because of the presence of edge that is not present in the direction of neighbours. Sometimes the corner detected may be an isolated point due to which the Moravec corner detector is considered vulnerable to noise. However using larger window size leads the algorithm to become more potent to the noise since the intensity variation of the true corner is larger than the isolated pixel. Moravec operator makes use of square window which do not provide enough good results. Moravec operator cannot be applied around the border as the window may "fall" off the image. Therefore it ignores the pixels around the border.

Moravec's corner detector is just a simple algorithm that was used by Moravec and others, but now is lesser in use. This is generally used for the false corners and hence also considered as sensitive to noise. However, for Moravec it proved to be enough useful and efficient as his work was based on a real-time application. Generally this may be used to detect similar regions.

#### B. *Harris Detector*

Harris Detector was given by C.Harris and Mike Stephens in 1988 [3] when they felt the need of getting richer information about the surfaces and objects and hence they once again took into account the Moravec detector [13]. This detector overcame the problems being faced by the Moravec detector and hence came up with corrective measures. Instead of using square window as in Moravec detector it uses Gaussian window. This detector efficiently distinguishes between the true and false corners. Harris detector is based on local auto-correlation function [3]. It is also a combined edge and corner detector.

In 1994, Shi and Tomasi [9] came up with a new method based on Harris and Stephens operator which relies on the calculation of the eigenvalues as well as eigenvectors of a small region. And from these calculated values two of the largest Eigen values are used for calculating some functions. Finally, the calculated function value and a threshold are used for corner detection. Harris and Stephens took into

account the differential of the corner score with respect to direction directly, rather than using shifted patches and thus, improved upon Moravec's corner detector.

Harris detector considers all the basic shifts in the window instead of 45 degree shift as in Moravec detector and hence for all the small shifts [u,v] there is a bilinear approximation:

$$E \ u, v \cong u, v \ M \begin{bmatrix} U \\ V \end{bmatrix}$$

where M is 2x2 matrix computed from image derivatives:

$$M = \begin{matrix} & w(x,y) & I_x^2 & I_x I_y \\ x, y & & I_x I_y & I_y^2 \end{matrix}$$

The eigenvalues of M are analyzed. Let the two eigenvalues are  $\lambda_1$  and  $\lambda_2$ . This characterization can be indicated in the following way: If M is having two "large" eigenvalues then it shows an interest point. On the basis of the magnitudes of the eigenvalues, there are following inferences:

- 1) If  $\lambda_1 \approx 0$  and  $\lambda_2 \approx 0$  then the pixel (x,y) has no features of interest.
- 2) If  $\lambda_1 \approx 0$  and  $\lambda_2$  has some large positive value, then there exists an edge.
- 3) If  $\lambda_1$  and  $\lambda_2$  have large positive values, then there exists a corner.

Since identical computation of the eigenvalues is computationally too high, and also requires the computation of roots of square. Thus Harris replaced the use of eigenvalues with the following function R, where k is empirical constant:

$$R = \det M - k (\text{trace } M)^2$$

$$\text{where } \det M = \lambda_1 \lambda_2$$

$$\text{trace } M = \lambda_1 + \lambda_2$$

Therefore, the algorithm does not need to actually calculate the decomposed matrix consisting of eigenvalues and eigenvectors for the matrix M and instead it is sufficient to evaluate the determinant and trace of M to find corners, or just the interest points.

Further Shi-Tomasi [9] improved this by directly calculating out the  $\min(\lambda_1, \lambda_2)$  because under certain conditions, the corners are more durable for tracking.

The biggest shortcoming of Harris detector is that it consumes lot of time for corner detection. This detector may find use for the purpose of identification of objects, scenes in videos, content-based image retrieval, or model-based recognition.

Hence based on the comparison of [7, 20] it can be said that this detector has high detection as well as high repeatability rate. This performs quite well for affine transformations though it might need larger number of resources for feature detection.

#### C. The SUSAN corner detector

SUSAN denotes the smallest univalue segment assimilating nucleus. SUSAN was given by S.M. Smith [17] in 1997. SUSAN corner detector is realized by a circular mask.

The area of mask having the same brightness as the nucleus can be defined only if the brightness of each pixel within a mask is correlated with the brightness of that mask's nucleus. This area of the mask

is known as the “USAN”, which denotes “Univalue Segment Assimilating Nucleus” [17]. Hence the corner detection can be done according to the area of USAN. Whenever the area of USAN is the smallest it shows that the nucleus lies on the corner. The SUSAN detector makes the use a circular mask.

For feature detection, using SUSAN a circular mask is placed over the pixel to be tested i.e. the nucleus. The region of the mask is  $M$ , and a pixel present in mask is represented by  $m \in M$ . The nucleus is present at  $m_0$ . Each and every pixel is compared to the nucleus using the given comparison function:

$$c_m = e^{-\frac{|I_m - I_{m_0}|^6}{t}}$$

where  $t$  determines the radius,  $I$  denotes the brightness of the pixel and the power of the exponent needs to be determined empirically.

For corner detection, Firstly, the centroid of the SUSAN is found. Usually a proper corner has the centroid lying far from the nucleus. The second step requires that all points on the line from the nucleus going through the centroid to the edge of the mask are coming in the SUSAN.

SUSAN corner detector does not require derivative. That is why it can work well when the noise is present. This corner detector relies on the usage of a mask to count the number of pixels having the brightness similar as the centre pixel. A comparison is made to find out the number of pixel having the same brightness as the centre pixel with a threshold, and thus the detector determines whether the centre pixel is a corner or not. This detector may be used for suppression of ringing artifacts, finding edges and corners.

Smith and Brady [17] claim that the SUSAN corner detector performs well even in the presence of noise as it does not need to calculate image derivatives and hence does not amplify noise.

#### D. FAST

FAST denoting Features from Accelerated Segment Test is an algorithm recommended primarily by Rosten and Drummond in 2005 for identification of the interest points in an image. An interest point in an image may be considered as a pixel having a well-defined position and which can be detected robustly. These interest points possess high local information content. For the best results these interest points are required to ideally repeat between different images.

FAST algorithm plays an important role as an interest point detector for the real time frame rate applications like SLAM (simultaneous localisation and mapping) being used on a mobile robot, which has very less computational resources. It helps in locating different objects in the corners.

In 2006 Rosten and Drummond [6] with help of machine learning enhanced a simple and repeatable segment test into the FAST-9 detector which provides better processing speed as well as excellent repeatability. Hence it leads to the FAST-ER detector being computationally effective and giving better results than the earlier one.

FAST is considered as a detector with good speed along with high quality feature detection. This detector has a high performance rate. However it is not robust to high levels noise and is dependent on a threshold. Sometimes numbers of features are detected which lies adjacent to one another.

Based on the studies of different paper we reach the following.

TABLE I  
 REVIEW BY DIFFERENT AUTHORS

Sr. No.	Author	Detector	Advantages	Disadvantages
1.	Hans P. Moravec[13]	Moravec	This helps in detection of the similar regions [13].	This operator is isotropic and also sensitive to noise [13].
2.	D. Parks[4]	Moravec	Moravec operator detects all of the corners [4].	D. Parks found that, this assigns large cornerness to diagonal edges [4].
3.	Shraddha S. Aher[1]	Harris	Harris detector proves to be better than SUSAN detector when compared in terms of complexity, stability, execution time [1].	
4.	D. Parks[4]	Harris	This performs excellent for affine transformations using isotropic gradient.	This detector gives poor localization at certain corners [4].
5.	Shraddha S. Aher[1]	SUSAN	SUSAN does not require derivative therefore, it works well even in presence of noise [1].	This uses a fixed global threshold instead of an adaptive threshold. The anti-noise capability of SUSAN detector is worse than Harris detector [1].
6.	S.M. Smith[17]	SUSAN	This allows image edges, lines, corners and junctions to be found quick and accurate [17].	
7.	Jie Chen[10]	SUSAN		The corner detector needs an improvement in adaptive threshold as well as the shape of mask. The robustness of the algorithm needs to be strengthened [10].
8.	E. Rosten[6]	FAST	It is many times faster compared to other corner detectors [6].	This is threshold dependent and also not robust to high level noise [6].
9.	Khairulmuzzamil Saipullah[11]	FAST	FAST has the best average performance in the terms of speed, number of key point and repeatability error. For the purpose of object detection in an embedded device, this achieves the real-time performance [11].	However this is insensitive to orientation and illumination change to some extent also the object detection performance of this is not much good as compared to other object detection methods [11].

#### IV. CONCLUSION

This paper discusses the different corner detectors each having different technique to detect the corners.

Moravec's corner detector is one of the earliest algorithms which find a least use in modern day applications. Harris detector can be considered far more efficient than the other corner detectors since it is able to detect corners as well as edges. SUSAN corner detector is fast as well as has a better repeatability rate though it is susceptible to noise. FAST provides good speed and has a high performance rate. Hence it can be said that each detector has its own pros and cons.

#### References

- [1] Aher, S. S. (2015). Analysis of different algorithms on edge and corner detection. International Journal of Multidisciplinary Research and Development, 2(6).
- [2] Bostjan Marusic, P. S.(2006). Application of the SUSAN filter to wavelet video-coding artifact removal. International Journal of Electronics and Communication, 56-64.
- [3] C.Harris, M. (1988). A Combined Corner and Edge Detector. Alvey Vision Conference.

- [4] D.Parks. Corner Detection.
- [5] Drummond, E. R. (2006). Machine Learning for High Speed corner detection. 9th Euproean Conference on computer vision , 430-443.
- [6] E. Rosten, T. D. (2006). Machine Learning for High - Speed Corner Detection. Lecture Notes in Computer Science.
- [7] F. Mohanna, F. Mokhtarian,Performance Evaluation of Corner Detection Algorithms under Similarity and Affine Transforms, Proceedings of British Machine Vision conference(2001).
- [8] Gao, C. (2012). Analysis and improvement of SUSAN algorithm., (pp. 2552-2559).
- [9] J. Shi, C. T. (1994). Good Features to Track. Proceedings of IEEE Conference Computer Vision and Pattern recognition .
- [10]Jie Chen, L.-h. Z.-h. (2009). The Comparison and Application of Corner Detection Algorithms. Journal of Multimedia .
- [11]Khairulmuzzammil Saipullah, N. A. (2013). Comparison of Feature Extractors for Real-time Object Detection on Android smartphone. Journal of Theoretical And Applied Information Technology, 135-142.
- [12]Luo, Z. (2013). Survey of Corner Detection Techniques in Image Processing. International Journal of Recent Technology and Engineering .
- [13]Moravec, H. (1977). Towards Automatic Visual Obstacle avoidance. 5th International Joint Conference on Artificial Intelligence .
- [14]Nestor Morales, J. t. (2011). Real-Time adaptive Obstacle Detection based on an Image Database. Computer Vision and Image Understanding .
- [15]Nilanjan Dey, P. N. (2012). A Comparative Study between Moravec and Harris Corner Detection of Noisy Images Using Adaptive Wavelet Thresholding Technique. International Journal of Engineering Research and Applications (IJERA) , 599-606.
- [16]Patel, T. P. (2014). Corner Detection Techniques: An Introductory Survey. IJEDR .
- [17]S.M. Smith, J. B. (1995). SUSAN - A New Approach to Low Level Image Processing.
- [18]Thareja, M. (n.d.). Performance Analysis of Edges, Corners and the genres: A Subjective Estimation. IOSR Journal of Electronics and Communication Engineering .
- [19]Yaoyu Cheng Shuxian Zhang, Y. H.-Y. Edge Detection Algorithm Based on SUSAN Operation on Auto Hub Image. 9th WSEAS International Conference on Multimedia Systems and Signal Processing.
- [20]Z. Zheng, H. Wang, Analysis of Grey Level Corner Detection, Pattern Recognition letters (1999).

# An Edge Detection Technique Using Bacterial Foraging Optimization Method

Harneet Kaur<sup>1</sup>

M.tech(Student)

Baba Banda Singh Bahadur Engineering

College Fatehgarh sahib, Punjab, India

[neetutung88@gmail.com](mailto:neetutung88@gmail.com)

Ishpreet Singh<sup>2</sup>

Assistant Professor

Baba Banda Singh Bahadur Engineering

College Fatehgarh sahib, Punjab, India

[ishpreet.virk@bbsbec.ac.in](mailto:ishpreet.virk@bbsbec.ac.in)

---

## ABSTRACT

Edge detection is very important field of image processing as well as in medical field. It helps in detecting various cracks in the body. In this research work, medical images i.e. X-rays are used to find the crack in the body. Discrete Wavelet Transformation (DWT) is used to divide the image into four sub-levels i.e. HH, HL, LL and LH. Bacteria Foraging Optimization (BFO) is used to get reduced feature set. The whole simulation is implemented in MATLAB 2010 environment. The whole simulation is tested using various parameters like Structural Similarity Index (SSIM), Normalized Absolute Error (NAE), Structural Content (SC), Entropy, Precision and Recall.

**Keywords:** Edge Detection, DWT, BFO, SSIM, NAE and SC.

---

## I. INTRODUCTION

These days computerized cameras are positively the most utilized gadgets to take the pictures. They are all over the place, including cell telephones, pocket PCs or palmtop PCs, robots and home security frameworks [1, 2, 3]. There is most likely the nature of the pictures got by advanced cameras, paying little heed to the setting in which they are utilized, has enhanced altogether since early days. Various problems that exist in pictures while capturing are as shown below;

1. Contrast deformities,
2. Chromatic deviations,
3. Vignetting (i.e., a decrease of a picture shine or immersion at the fringe contrasted with the picture focus)  
Geometrical contortions,
4. Shading demosaicing and
5. Centre imperfections [4, 5].

In day to day life many such cases occurs in which minute hair line fracture may get un-noticed in the X-ray by the Doctors in such cases Bone fracture detection using image processing will help the doctor to avoid such errors. Digital image processing is an expanding area with application regarding to our daily lives, especially in progressive transmission of images video coding, digital libraries image database, remote sensing, and other image database, remote sensing, and other analysis techniques have been developed to aid the interpretation of remote sensing images and to extract as much information as possible from the image [6, 7]. The huge collection of digital images are collected due to the improvement in the digital storage media, image capturing devices like scanners, web cameras, digital cameras and rapid development in internet. This leads to rapid and efficient retrieval of these images for visual information in different fields of life like medical, medicine, art, architecture, education, crime preventions, etc [8, 9].

## II. EDGE DETECTION

Edge detection is one of the most frequently used techniques in digital image processing. Edge detection process has three steps: filtering, enhancement and detection. Images may be affected by different types of noise. The most widely studied two types are the white noise, impulse noise and “salt and pepper” noise. To reduce the influence of noise a filtering step (for example Gaussian filtering) is necessary before edge detection. Enhancement techniques have the role of improving the quality of a digital image. Enhancement is usually performed by computing the gradient magnitude. Detection methods are used to determine which points are edge points or not. Usually, thresholding provides the criterion used for detection. The most frequently used edge detection methods are: Sobel edge detection, Prewitt edge detection, Roberts edge detection, Laplacian of Gaussian (LoG) edge detection and Canny edge detection [10, 11].

Edge detection can be done using:

- a) First order derivative
- b) Second order derivative

First order derivatives in an image are computed using the Gradient and second order derivatives are obtained using the Laplacian.

### A. GRADIENT EDGE DETECTORS

The first derivative assumes a local maximum at an edge. For a gradient image  $f(x, y)$ , at location  $(x, y)$ , where  $x$  and  $y$  are the row and column coordinates respectively, we typically consider the two directional derivatives. The two functions that can be expressed in terms of the directional derivatives are the gradient magnitude and the gradient orientation. 24 The gradient magnitude is defined by

$$G = |G_x| + |G_y| \quad (1)$$

This quantity give the maximum rate of increase of  $f(x,y)$  per unit distance in the gradient orientation of  $\nabla f$ . The gradient orientation is also an important quantity. The gradient orientation is given by

$$G = \tan^{-1} | G_x | + | G_y | \quad (2)$$

where the angle is measured with respect to the x- axis. The direction of the edge at  $(x, y)$  is perpendicular to the direction of the gradient vector at that point.

#### B. ROBERTS EDGE DETECTOR

The calculation of the gradient magnitude and gradient magnitude of an image is obtained by the partial derivatives  $\frac{\partial f}{\partial x}$  and  $\frac{\partial f}{\partial y}$  at every pixel location. The simplest way to implement the first order partial derivative is by using the Roberts cross gradient operator.

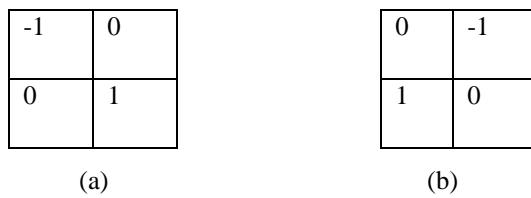


Fig. 1 Roberts Edge Detector

#### C. PREWITT EDGE DETECTOR

The Prewitt edge detector is a much better operator than the Roberts operator. This operator having a 3x3 masks deals better with the effect of noise. An approach using the masks of size 3x3 is given by considering the below arrangement of pixels about the pixel  $[i, j]$

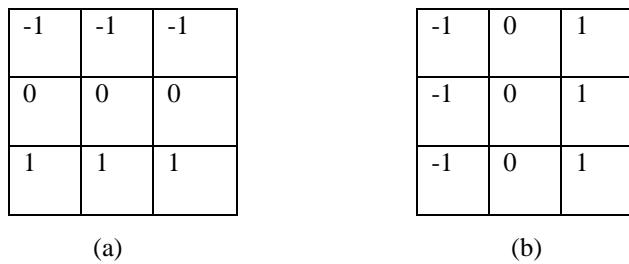


Fig.2 Prewitt Edge Detector

#### D. SOBEL EDGE DETECTOR

It is used in practice for computer digital gradients. The Prewitt masks are simpler to implement than the Sobel masks, but the latter have slightly superior noise suppression characteristics.

-1	-2	-1
0	0	0
1	2	1

(a)

-1	0	1
-2	0	2
-1	0	1

(b)

Fig. 3 Sobel Edge Detector

#### E. The Laplacian Operator

In the Laplacian method we calculate the second derivative of the signal and the derivative magnitude is maximum when second derivative is zero. In short, Laplacian method searches for zero crossings in the second derivative of the image to find edges. An edge map detected from its original image contains major information, which only needs a relatively small amount of memory space to store. The original image can be easily restored from its edge map.

The Laplacian  $L(x,y)$  of an image with pixel intensity values  $I(x,y)$  is given by:

$$L(x, y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \quad (3)$$

Since the input image is represented as a set of discrete pixels, we have to find a discrete convolution kernel that can approximate the second derivatives in the definition of the Laplacian. Two commonly used small kernels are shown in Figure 4 .

0	-1	0
-1	4	-1
0	-1	0

(a)

-1	-1	-1
-1	8	-1
-1	-1	-1

(b)

Fig. 4 The Laplacian

### III. BACTERIAL FORAGING OPTIMIZATION (BFO)

Bacteria Foraging Optimization Algorithm is a new algorithm to the nature inspired family of optimization-algorithms [12, 13]. For over the last five eras, some optimization-algorithms such as GAs, Evolutionary-Programming, Evolutionary-Strategies that draw their stimulus from evolution as well as natural-genetics, which have been controlling the realm of optimization-algorithms. BFOA is a new-group of geographically-confident-stochastic-international search-technique based on impersonator the foraging-behavior of E. coli bacteria. This technique is utilized for locate, handling, and ingesting the food. During foraging, a bacterium can exhibit two

different actions: tumbling or swimming. The tumble action modifies the compass reading of the bacterium. During swimming means the chemo taxis step, the bacterium will move in its recent direction. Chemo taxis movement is continuous until a bacterium goes in the direction of positive nutrient rise. After a definite number of complete swims, the best halves of the inhabitants undergo the original and eliminate the rest of the population. Algorithm of BFO

Step1: Initialize parameters  $p, S, N_c, N_s, N_r, N_{ed}, P_{ed}C(i)$  ( $i=1, 2, \dots, S$ ),  $\theta^i$ .

Step 2: Elimination dispersal loop = l+1

Step 3: Reproduction loop = k+1.

Step 4: Chemo taxis loop = j+1.

[a] For  $i=1, 2, \dots, S$  take a chemotactic step for bacterium I as follows.

[b] Compute fitness function  $(i, j, k, l)$ .

$$\text{Let } J(i, j, k, l) = J(i, j, k, l) + J_{cc} (\theta^i(j, k, l), P(j, k, l)).$$

[c] Let  $J_{last} = J(i, j, k, l)$  to save this value since we may find a better cost via a run.

[d] Tumble: generate a random vector  $\Delta(i) \in R^P$  each element  $\Delta_m(i), m=1, 2, \dots, p$ , a random number on [-1, 1].

[e] Move: Let

$$\theta^{i+1, k+1} = \theta^{i, k, l} + C(i) \frac{\Delta(i)}{\Delta^T i \Delta(i)}$$

This results in a step od = f size C (i) in the direction of the tumble for Bacterioum i.

[f] Compute  $J(I, j+1, k, l)$  and let

$$J(i, j+1, k, l) = J(i, j, k, l) + j_{cc} \square^i j+1, k, l, P(j+1, k, l))$$

i. Let  $m=0$

ii. While  $m < N_s$

Let  $m=m+1$

$$\text{If } j(I, j+1, k, l) < J_{last}, \text{ let } J_{last} = J(I, j+1, k, l) \text{ let, } \theta^{i+1, k+1} = \theta^{i, k, l} + C(i) \frac{\Delta(i)}{\Delta^T i \Delta(i)}$$

To count, the new  $J(i, j+1, k, l)^i, \theta^i(j+1, j, k)$  is used as in [f].

iii. Else, let  $m=N_s$ . This is the end of the while statement.

[h]. Go to next Bacterioum (i+1) if  $i \neq S$ .

Step 5: If  $j < N_c$ , go to step 4

Step 6: Reproduction:

[a]. For the given  $k$  and  $l$ , used for every  $h$   $i=1, 2, \dots, S$ . take  $J_{health}^i = \sum_{j=1}^{N_c+1} J(i, j, k, l)$

Be the health of bacterium i. Sort bacteria and Chemo tactic parameters  $C(i)$  in order of ascending cost  $J_{health}$ .

[b]. The  $S_r$  bacteria with the highest  $J_{\text{health}}$  values die and the remaining  $S_r$  bacteria with the best values split.

Step 7: if  $k < N_{\text{re}}$ , go to step 3.

Step 8: Elimination-dispersal: For  $i=1,2,\dots,S$  with probability  $P_{\text{ed}}$ , eradicate and scatter every bacterium. For this, a bacterium need to be eliminated; so scatter one more to an arbitrary position on the optimization sphere. Go to step 2 if  $l < N_{\text{ed}}$  else end.

#### IV. PROPOSED METHODOLOGY

X-ray medical imaging plays a vital role in diagnosis of bone fracture in human body. The X-ray image helps the medical practitioners in decision making and effective management of injuries. In order to improve diagnosis results, the stored digital images are further analyzed using medical image processing. The most common ailment of the human bone is fracture. Bone fractures are nothing but the cracks which occur due to accidents. There are many types of bone fractures such as normal, transverse, comminuted, oblique, spiral, segmented, avulsed, impacted, torus and greenstick. So, this work mainly proposed the computer aided diagnosis of bone fracture detection in X-ray images using edge detection operators in addition with DWT and BFO optimization algorithm and the proposed model's performance is evaluated using various metrics like SSIM, NAE and SC.

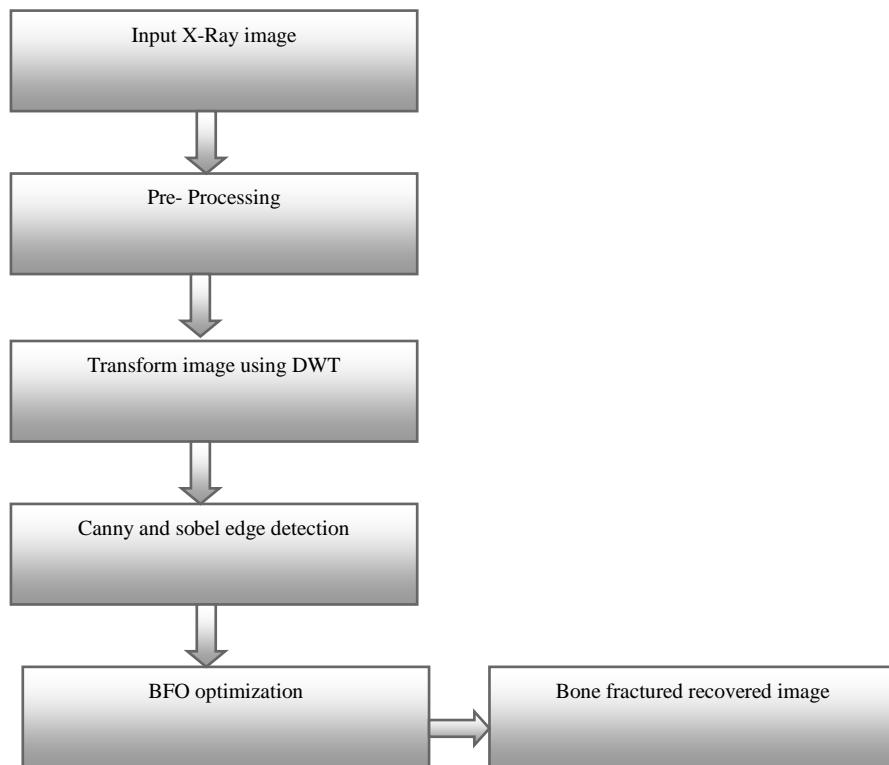


Fig. 5 Proposed Flow chart

Step: 1      Upload X-ray image.

Step: 2      2D DWT decomposition separates an image into the four parts (HL, LL, LH, HH), each of them contains different information of the original image. Detail coefficients represent edges in the image, approximation coefficients are supposed to be a noise. A proper modification of approximation coefficients is the easiest way for edge detection.

Step: 3      Then the principle of the simplest method of edge detection is based on replacing of all approximation coefficients by zeros using canny and sobel. This modification removes low frequencies from the image. The image is reconstructed using only the remaining wavelet coefficients. By means of this method the most expressive edges are found.

Step: 4      The image is reconstructed from remaining coefficients and from modified approximation coefficients using BFO. This method provides sufficient results.

Step: 5      Find SSIM, NAE and SC.

## V. PROPOSED EDGE DETECTION ALGORITHM

---

```
global Fimg Rimg Eimg

set(handles.dip,'string','Please wait...');

soi = size(Fimg);

thresh = 0.1;

Bimg = im2bw(Fimg,thresh);

[LL,LH,HL,HH] = dwt2(Bimg,'haar');

pbst = bfo(Fimg);

[rh ch] = size(HH);

for i = 1:rh

    for j = 1:ch

        HH(i,j) = min(min(pbst)).* zeros(1);

    end

end

Rimg = uint8(idwt2(LL,LH,HL,HH,'haar',soi));

usr = input('Select 1 for Sobel 2 for Log & 3 for Canny : ');

if usr == 1
```

```

type = 'sobel';

elseif usr == 2

type = 'log';

elseif usr == 3

type = 'canny';

end

```

---

## VI. RESULTS AND DISCUSSION

### A. Parameters

#### SSIM

SSIM is used for measuring the similarity between two images. The SSIM index is a full reference metric; in other words, the measurement or prediction of image quality is based on an initial uncompressed or distortion-free image as reference. SSIM is designed to improve on traditional methods such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE), which have proven to be inconsistent with human visual perception.

$$\text{SSIM} = \frac{(2m_1m_2+c_1)(2s_1s_2+c_2)}{m_1^2+m_2^2+c_1(s_1^2+s_2^2+c_2)} \quad (4)$$

Where  $m_1$  and  $m_2$  average of row data and column data respectively,  $s_1$  and  $s_2$  variance of row data and column data respectively,  $c_1$  and  $c_2$  are variables to stabilize the division with weak denominator.

In proposed work we obtained SSIM value:

$$\text{SSIM} = \frac{(2m_1m_2+c_1)(2s_1s_2+c_2)}{(m_1^2+m_2^2+c_1)(s_1^2+s_2^2+c_2)} = 57.65 \quad (5)$$

#### NAE

Normalized Absolute Error is on same scale of data being measured. This is known as a scale-dependent accuracy measure and therefore cannot be used to make comparisons between series on different scales.

$$\text{NAE} = \sum \sum [(O_i - E_i)/O_i] \quad (6)$$

In proposed work we obtained NAE value:

$$\text{NAE} = \sum \sum [(O_i - E_i)/O_i] = 0.135 \quad (7)$$

### *Structural Content*

Structural Content is the ratio of image pixels to find the how many pixels are losses during the operation on image.

$$SC = \frac{(O_i)^2}{(E_i)^2} \quad (8)$$

In proposed work we obtained SC value:

$$SC = \frac{(O_i)^2}{(E_i)^2} = 21.66 \quad (9)$$

### *Entropy*

The entropy is a valuable tool to measure the richness of the details in the output image. By using given formula we calculate the entropy of original and enhanced image.

$$\text{Entropy} = \text{sum}(p.*\log_2(p)) \quad (10)$$

In proposed work we obtained entropy value for original image is:

$$\text{Entropy} = \text{sum}(3.65*\log_2(3.65)) = 6.758 \quad (11)$$

In proposed work we obtained entropy value for enhanced image is:

$$\text{Entropy} = \text{sum}(0.094.*\log_2(0.094)) = 0.320 \quad (12)$$

### *Precision*

In proposed work we obtained precision value:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (13)$$

$$\text{Precision} = \frac{255}{255+1} = 0.99 \quad (14)$$

### *Recall*

In proposed work we obtained recall value:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (15)$$

$$\text{Recall} = \frac{255}{255+65.91} = 0.795 \quad (16)$$

### *Accuracy*

Accuracy is a general term used to describe how accurate a biometric system performs. Its formula is given as:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \times 100 \quad (17)$$

$$\text{Accuracy} = \frac{255+0.0849}{255+1+0.0849+65.91} \times 100 = 79.29 \quad (18)$$

### *B. Database*



Fig. 6 Sample of Database

Above images has been utilised for running and implementation of the algorithm.

### *C. Results*

Below Table shows the values of 7 different parameters for obtained enhanced output image.

TABLE.1 Proposed Parameters I

Image no.	Entropy for original image	Entropy for enhanced image	SSIM	NAE	SC
1.	6.758	0.320	0.515	0.135	21.66
2.	4.64	2.81	0.628	0.618	112.91
3.	6.96	0.124	0.223	0.496	71.88
4.	5.38	0.105	0.536	0.817	33.74
5.	6.26	0.302	0.523	0.814	16.67

TABLE.2 Proposed Parameters II

Image no.	Precision	Recall	Detection Accuracy
1.	0.99	0.795	79.29
2.	0.97	0.756	84.53
3.	0.98	0.753	76.38
4.	0.99	0.734	78.43
5.	0.96	0.723	86.23

## VII. CONCLUSION AND FUTURE SCOPE

A combination of Histogram Equalization and Bacteria Foraging optimization algorithm based edge detector has been developed and implemented to produce better edge detection results than traditional detectors. Such that BFO algorithm is used to choose the optimal value for canny edge detector, Sobel and log edge detector to produce more accurate and satisfactory edge detection results.

Results of the proposed edge detector and the edge detectors based on Entropy, SSIM, NAE and SC method were presented both qualitatively and quantitatively. From the qualitative and quantitative results of edge detectors on medical images, It is concluded that the proposed edge detector clearly outperform all the other methods in study. It has been observed that proposed method obtain optimal results but with high computational effort. So the BFO based technique produced more accurate results than other studied techniques. Finally obtained results indicate that the proposed method have a high effectiveness on a large category of image applications. The effectiveness of method is checked by simulating the test images on MATLAB. The proposed method provides the superior edge detection results to existing edge detection techniques based on the Entropy, SSIM, NAE and SC. In addition, the proposed method provides better quality in visualization by obtaining maximum Precision, Recall and detection accuracy value. Future work lies in the utilization of other wavelet methods.

## REFERENCES

- [1] R. M. Haralick, "Digital Step Edges from Zero Crossing of Second Directional Derivatives", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 1984, pp. 58-68.
- [2] F. Ulupinar and G. Medioni, "Refining Edges Detected by a LoG operator", Computer vision, Graphics and Image Processing, 51, 1990, pp. 275-298.
- [3] D. J. Willians and M. Shan, "Edge Contours Using Multiple Scales", Computer Vision, Graphics and Image Processing, 51, 1990, pp. 256-274.
- [4] M. Roushdy, "Comparative Study of Edge Detection Algorithms Applying on the Grayscale Noisy Image Using Morphological Filter", GVIP, Special Issue on Edge Detection, 2007, pp. 51-59

- [5] J. Canny, "A Computational Approach to Edge Detect", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.PAMI-8, No.6, 1986, pp. 679-698.
- [6] C. H. Li, C.K. Lee, "Minimum cross entropy thresholding", Pattern Recognition, Vol. 26, No. 4, 1993, pp.
- [7] Jin Li, Lei Wang, PeihuaBao, "An Industrial CT Image Segmentation Algorithm Based on Non Estimation", Journal of Computational Information Systems 6:9, 2010, pp. 3103-3109.
- [8] C. H. Li, P. K. S. Tam, " An iterative algorithm for minimum cross entropy thresholding", Letters 19, 1998, pp. 771-776.
- [9] El-Owny, Hassan Badry Mohamed, "Edge Detection in Gray Level Images Based on NonInternational Journal on Computer Science and Engineering (IJCSE), Vol. 5, No. 12, 2013, pp. 932-939.
- [10] El-Owny, Hassan Badry Mohamed, "A novel non edge detection algorithm for noisy images". International Journal of Computer Science and Information Security (IJCSIS), Vol. 11, No. 12, 2013, pp. 8-13.
- [11] K. Padmapriya, and T. K. Bino, "Boundary Detection using Edge Following Algorithm and Enhancement of the Image", International Conference on Computing and Control Engineering (ICCCE 2012), 12-13 April 2012.
- [12] Nitin Kumar Jhankal, DipakAdhyaru "Bacterial foraging optimization algorithm: A Derivative free technique" Institute Of Technology, Nirma University, Ahmedabad,Gujarat IEEE, pp. 1-4, December,2011.
- [13] P.D. Sathya and R. Kayalvizhi "Optimum Multilevel Image Thresholding Based on TsallisEntropy Method with Bacterial Foraging Algorithm" IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, pp. 336-343, September 2010.

# Automatic Parameter Tuning of Support Vector Clustering

Madhurbain Singh, Husanbir Singh Pannu

Thapar Institute of Engineering and Technology, Patiala India

Email: mbs22madhur@gmail.com

**Abstract**—Clustering is an unsupervised technique to group the data according to their mutual similarities. In this research paper we have introduced an improved version of Support Vector Clustering (SVC) technique with automatic parameter tuning. Using Gaussian kernel data points are mapped into a higher dimensional feature space. In Gaussian kernel we look for the minimal enclosing sphere. This sphere, when mapped back to the data space, separates into several components with irregular boundaries, enclosing separate clusters of points. There are two tuning parameters in SVC, soft margin penalty constant and width of the Gaussian kernel which are varied and used to attain smooth cluster boundaries. It is difficult to accurately select both of these parameters manually (by k-fold Cross Validation). So, we have employed an automatic RBF sigma parameter tuning technique which was originally proposed by Cheng Hsuan Li et.al for Support Vector Machines (SVM). It finds out the optimal kernel RBF sigma value for SVC, which is more efficient and accurate than k-fold cross validation method.

**Keywords** - Support Vector Clustering, Gaussian kernel, unsupervised learning, cross validation, support vector data description

## 1. Introduction

A Cluster may be defined as the collection similar type of data points. The data and objects present in the same cluster are all but similar to each other when compared to other clusters. So, the goal is to create homogeneous groups of the data objects.

Major research areas of clustering are exploratory data mining and analyzing data statistically in the field including Machine Learning, recognition of different patterns, analyzing images, retrieval of information, bioinformatics, data compression, and computer graphics. Clustering has an application in anthropology which was derived by Driver and Kroeber in 1932, also A Psychologist Zubin found its application in psychology [1][2]. It was also applied by Cattell in the 1943[3] for trait theory classification in personality psychology.

There are different types of clustering algorithms such as: connection based clustering, centroidal based, density based, distribution based clustering. In connectivity based clustering, the cluster indices are decided based upon relative distances among the objects. It is also known as hierarchical clustering. In centroidal based clustering, clusters are represented by centroidal vectors. For given  $k$  clusters,  $k$  centers are found and then individual data points are assigned respective cluster indices based upon their relative distances from those centroids. Distribution based clustering involves grouping the data objects depending upon their underlying data distribution characteristics. For example an Expectation Maximization (EM) algorithm works well for the Gaussian distribution because it uses Gaussian model for clusters. Density based clustering involves grouping of the data which has more density than rest of the dataset. Density based spatial clustering of applications with noise (DBSCAN) is mostly used algorithm of this category.

### **Internal evaluation**

When the result of clustering is decided by the data used inside the cluster itself is called internal evaluation. This type of clustering is helpful when different clusters are very different from each other. The Davies–Bouldin index can be evaluated by the following formula:

$$DB = \frac{1}{n} \sum_{i=1}^n \max_{i \neq j} \sigma(i) + \sigma(j)/d(c(i), c(j))$$

where  $n$ ,  $c_x$ ,  $\sigma(x)$ ,  $d(c_i, c_j)$  represent the count of clusters, centroid of  $x$ , average distance between all elements in  $x$  to centroid( $c_x$ ) and distance between  $c_i$  and  $c_j$  respectively.

### **External Evaluation**

It is defined as when the result of clustering is evaluated by the data that was not used in clustering. The Rand Index can be computed by the following formula:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

where  $TP$  = True positive,  $TN$  = True Negative,  $FP$ =False Positive,  $FN$  = False Negative

In recent years, a modern clustering technique has been developed, known as Support Vector Clustering (SVC) [4]. It is quite efficient in detecting the clusters with irregular boundaries using non linear mapping of data points into the higher dimensional feature space using Gaussian kernel [2]. In the higher dimensional feature space, the objective is to include maximum data points into the minimum enclosing sphere. This sphere when mapped back into the original space gets split into different irregular boundary contours which enclose the data points known as clusters. The performance of SVC is based on parameters used for Gaussian kernel and soft margin penalty ( $C$ ) [6][9]. Selecting these parameters manually by method of k-cross validation method (CV) is very tiring and for large data sets it is almost impossible to get an ideal result. In this paper an automatic technique for tuning of the RBF kernel function parameters is proposed.

The paper is organized in 6 sections. Section 2 explains about related works, section 3 discusses SVC technique, section 4 explains about proposed automatic tuning for SVC, section 5 explain experimental results and section 6 discusses conclusion and future works.

## **2. Related Work**

In recent years there have been studies on the automatic parameter selection of kernel function parameters [5]. It was used in SVM for automatic parameter tuning successfully. In [8] a clustering approach based upon Information Maximization clustering based upon tuning parameter selection is discussed. It is based upon squared-loss variant of mutual information and eigenvalue decomposition. Another advanced method for automatic parameter tuning is used in [13] in context of SVM in which the RBF kernel function sigma value is calculated through optimization. In this research, we propose this automation for clustering technique for automated tuning of parameters in SVC.

## **3. Support Vector Clustering**

Our main task is to find a sphere of the smallest radius which can contain the given data sets in the higher dimensional kernel feature space mapped by  $\phi(x)$ . The objective is to minimize  $R^2$  where  $R$  is radius of sphere w.r.t. constraints

$$\begin{aligned} R^2 + \epsilon_j - \|\phi(x_j) - a\|^2 &\geq 0 \\ \epsilon_j &\geq 0 \end{aligned}$$

The Lagrange's Multiplier are calculated as follows:

$$L = R^2 - \sum_j (R^2 + \epsilon_j - \|\phi(x_j) - a\|^2) \beta_j - \sum_j \epsilon_j \mu_j + C \sum_j \epsilon_j$$

To find out the dual objective function, first we need to minimize  $L$ . For this we will find derivatives of  $L$  w.r.t.  $R$ ,  $\epsilon$  and  $a$

$$\begin{aligned} L_r &= 2R - \sum_j 2R \beta_j = 0 \\ \sum_j \beta_j &= 1 \end{aligned} \tag{3}$$

$$\begin{aligned} L_\epsilon &= C - \sum_j \mu_j - \sum_j \beta_j \\ C &= \mu_j + \beta_j \end{aligned} \tag{4}$$

$$\begin{aligned} L_a &= 2 \sum_j \|\phi(x_j) - a\| \beta_j = 0 \\ a &= \sum_j \beta_j \phi(x_j) \end{aligned} \tag{5}$$

Complementary conditions=

$$\epsilon_j \mu_j = 0 \tag{6}$$

$$(R^2 + \epsilon_j - \|\phi(x_j) - a\|^2) \beta_j \tag{7}$$

By using Eq. (7) it can be seen that the point  $x_i$  lies outside the feature space only if  $\epsilon_i > 0$  and  $\beta_i > 0$ . From Eq.(6) we can say that  $\mu_i = 0$  as  $\epsilon_i > 0$ , hence from Eq. (5) we can conclude that  $\beta_i = C$ . This is called BSV (Bounded support vector). When  $\epsilon_j = 0$  any point  $x$  would be marked on the surface or inside the feature space. From Eq.(7) it can be said that its image  $\phi(x_i)$  is present on the boundary of the feature space under the constraint  $0 < \beta_i < C$ . This point is known a support vector (SV). So, now we can say that Support Vectors(SV) lie on the boundary whereas the boundary support vectors(BSV) lie outside the boundary and the rest lie inside the boundary line. There exists no BSV's due to constraint 3, provided that  $C \geq 1$ . Using the above relations we eliminate  $R$ ,  $a$  and  $\mu_j$ , and thus convert the Lagrange's form to the Wolfe's Duals

$$W = \sum \phi(x_i) 2\beta_j - \sum \beta_i \beta_j \phi(x_i) \phi(x_j) \tag{8}$$

Using Constraint,

$$0 \leq \beta_j \leq C, j=1, 2, 3, \dots, N \tag{9}$$

From non linear SVM extension we can convert the dot product of  $\phi(x_i) \cdot \phi(x_j)$  by a certain Kernel function represented by  $K(x_i, x_j)$ . We will use the concept of kernel function in this research paper.

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2} \quad (10)$$

where  $\sigma$  is width parameter. The Lagrangian  $W$  is

$$W = \sum_{i=1}^n K(x_i, x_i)\beta_i - \sum_{i=1, j=1}^n \beta_i \beta_j K(x_i, x_j)$$

the image distance of each point  $x$  from the center of the sphere created in the feature space is represented by:

$$R^2(x) = \|\phi(x) - a\|^2 \quad (12)$$

According to Eq.(4) And using kernel function :

$$R^2(x) = K(x, x) - 2 \sum_j \beta_j K(x_j, x) + \sum_{i=1, j=1} \beta_i \beta_j K(x_i, x_j) \quad (13)$$

Sphere's radius is given by  $R$

$$R = \{R(x_i) / x_i \text{ is SV}\} \quad (14)$$

The boundary that encloses the points in the data space is defined by the set

$$\{x / R(x) = R\} \quad (15)$$

They are interpreted by us as forming cluster boundaries. In view of equation (14), SVs lie on cluster boundaries, BSVs are outside, and all other points lie inside the clusters.

### Cluster Assignment

The theory which we have discussed beforehand does not differentiate between points of different clusters. To do this we use a geometrical theory in which we will include  $R(x)$ , now by using the following observation: we have been given a pair of data points which are from distinct clusters; so any of the paths which connect them has to exit from the sphere in feature space. Thus, a pathway containing a segment of the points  $y$  such that  $R(y) > R$  should lie within the sphere. We end up with the definition of the adjacency matrix  $B_{ij}$  among the points  $x_i$  and  $x_j$  whose images lie inside or on the boundary of the sphere in feature space:

$$B_{ij} = 1, \text{ for all values of } y \text{ and on the line segment joining } x_i \text{ and } x_j \text{ when } R(y) \leq R \text{ and} \\ B_{ij} = 0, \text{ Otherwise}$$

Because of this square matrix representation during this SVC's cluster assignment phase, it becomes a bottleneck during the calculations for the space and time complexity. So SVC is efficient to detect quite irregular cluster boundaries for smaller datasets only.

Now we briefly discuss an important existing clustering method known as K-means clustering. This would help to contrast the features of SVC over other conventional clustering methods. In recent times there have been detailed studies on them [5].

### K-means Method

It is very simple unsupervised learning algorithms which help us solve a well-known problem of clustering. In this we first differentiate between a given data set through a definite number of clusters ( $k$  clusters) and then name  $k$  centers, one for each cluster. These centers need to decide in such a way that each location or place causes un-

similar results. It is useful to place them at a maximum possible distance from each other. Firstly, we need select each point from a defined data set connect it to the closest k-center. When all the points have been connected the early part of our job is done. Now, we need to find out new k-centroids as barycenter of the clusters found earlier. Now that we have found the new k-centroids we need to establish a new binding between the data set and the closest centers. As a result a loop was produced. After this we notice that the k-center have changed their location, now as no more changes are done we observe that centers do not move anymore. Finally, using this algorithm we aim to minimize an objective function (squared error function):

$$J(V) = \sum_{i=1}^c \sum_{j=1}^{c_i} \|x_i - v_j\|^2$$

where,

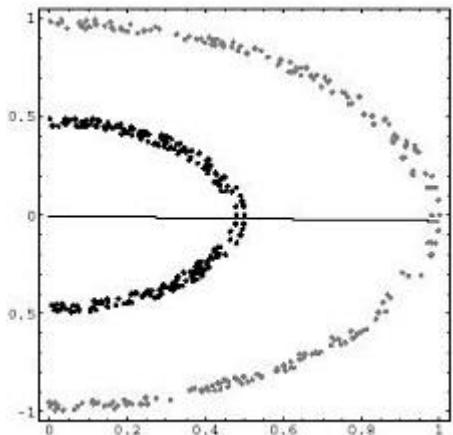
' $\|x_i - v_j\|$ ' is the Euclidean distance between  $x_i$  and  $v_j$ .

' $c_i$ ' is the number of data points in  $i^{th}$  cluster.

' $c$ ' is the number of cluster centers.

### Disadvantages

1. This method of clustering is not robust to non-linear data distribution. It gives different results with different representations. This is represented in the form of coordinates.
2. To learn this algorithm we require apriori specification of the number of  $k$ 's i.e. the cluster centers.
3. Overlapping data cannot be resolved by k-means method.
4. It only is defined for a definite mean which means it will fail for categorized data.
5. A Cluster center cannot be chosen randomly.
6. This algorithm is not applicable for non-linear data set.
7. This algorithm is not able to handle noisy data as well as outliers.



**Fig 2:** This figure shows that k-means method fails for non-linear data set.

Due to the above disadvantages, we choose SVC and employ automatic parameter tuning to make it self-dependable clustering algorithm. It would take care of above limitations and there is accurate selection of Kernel Parameters.

## 8. Proposed Algorithm for Automated Tuning of Kernel Function Parameters

Using this algorithm we propose that one should be able to find out the most ideal parameters for our Kernel Function used in SVC. This theorem was previously used in Support Vector Machines to for automatic tuning of parameters of Kernel Function [5], but in this research paper we try to apply it on SVC.

Let us assume that  $A_i$  is the set of training samples in the set  $i$ , where  $i=0,1,2,\dots,L$ . There are two important properties of the RBF kernel function: (1)  $K(x_i, x_i, \sigma) = 1$  for all  $i=1,2,\dots,n$ , i.e., the norm of every sample in the feature space is 1 and (2)  $0 < K(x_i, x_j, \sigma) \leq 1$  for all  $i,j=1,2,\dots,n$ , i.e., the cosine value of two training samples and in the feature space can be computed by and it determines the similarity between these two samples.

RBF Kernel function

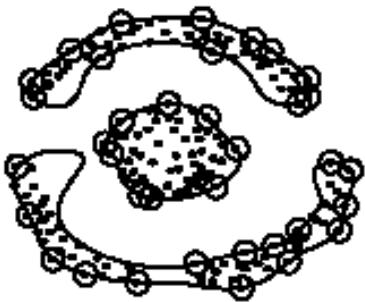
$$w(\sigma) = \frac{1}{\sum_{i=1}^L \|w_i\|^2} \sum_{i=1}^L \sum_{x \in w_i} \sum_{z \in w_i} K(x, z, \sigma)$$

Where  $\|w_i\|$  is the number of training samples in class  $i$ . The parameter  $\sigma$  should be determined such that  $w(\sigma) \leq 1$  and  $w(\sigma) \geq 0$ . Hence, the optimal  $\sigma$  can be obtained by solving the Following optimization problem:  $\min_{\sigma \rightarrow 0} J(\sigma) \equiv 1 - w(\sigma)$

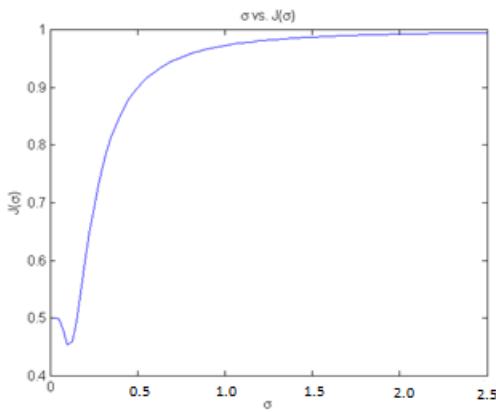
## 9. Experimental Results

Quantitative testing of the proposed technique is performed on an artificial and publically available datasets. The artificial dataset contains 200 points in two dimensions as shown below in Fig 3. Using the parameter automation approach [13] we got three clusters detected as shown in Fig. 3. Values of parameters are listed in the Table 1. Fig.4 shows the optimal RBF sigma calculated through [13] for the artificial dataset.  $\sigma = 0.15$  gives the minimum value of  $J(\sigma)$ .

The experimental results have also been illustrated for IRIS and AAAI 2014 Accepted Papers Data Set available at UCI repository [11]. We have performed the experiment on 10 other standard datasets and obtained results with high accuracy through automatic parameter tuning. The automation of RBF sigma [11] saves a great deal of time complexity as compared to traditional Cross Validation (CV) technique. The overall average clustering accuracy from comes out to be 88% which makes it well suitable to detect clusters of irregular boundaries. K-means algorithm fails to detect such irregular clusters successfully.



**Fig 3:** Artificial dataset with 200 points using proposed technique detected 3 clusters, with RBF sigma = 0.15



**Fig 4:** RBF sigma value calculated through [13] for the artificial dataset

Dataset	N	Sigma	C	Accuracy
Artificial	183	0.15	100	0.921
IRIS	150	0.28	0.1	0.844
AAAI	399	0.45	12	0.898

**Table1:** Clustering results on three datasets

## 10. Conclusion

In this paper, we have proposed an automatic parameter tuning method for Support Vector Clustering. SVC is quite efficient to detect clusters with irregular boundaries compared to other conventional algorithms such as k-means. Kernel parameter in SVC is done using a Chen-Hsuan et al.'s technique which was proposed for SVM. Results show that clusters have been detected with high accuracy of average 89%. SVC although is a good algorithm for clustering, but it uses a square matrix during the cluster assignment phase. So it is not efficient for time and space complexity. In future we are planning to extend SVC for massive datasets by using advanced cluster assignment strategies.

## REFERENCES

1. Bailey, Ken (1994). "Numerical Taxonomy and Cluster Analysis". *Typologies and Taxonomies*. p. 34. ISBN 9780803952591.
2. Tryon, Robert C. (1939). *Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality*. Edwards Brothers.I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963
3. Cattell, R. B. (1943). "The description of personality: Basic traits resolved into clusters". *Journal of Abnormal and Social Psychology* 38: 476–506. doi:10.1037/h0054116.
4. Asa Ben-Hur, David Horn, Hava T. Siegelmann, Vladimir Vapnik; 2(Dec):125-137, 2001.M.
5. Mingyuan Zhao,; Ke Tang,; Mingtian Zhou,; Fengli Zhang,; Ling Zeng, "Model parameter selection of support vector machines", *Cybernetics and Intelligent Systems*, 2008 IEEE Conference on, On page(s): 1095 – 1099
6. An Efficient k-means Clustering Algorithm: Analysis and Implementation by Tapas Kanungo, David M. Mount,Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu.
7. Ortega, Joaquín Pérez, Ma Del Rocío Boone Rojas, and María J. Somodevilla. "Research issues on K-means Algorithm: An Experimental Trial Using Matlab."
8. Wu, Xindong, et al. "Top 10 algorithms in data mining." *Knowledge and information systems* 14.1 (2008):1-37.
9. Teknomo, Kardi. "K-means clustering tutorial." *Medicine* 100.4 (2006): 3.k-means clustering by kechen.
10. Chapelle, Olivier, et al. "Choosing multiple parameters for support vector machines." *Machine learning* 46.1-3 (2002): 131-159.
11. Bache, K., and M. Lichman. "UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. University of California, School of Information and Computer Science." Irvine, CA (2013).
12. Sugiyama, Masashi, et al. "On information-maximization clustering: Tuning parameter selection and analytic solution." *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011.
13. Li, Cheng-Hsuan, et al. "An automatic method for selecting the parameter of the RBF kernel function to support vector machines." *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International. IEEE*, 2010.

# A Comprehensive Review of Sobel Edge Detector using Gray Scale Images

<sup>1</sup>Himanshu Rana, <sup>2</sup>Nirvair Neeru

<sup>1</sup>M.Tech Scholar, Department of Computer Engineering, Punjabi University, Patiala (Punjab), India.  
<sup>2</sup>Assistant Professor, Department of Computer Engineering, Punjabi University, Patiala (Punjab), India.

**Abstract-** Edge in digital image processing is a change in the sharpness present in the digital image. These changes can be the border between a gray color image i.e. either white or black image. The main property of any of the edge detector is to extract the key information hidden in the image that we can't get using simple methods. Edge detection is basically based on the range non-continuity. Various operators are present for the purpose of the edge detection but here we discuss about the one of the prominent edge detector i.e. Sobel Edge Detector. This paper is created to guide the initial researches in the area of the edge detection and researchers can use its information for the purpose of the basic knowledge about the Sobel Edge Detector. Edge detection is used in various fields very prominently like Brain Tumor, MRI etc. The efficiency and the accountability of this operator have been discussed in this paper. All of the experimental results and simulations shown in this paper have been done on MATLAB R2012b. Various images have been taken for the experiment purposes and results of the operations have been discussed.

**Keywords**— *Digital Image Processing, Edge Detection, Noise, Sobel Edge Detector.*

## I. INTRODUCTION

In image detection we can achieve the desired target or we can say that we can get the desired information without affecting the other parts of the image. Sometimes the information present in the image is so crucial that we have to get that information at any cost, but due to lack of the operators present we cannot get that in an accurate and efficient manner. The most basic feature of an image is its edge as it contains the internal information about the image. Therefore the edge detection is one of the key research areas in the field of digital image processing.

Edge is a boundary between the two homogeneous regions and in these regions there will be a sharp change in the intensity. Image processing can solve the two objectives: One, we can create the image that will be suitable for the people to observe and identify. Second, we can also wish that particular image can be easily recognizable by the computing device. The edge is we can say a set of those pixels whose grey have step change and roof change, and they exist between object and object and in some cases in elements also [1]. Edge detection enhances the pictorial information for the human intervention and also provides the best results in real time objects. Various places uses the facility of the edge detection whether it is CCTV in night vision, vehicle detection mechanism installed on the traffic lights and also in case of fractured bones of the patients, it helps us to provide the best and accurate results so as that proper actions can be taken.

Image processing is a form of signal processing for which the input is an image such as the photograph or video frame, the output may be either an image or a set of characteristics or parameters related to the image. Sometimes noise is present in image and that also removed with the help of edge detectors [5].

## II. PRINCIPLE OF EDGE DETECTION

Edge detection in digital image processing is one of the most frequently used techniques. The so-called edge is a collection of pixels whose brightness at the different local areas changes frequently and also its grey value also got changed at roof or step. Detection is an important step in the image processing as with its help we can easily recognize the objects. The edge in an image is reflected as the change in the grey level of the image.

Edge detection helps us in significantly reducing the data that is not of use and also it preserves the important and crucial information. So, the general method for the edge detection is to study those changes of each and every single pixel present in the grey area of the concerned image and then by taking its first-order and second-order derivatives we can detect the edge. This method is generally referred as the edge detection method for local operator.

According to the edge and line present in the image we can get the structure of the object. Therefore, we can say that the edge extraction plays a vital role in image or graphics processing and feature extraction.

Edge basically consists of important and meaningful information and features. Basically all of the edge detectors determine the image's boundary information which represents the interior objects of image. We can define an edge as the borderline pixels that connect the two mutual exclusive regions which are having different values of their luminance [1].

### III. AN EDGE DETECTION MODEL: SOBEL OPERATOR

Compared to other edge operator, Sobel has two main advantages:

1. Since the introduction of the average factor, it has some smoothing effect to the random noise of the image.
2. Because it is the differential of two rows or two columns, so the elements of the edge on both sides have been enhanced, so that the edges seem thick and bright [2].

In the airspace, edge detection is usually carried out by using the local operator. What we usually use are orthogonal gradient operators, directional differential operator and some other operators relevant to second-order differential operator. Sobel operator is a kind of orthogonal gradient operator. Gradient corresponds to first derivative, and gradient operator is a derivative operator. For a continuous function  $f(x, y)$ , in the position  $(x, y)$ , its gradient can be expressed as a vector (the two components are two first derivatives which are along the X and Y direction respectively) [3]:

$$\nabla f(x, y) = [G_x \ G_y]^T = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix} \quad (i)$$

The magnitude and direction angle of the vector are:

$$G_{mag} = \sqrt{(G_x^2 + G_y^2)} \quad (ii)$$

$$\theta(x, y) = \arctan(\frac{G_x}{G_y}) \quad (iii)$$

The partial derivatives of the formulas above need to be calculated for each pixel location. In practice, we often use small area template convolution to do approximation.  $G_x$  and  $G_y$  need a template each, so there must be two templates combined into a gradient operator. The two  $3 \times 3$  templates used by Sobel are showed in Figure A .Every point in the image should use these two kernels to do convolution. One of the two kernels has a maximum response to the vertical edge and the other has a maximum response to the level edge. The maximum value of the two convolutions is used as the output bit of the point, and the result is an image of edge amplitude [1][4].

-1	0	+1
-2	0	+2
-1	0	+1

G<sub>x</sub>

+1	+2	+1
0	0	0
-1	-2	-1

G<sub>y</sub>

Figure A: Mask for Sobel Edge Detector

Each of these masks responds to the edges in horizontal and in vertical directions in its maximum possible ways. These masks can be operated horizontally and vertically to obtain the gradients in the corresponding directions.

### IV. IMPLEMENTATION AND DISCUSSIONS

All of the implementations has been done with the help of the MATLAB software and this software is very useful whenever we are working on the image processing also it provides maximum number of inbuilt functions in it. Every new version provides some new functionality with new launch and makes the work of the researchers very easy. This time the version MATLAB R2012b has been used and all of the operations are performed in very effective way. Many images have been taken and the results of the few have been showed. It is found that when the noise is present in the image sobel edge detector not perform well. Although it gives the edges present in the image but at a large extent and sometimes it becomes difficult to recognize the appropriate edges that are required for the

further processing may get lost. So we can say that the sobel edge detector works well for the images with the less noise and provides the good results. Results of the various images have been showed below which is of both noiseless as well as noiseless images.

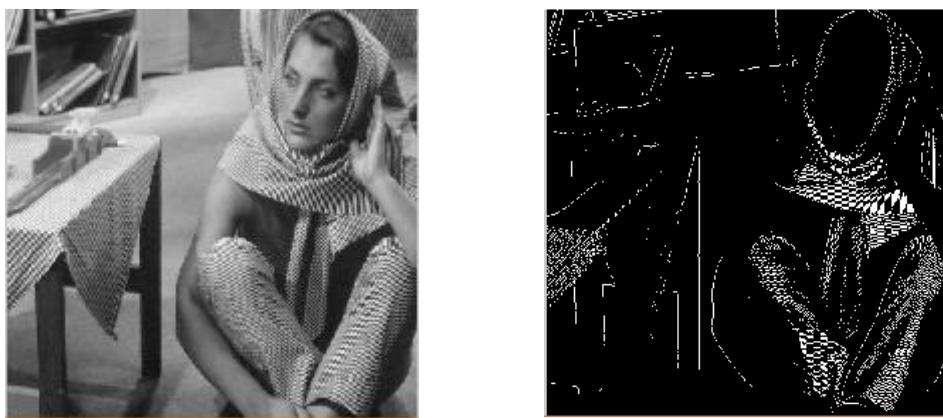


Figure 1. Before and After Sobel Detector Operation (Noiseless)

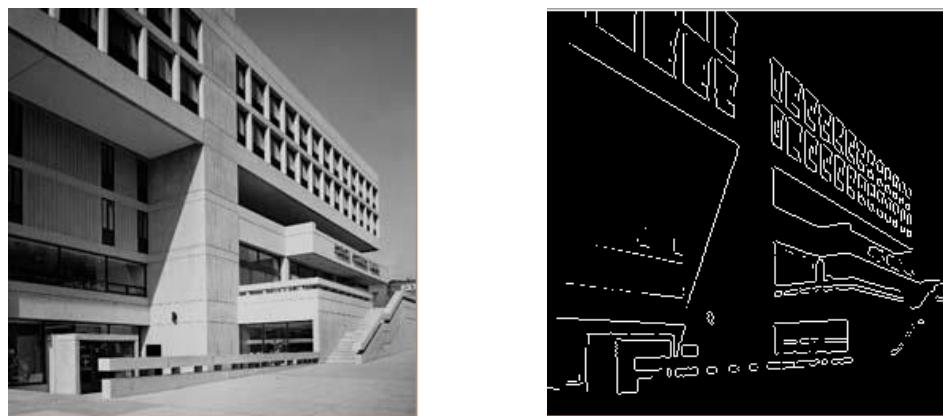


Figure 2. Before and After Sobel Detector Operation (Noiseless)



Figure 3. Before and After Sobel Detector Operation (Noiseless)



Figure 4. Before and After Sobel Detector (Noisy Image)

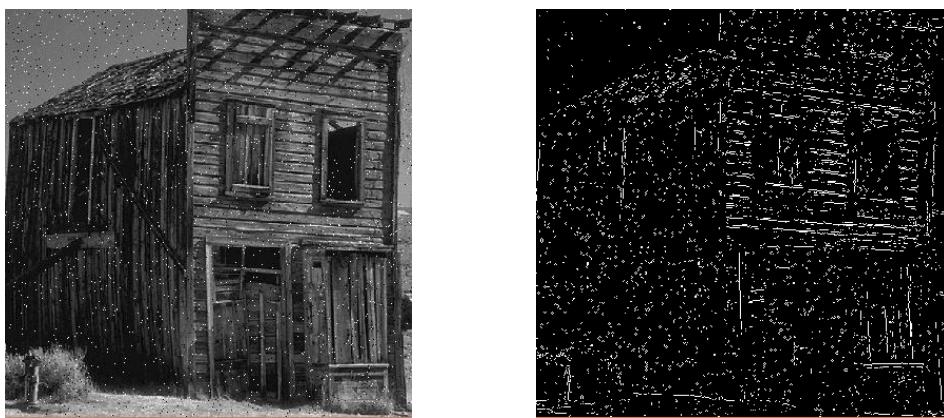


Figure 5. Before and After Sobel Detector (Noisy Image)

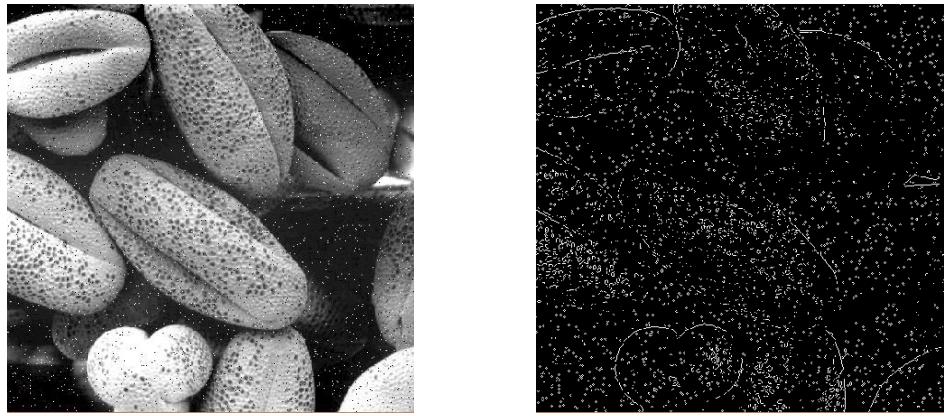


Figure 6. Before and After Sobel Detector (Noisy Image)

#### CONCLUSIONS

Edge detection is one of the most important techniques that has been implemented commonly now a days in the field of image processing. It is used for the image segmentation, and image identification. In this paper we have studied the Sobel Edge Detector to detect the edges and to get the appropriate information. The basic purpose of this review paper is to check the each and every functionality provided in the Sobel Edge Detector and also it can be used to make the new researchers aware about this detector. On the basis of its functions we can say that the

algorithm of this detector is very easy to implement. Sometimes it appears while implementing this that it is sensitive to the noise but in whole it can be removed. As per the operations we found that the noisy images may loss the information as compared to simple images. In future more work can be done on the improvement towards the noisy images because some of them may contain the crucial information that can be used for the further problem solving methods.

#### ACKNOWLEDGMENT

The authors are greatly indebted towards the Department of Computer Engineering, University College of Engineering, Punjabi University, Patiala (Punjab) for providing the excellent lab facilities that made this work possible.

#### REFERENCES

- [1] R.C. Gonzalez and R.E. Woods, “*Digital Image Processing*”, Addison-Wesley Publishing Company, 1992.
- [2] Ajeet Singh et.al, “*Implementing Edge Detection for Medical Diagnosis of a Bone in Matlab*”, 5th ICCICN, IEEE, 2013, 978-0-7695-5069-5/13.
- [3] Wenshuo Gao et.al, “*An Improves Sobel Edge Detection*”, ICICT 2010, 978-1-4244-5540-9/10 IEEE.
- [4] Diplaxmi R. Waghule et.al, “*Overview on Edge Detection Methods*”, ICESSPCT 2014, 978-1-4799-2102-7/14 IEEE.
- [5] C.Pavithra et.al, “*An Efficient Edge Detection Algorithm for 2D-3D Conversion*” ICCPEIC 2014, 978-1-4799-3826-1/14 IEEE.
- [6] Amphy Jose et.al, “*Performance Study of Edge Detection Operators*” ICES 2014, 978-1-4799-5026-3/14 IEEE.
- [7] Kiranjeet Kaur et.al, “*A Survey on Edge Detection using Different Techniques*” IJAIEM Volume 2, Issue 4, April 2013.
- [8] N. Hamid et.al, “*An FPGA implementation of gradient based edge detection algorithm design*” ICCTD ’09, Kota Kinabalu, vol. 2, pp. 165-169, Nov. 2009.
- [9] Sourav Roy et.al, “*An Approach towards Detection of Indian Number Plate from Vehicle*” IJITEE, Vol. 2, Issue 4, March 2013.
- [10] A.E.A. Elaraby et.al, “*A Noval Algorithm for Edge Detection of Noisy Medical Images*” IJSIP, Vol. 6 No. 6. pp. 365-374, 2013
- [11] Isra Javaid Alam et.al, “*Detecting Edges in ana Image with the Help of Fuzzy Parameters*” ICFIT 2014, 978-1-4799-2293-2/13 IEEE.

# Fruit Disease Detection by using Naïve Bayes Classifier

Himani Kakkar, Lakhwinder Kaur  
Department of Computer Engineering, Punjabi University, Patiala  
[er.himanikakkar@gmail.com](mailto:er.himanikakkar@gmail.com), [mahal2k8@yahoo.com](mailto:mahal2k8@yahoo.com)

**Abstract--** Diseases in fruit cause a catastrophic problem and leads to economic and agricultural industry loss. Earlier infected fruit had detected manually but now with the advancement in technology image processing techniques have been developed. This framework works in two phases: training and testing. In training phase, all the data related to the non-infected and infected fruit is stored and in testing phase, it is analyzed that whether the fruit is infected or not and if yes then by which disease. In this paper, the technique was developed by combining K-mean clustering algorithm, speedup robust feature (SURF) feature detector and Naïve Bayes Classifier and implemented to detect the infected and non-infected fruit. The experiments are performed on fruit database and results are compared with neural network. The results show the superiority of the method with Naïve Bayes Classifier.

Keywords—k-mean, naïve bayes, SURF (speedup robust feature), NN(Neural Network) , blob detector.

## I. INTRODUCTION

Agriculture images are important source of information and data for horticulture industry. The use of image processing methods is of great significance for the analysis of agricultural problems. Fruit disease detection is one of the major applications that can be used by farmers for the detection of plant illness at growing stage. Detection of infection and plant health at growing stage of plants can help in controlling the fruit diseases by providing proper management approaches to plants at early stage i.e. disease-specific chemical applications and pesticide applications; and that leads to improvement in productivity of the plant. Automatic detection of fruit diseases is of great significance as it automatically detects the symptoms of disease from the images captured, as early as disease appear on the growing fruits. The diseases in fruits not only reduce the yield but also deteriorate the variety [9] and it may result in withdrawal from the cultivation. Diseases on fruit appear as spots in beginning and if not treated on time, cause the severe economic loss. Excessive uses of pesticide for the treatment of fruit disease may increase the danger of toxic residue level on soil and pesticides are identified as a major contributor to the groundwater contamination. Therefore, a framework is presented which can detect the diseases in the fruits as soon as they produce their symptoms on the fruits such that proper treatment can be applied on the plant [10]. With smart farming today's farmer can use decision tools and automation techniques which seamlessly integrate product, knowledge and services for better productivity, grading and surplus yield.

In this work the problem of the automatic detection and classification of fruit disease is considered and a framework for this is presented. The system works on two diseases of apple i.e. apple rot and apple scab and two diseases of mango i.e. mango anthracnose and mango fruit fly. In this framework K-mean algorithm, for the image segmentation is used, it converts the image into clusters and segments the infected part of the fruit. Then the feature extraction technique is used for extracting the feature of infected part. For feature extraction SURF is used that gives the better results as it uses the blob detection technique. Then for image classification Naïve

Bayes classifier is used which have given better results as compared to neural network in case of time consumption and confusion matrix and the Naïve Bayes classifier is more precise and relatively faster in terms of implementation. When the fruit diseases are detected, then appropriate treatment are accordingly recommended which will help the farmers in controlling the disease as well as in increasing the productivity of fruit.

## II. RELATED WORK

KutibaNanaa et al. [1] presented a method for detecting mango fruits from images. The main contribution was to employ the elliptical mango shape in detection method. Method was based on pre-processing operators on image includes converting to gray image, finding edges, calculating distances to edges, opening morphology and converting to binary color image. To take advantage of the elliptical shape of mango fruit, they employ Randomized Hough Transform to find oval shapes in input image. Back propagation Neural Network was used to classify the mango fruit from their proposed oval shapes. Three layers of neural network were used. Input layer used 450 neurons to forward values of the cropped oval shape image, one hidden layer included 50 neurons, and output layer. Experimental results show that mango detection rate is up to 96.26% in the case of clear appearance of mango and 90% in the case of ripped mango. Han li et al. [2] proposed a novel method termed ‘extended spectral angle mapping (ESAM)’ to find citrus greening disease (Huang long bingor HLB). A research was carried out using a HS image attained by an airborne HS imaging system, and a multispectral image. Gabriel et al. [3] proposed a pattern recognition method to automatically differentiate stem and calyx ends and detect damaged berries. First, blue berries were imaged under standard conditions to extract color and geometrical features. Second, five algorithms were tested to select the best features to be used in the subsequent evaluation of classification algorithms and cross-validation. The best classifiers were Support Vector Machine and Linear Discriminant Analysis [3]. Shiv Ram Dubey et al. [4] presented an adaptive approach for the identification of fruit diseases. This approach depends on the image processing which consist of some stages such as segmentation, feature extraction and image classification. The result analysis shows that the presented approach may considerably maintain accurate detection and automatic identification of fruit diseases. Juan Gómez-Sanchis et al. [5] presented a system for detecting two types of fungi which belongs to the Penicillium genus in citrus fruits. The goal is to avoid or at least reduce associated economic losses. They had used the MRMR (Minimum redundancy and maximum relevance) method that has reduced the number of features considerably. Rajesh. Yakkundimath et al. [6] proposed a reduced feature set based approach for recognition and classification on images of fruits into normal and affected. The average accuracy of 93.15% for normal type and 89.50% for affected type is obtained using 2 texture features .They finds application in developing a machine vision system in agriculture and horticulture fields. Anand Singh Jalal et al. [7] proposed an image processing-based apple fruit disease classification approach is introduced and validated. The approach comprised of the four steps. K-means clustering-based defect segmentation method was used in the first step for a region of interest extraction. In the second step, state-of-the-art color-, texture- and shape-based features were drawn from the segmented apple diseases. The different types of features were combined to form the more distinctive feature in the third step. In the last step, the training and classification was done using a MSVM (Multi-class support vector machine\_. Three kinds of apple diseases, including apple blotch, apple rot, and apple scab as well as normal apples were considered as the case study for the experimentation. The experiments and

results pointed out the significance and distinctiveness of their method for apple disease classification problem. Based on the classification results, they had concluded that the normal apples were easily distinguishable as compared to the infected apples and the combinations of the color, texture and shape based features outperform the state-of-the-art colour, texture and shape features standalone with less contribution from shape feature. Shiv Ram Dubey et al. [8] introduced and evaluated an approach to recognize the fruit and vegetable from the images. The described framework operates in three steps, background subtraction, feature extraction and training and classification. Background subtraction was performed using K-means clustering-based segmentation technique. They extracted some state-of-art color and texture features from the foreground image and fused them together. The fusion of color and texture information makes the resultant feature more discriminative than color and texture feature individually. They used a MSVM (Multi-class Support Vector Machine) for the training and classification. They had compared the performance fused features for SVM (Support Vector Machine) and nearest neighbour classifier and indicate that support vector machine is better choice for training and classification.

### III. METHODOLOGY

The main purpose is to supervise the diseases on fruit and suggest alternate solution for healthy yield and good productivity. Image acquisition is consistently the initial condition for the work flow series of image processing because as processing is possible only with the help of an image. For image segmentation, K-Means clustering technique is used. Feature vectors such as image color, morphology, texture and structure of hole are applied for extracting features of each image and for diagnosis of disease morphology gives accurate result. SURF algorithm is used as locator and descriptor for extracting the features and for classification of disease naïve bayes classifier is used. The algorithm consists of following steps (see figure 1):

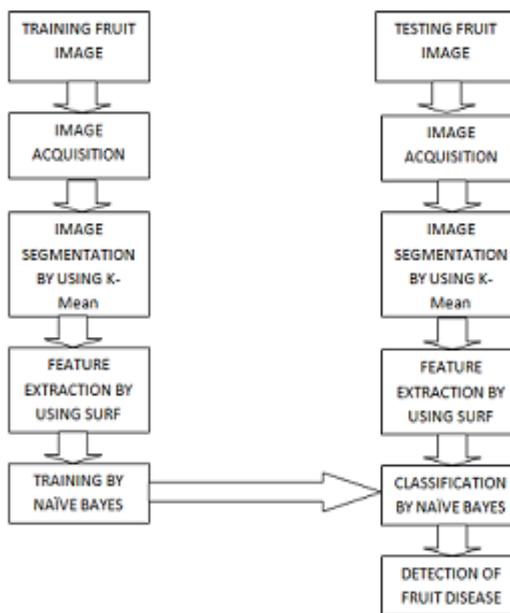


Figure 1: Steps Involved In Detection of Fruit Disease

- *Image acquisition:* Image acquisition in image processing may be termed as the action of retrieving an image from some source.
- *Segmentation:* Segmentation is a technique which partitions an image into distinct regions. Regions and each region will contain pixel with similar attributes. Segmentation is the first step from low-level image processing for transforming a grayscale or color image into one or more other images to high-level image description in terms of features, objects, and scenes [15]. K-mean segmentation technique is used in the present work.
- *Feature Extraction:* Four feature vectors are considered namely color, texture, morphology and structure of hole of the fruits. SURF (Speed up Robust Feature) algorithm is applied for extracting the features. It is used as local descriptor and blob detector.
- *Classifier:* For classification whether fruit is defected or not Naïve Bayes (NB) Classifier is used.

### *III a. Speeded Up Robust Features (SURF)*

SURF is one of the feature detector and descriptor which is utilized further for the task like classification, recognition etc. It has been proposed by Herbert Bay in the year 2006 at European Conference on computer vision in USA [11]. SURF is motivated by the scale invariant feature transform (SIFT) descriptor. In comparison with SIFT, SURF descriptor is faster and robust.

SURF is used as integer approximation of determinant of hessian blob detector that may be calculated with three operations. Moreover SURF descriptor is being used to discover objects, recognize faces and re-built 3-Dimensional scenes and to track object.

In SURF [12], square-shaped filters are used as an approximation of Gaussian smoothing. The image is being filtered with a square is much faster only if integral images can be used. It may be explained as (1):

$$S(x, y) = \sum_{i=0}^x \sum_{j=0}^y I(i, j) \quad (1)$$

By using the integral images, the addition of real images with in the rectangle may be evaluated. It needed four evaluations at the corners of the rectangle.

SURF can also be utilized as blob detector relied on the hessian matrix. The determinant of hessian matrix is useful for measuring the local change around the points. These points are selected if the determinant is maximum. SURF uses the determinant of the Hessian for choosing the scale.

Given a point  $p=(x, y)$  in an image  $I$ , the Hessian matrix  $H(p, \sigma)$  at point  $p$  and scale  $\sigma$ , is defined as follows [13]:

$$H(p, \sigma) = \begin{pmatrix} L_{xx}(p, \sigma) & L_{xy}(p, \sigma) \\ L_{xy}(p, \sigma) & L_{yy}(p, \sigma) \end{pmatrix} \quad (2)$$

Where  $L_{xx}(p, \sigma)$  etc. are the second-order derivatives of the grayscale image.

### *IIIb. Blob Analysis*

Blob detection methods [14] are aimed at detecting regions in digital images that differ in properties, such as brightness or color, compared to surrounding regions. Region of an image in which some properties are approximately constant can be called as Blob. The Blob Analysis consists of following stages:

- Extraction: It uses image thresholding technique to detect a region corresponding to single object or objects.
- Refinement: For region refinement, it uses transformation techniques.
- Analysis: It is final step after refinement, for analysis. If the region shows multiple objects then divide it into separate blobs for inspection.

*IIIc. Pattern Matching:* Pattern matching is the act of checking a given sequence of tokens for the presence of the constituents of some pattern [16]. In present work, Naïve Bayes concept is applied for pattern matching, which classifies disease.

*IIId. Naïve Bayes (NB) Classifier [16]*

Naïve bays classifier is a classification approach depends on bayes theorem. Naive Bayes model is very useful for large data sets. Generally, it outperforms highly sophisticated classification methods. These models are popular in machine learning applications. Naive Bayes classifier uses the concept that the presence of a particular feature in a class is unrelated to the presence of any other feature. This property allows every feature towards final result to contribute equally and independently from the other features. These models are suitable for numerous domains. When making decision, NB classifier are computationally fast. These classifiers are used for multi class prediction.

#### IV. RESULTS AND DISCUSSIONS

The work presented in this paper is implemented in Mat lab for fruit disease detection. Apple and mango fruit are chosen as a case study considering apple rot, apple scab, mango anthracnose and mango fruit fly diseases. Images of non-infected fruit samples and infected fruit samples of apple and mango are collected and stored in jpeg format.

These images are firstly resized to 480X640 resolutions, then the resized images are converted into lab colour format from srgb (standard rgb format), then the image is divided into clusters by using k-mean segmentation technique. After segmentation feature extraction is done by using SURF (Speedup Robust Feature Extractor) and four features are considered i.e. colour, texture, morphology and structure of hole. We had extracted features of all the images of dataset and saved them in the form of a matrix by using a command save ('new name.mat', 'feature variable'). Then depending upon these features, the naïve Bayes classifier will match the query image feature with the features matrix and provide an output based on it. Results are shown in Figures 2 to 12.

The Naïve Bayes (NB) classifier detects whether the image is defected or not and if it is then by which disease apple rot or apple scab. In case of mango classifier will detect rather the query image is detected by mango anthracnose or by mango fruit fly. This system also provides the solution of disease i.e. which pesticide is to be applied and how to increase productivity of the yield. The analysis shows that the naïve bayes classifier is 57.63% accurate in case of apple and NN is 42.2% accurate and for mango naïve bayes is 75.7% and NN is 46.5.

Confusion matrix of NB classifier tells how many images of a particular class, out of total number of images are detected accurately. As first column and first row of the confusion matrix tells how many images are detected accurately from total no of apple rot images. i.e. from 23 images 15 images figure 11. NB can detect accurately

and so on. The graph (figure 11 and figure 12) represents performance of NB based approach and the feed forward neural network i.e. the base technique. The blue solid line in graph represents how NB based technique detects the various diseases of apple and mango accurately and the red dotted line represent how NN based technique detects diseases.

The NB classifier is more precise and relatively faster in terms of implementation as compared to the Feed Forward Neural Network. The time consumed by Naïve Bayes Classifier to detect a disease is comparatively less than the Neural Network i.e. 0.0117 and 0.4199 respectively. Confusion Matrix of NN depicts how accurately NN can classify the disease. The Green cells represent the accurate classification and the red cell represents the in-accurate classification.

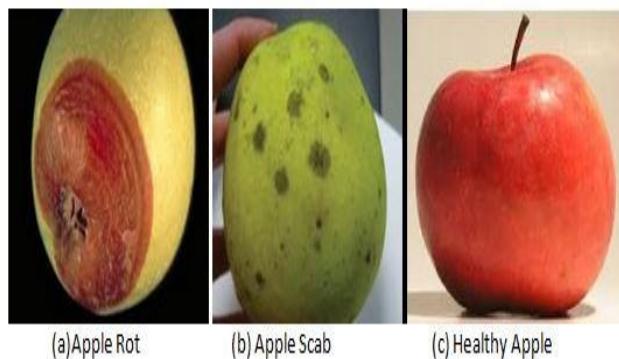


Figure 2: Images of Apple Fruit Samples

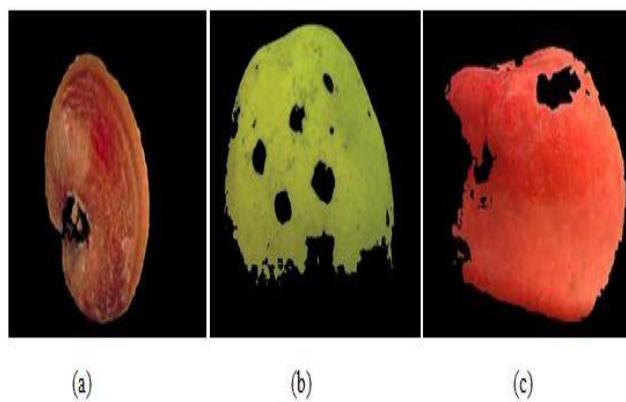


Figure 3: Segmented portions of the Apple sample images

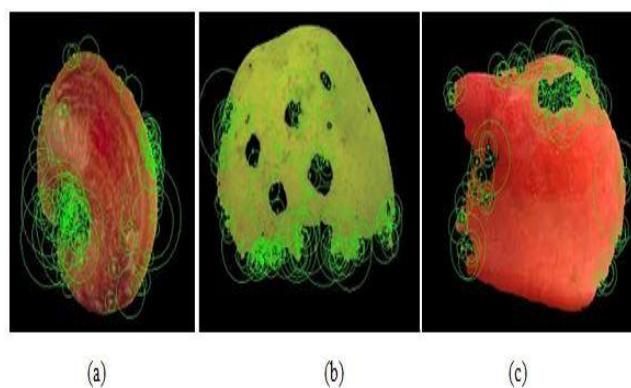
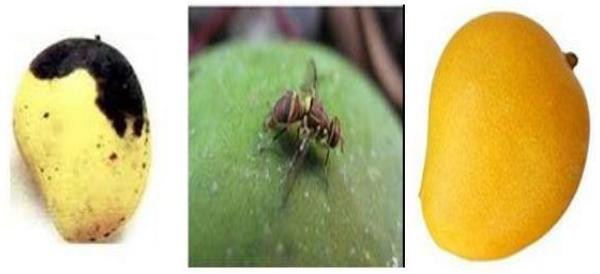


Figure 4: Extracted Features of Segmented Image of Apple Using SURF



(a) Mango Anthracnose      (b) Mango Fruit Fly      (c) Healthy Mango

Figure 5: Images of Mango Fruit Samples

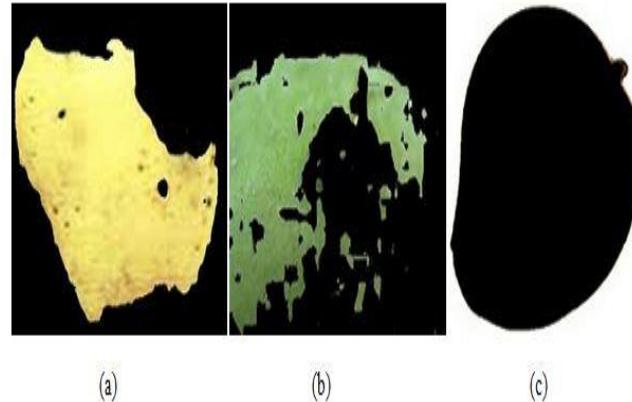


Figure 6: Segmented Portions of Mango Sample Images

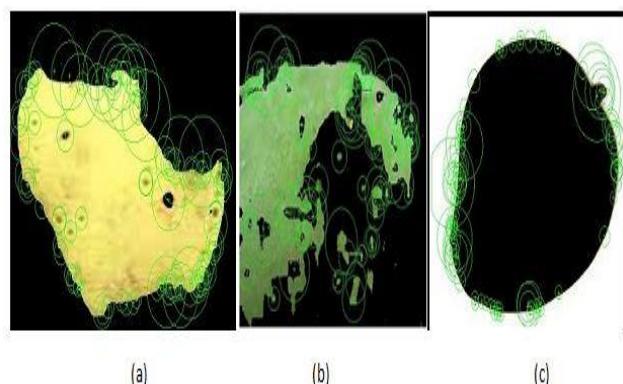


Figure 7: Extracted Features of Segmented Image of Mango using SURF

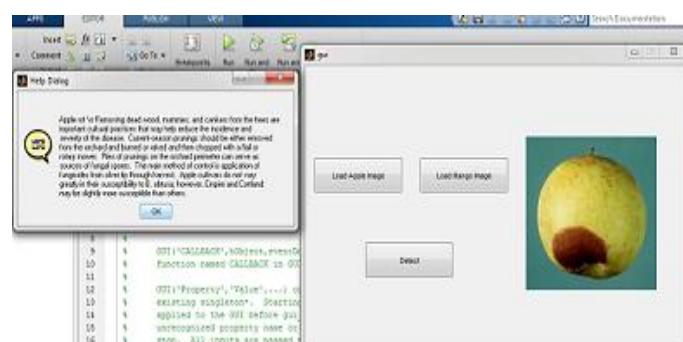


Figure 8: GUI Screenshot for Apple Rot Disease Detection

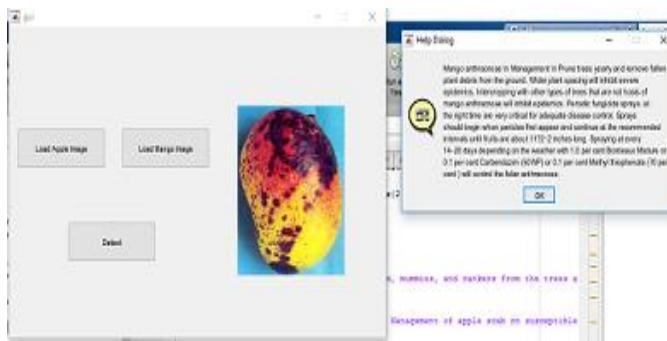


Figure 9: GUI Screenshot for Mango Anthracnose Disease Detection

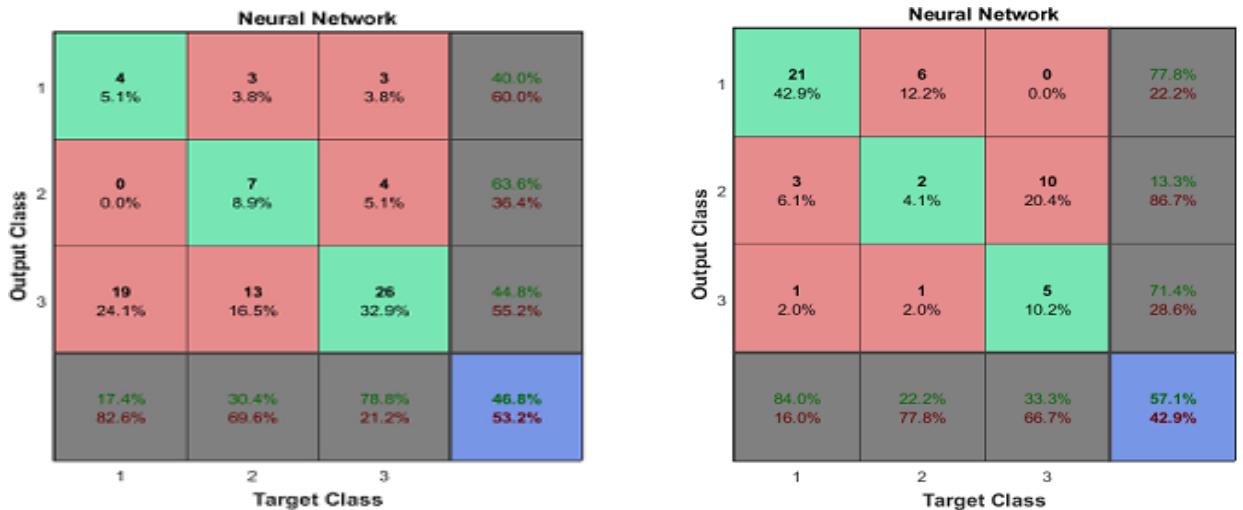


Figure 10: Confusion Matrix for Apple Disease Detection and Mango Disease Detection respectively by using Neural Network.

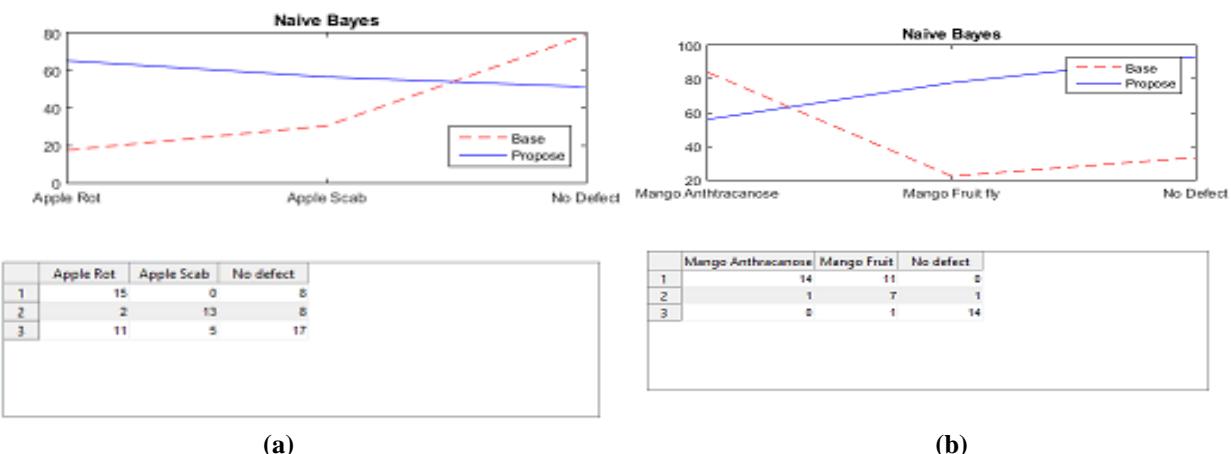


Figure 11. Performance of Naïve Bayes for (a) Apple disease detection, and (b) Mango disease detection.

#### ACKNOWLEDGMENT

I want to thank the Almighty God then I wish to express my gratitude to Dr. Lakhwinder Kaur , H.O.D Computer Engineering for providing me an opportunity to do my research work in “Fruit Disease Detection by

using Naive Bayes Classifier" and for her guidance . I also wish to express my gratitude to all the professors and assistant professors of University College of Engineering, Punjabi University Patiala for supporting me during the period of my research work. I wish to thank all my mates for supporting and encouraging me throughout the research work.

## REFERENCES

- [1] Nanaa, Kutiba, Mohamed Rizon, AbdRahman, MohdNordin, Yakubu Ibrahim, Abd Aziz, and AzimZaliha. "Detecting mango fruits by using randomized Hough transform and back propagation neural network." in Information Visualisation (IV), 2014 18th International Conference on, pp. 388-391. IEEE, 2014.
- [2] Li, Han, Won Suk Lee, Ku Wang, Reza Ehsani, and Chenghai Yang. "‘Extended spectral angle mapping (ESAM)’ for citrus greening disease detection using airborne hyper spectral imaging." Precision agriculture 15, no. 2 (2014): 162-183.
- [3] Leiva-Valenzuela, Gabriel A., and José Miguel Aguilera. "Automatic detection of orientation and diseases in blueberries using image analysis to improve their postharvest storage quality." Food Control 33, no. 1 (2013): 166-173.
- [4] Dubey, Shiv Ram, and Anand Singh Jalal. "Adapted approach for fruit disease identification using images." arXiv preprint arXiv:1405.4930 (2014).
- [5] Gómez-Sanchis, Juan, José D. Martín-Guerrero, Emilio Soria-Olivas, MarcelinoMartínez-Sober, Rafael Magdalena-Benedito, and José Blasco. "Detecting rottenness caused by *Penicillium* genus fungi in citrus fruits using machine learning techniques." Expert Systems with Applications 39, no. 1 (2012): 780-785.
- [6] Rajesh Yakkundimath, Pujari, and A. S. Byadgi. "Reduced colour and texture features based identification and classification of affected and normal fruits' images." International Journal of Agricultural and Food Science 3, no. 3 (2013): 119-127.
- [7] Anand Singh Jalal, dubey "Apple disease classification using colour, texture and shape features from images." Signal, Image and Video Processing (2015): 1-8.
- [8] Dubey,Shiv Ram, and Anand Singh Jalal. "Fruit and vegetable recognition by fusing colour and texture features of the image using machine learning." International Journal of Applied Pattern Recognition 2, no. 2 (2015): 160-181.
- [9] Shiv Ram Dubey, Anand Singh, Jalal "Automatic Fruit Disease Classification Using Images".
- [10] Manisha Bhangea , H.A.Hingoliwala "Smart Farming: Pomegranate Disease Detection Using Image Processing" Second International Symposium on Computer Vision and the Internet (VisionNet'15).  
[https://en.wikipedia.org/wiki/Speeded\\_up\\_robust\\_features](https://en.wikipedia.org/wiki/Speeded_up_robust_features).
- [11] [http://broom02.revolvy.com/main/index.php?s=Speeded%20Up%20Robust%20Features&item\\_type=topic](http://broom02.revolvy.com/main/index.php?s=Speeded%20Up%20Robust%20Features&item_type=topic)
- [12] Sreelekshmi K R, Dr. Shyama Das "Interactive Example-based Colour Transfer using Speeded Up Robust Feature" International Journal of Engineering Research and General Science Volume 4, Issue 3, May-June, 2016.
- [13] <https://www.cs.auckland.ac.nz/courses/compsci773s1c/lectures/ImageProcessing-html/topic3.htm>.
- [14] [https://en.wikipedia.org/wiki/Pattern\\_matching](https://en.wikipedia.org/wiki/Pattern_matching).
- [15] <https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained>.
- [16] <https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained>.

# COMPARATIVE ANALYSIS OF TANAGRA AND R DATA MINING TOOL FOR DIABETIC DIAGNOSIS USING K MEAN CLUSTERING AND GENETIC ALGORITHM BY INTEGRATING WITH S.V.M

Ramanpreet kaur<sup>1</sup>, Gurpreet Singh<sup>2</sup>

<sup>1</sup>Student M. Tech, Computer Engineering Department, Punjabi University, Patiala, India

<sup>2</sup>Assistant Professor M. Tech, Computer Engineering Department, Punjabi University, Patiala, India

**Abstract:** Diabetes mellitus is a chronic disease and a major public health challenge worldwide. According to the International Diabetes Federation, there are currently 246 million diabetic people worldwide, and this number is expected to rise to 380 million by 2025. Diabetes is a standout amongst the most well-known non-transmittable diseases in the world. Vast amount of data available in health care industry is difficult to handle, hence mining is necessary to find the necessary pattern and relationship among the features available. Medical data mining is one major research area where evolutionary algorithms and clustering algorithms play a vital role. Several data mining and machine learning methods have been used for the diagnosis, prognosis, and management of diabetes. Several researchers are using statistical and data mining tools like rapid miner ,weka , KNIME etc. to help health care professionals in the diagnosis of diabetes. The data source for this research is taken from UCI repository. Various experiments are made iteratively by using various techniques on Tanagra and R tool. In this research work, K-Means is used for removing the noisy data and genetic algorithms for finding the optimal set of features with Support Vector Machine (SVM) as classifier for classification. It shows that the proposed method using Tanagra with an accuracy of 76.44 % has attained better results compared to R Tool with an accuracy of 75.39 %.

## 1. INTRODUCTION:

The abundance of data, together with the need for powerful data analysis tools in many countries has been described as data rich but information poor society. Data mining (sometimes called knowledge discovery in database) is an extraction of information from huge set of data or we can say that data mining is mining of knowledge from data. Data mining techniques can be implemented on massive data in automated manner whereas traditional statistical methods require custom work by experts. Diabetes mellitus is a chronic disease and a major public health challenge worldwide. Diabetes is often called a modern-society disease because widespread lack of regular exercise and rising obesity rates are some of the main contributing factors for it. According to the International Diabetes Federation, there are currently 246 million diabetic people worldwide, and this number is expected to rise to 380 million by 2025. This research work proposes K-Means clustering based outlier detection followed by GA for feature selection

with SVM as classifier to classify the dataset. The rest of the paper is organized as follows: Section 2 shows the literature review. The preliminaries used is given in section 3 followed by data source description in section 4. The proposed model is described in section 5. Section 6 reports the experimental analysis and the final section deals with a conclusion.

## 2. LITERATURE SURVEY:

*Azra et al* proposed a method to create a prediction model using data mining approach to identify low risk individuals for incidence of type 2 diabetes, using the Tehran Lipid and Glucose Study (TLGS) database. *C. M. Velu et al* uses Expectation-Maximization (EM) algorithm for sampling. The study of classification of diabetic patients was main focus of this research work. Diabetic patients were classified by data mining techniques for medical data obtained from Pima Indian Diabetes (PID) data set. This research was based on three techniques of EM Algorithm, h means+ clustering and Genetic Algorithm (GA). These techniques were employed to form clusters with similar symptoms. Result analyses proved that h-means+ and double crossover genetics process based techniques were better on performance comparison scale. The simulation tests were performed on WEKA software tool for three models used to test classification.

*Emirhan et al* proposed a method is to develop a data mining model for which will predict a suitable dosage planning for diabetes patients. Medical records of 89 different patient records were used in this study. 318 diabetes assays were extracted using these patient records. ANFIS and Rough Set methods were used for dosage planning objective.

*Bjørnar et al* proposed a second edition of the Diabetes Atlas extends coverage to 212 countries and territories around the world. It provides current estimates of the prevalence of diabetes and impaired glucose tolerance (IGT) as well as forecasts the estimates for 2025.

*Jianchao et al* proposed a model for the data pre-processing, including attribute identification and selection, outlier removal, data normalization and numerical discretization, visual data analysis, hidden relationships discovery, and a diabetes prediction model

*K Selvakuberan et al* propose a Feature Selection approach using a combination of Ranker Search method. The classification accuracy of 81% resulted from our approach proves to be higher when compared with previous results.

*Mai Shouman et al* proposes a model to systematically close those gaps to discover if applying data mining techniques to heart disease treatment data can provide as reliable performance as that achieved in diagnosing heart disease.

*Marios Skevofilakas et al* proposes a design and develop a Decision Support System (DSS) closely coupled with an Electronic Medical Record (EMR), able to predict the risk of a Type 1 Diabetes Mellitus (T1DM) patient to develop retinopathy

*Marios et al* proposes a model that is focused on the integration of state of- the-art technologies in the fields of telecommunications, simulation algorithms, and data mining in order to develop a Type 1 diabetes patient's semi to fully – automated monitoring and management system.

*Nahla H. Barakat et al* proposes utilizing support vector machines (SVMs) for the diagnosis of diabetes. In particular, we use an additional explanation module, which turns the “black box” model of an SVM into an intelligible representation of the SVM’s diagnostic (classification) decision.

*P. Kasemthaweesab et al* has presented a basic method of discovering an association of diabetes mellitus with complication states by applying gender, age and occupation factors and testing to find out a relationship of diagnostic data.

*Sarojini Balakrishnan et al* propose a feature selection approach for finding an optimum feature subset that enhances the classification accuracy of Naive Bayes classifier.

*Abdullah A. Aljumah et al* concentrates upon predictive analysis of diabetic treatment using a regression-based data mining technique. The Oracle Data Miner (ODM) was employed as a software mining tool for predicting modes of treating diabetes.

*Akira Hara et al* report the overview of the CHD DB and the soft computing methods, and discuss the features of respective methods by comparison of the experimental results.

*Arianna Dagliati et al* research has been performed within the EU project MOSAIC, which gathers T2D patients' data coming from three European hospitals and a local health care agency. The main idea underlying our approach is to use a suite of temporal data mining methods in order to derive healthcare pathways.

### 3. PRELIMINARIES

#### 3.1. K-MEANS CLUSTERING ALGORITHM:

It was developed in 1976 by MacQueen. It is a unsupervised clustering algorithm generates a specific number of disjoint, flat (non-hierarchical) clusters. The procedure follows simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. K-Means algorithms randomly choose k objects, representing the k initial cluster centre. The next step is to take each point belonging to a given data set and associate it to the nearest centre based on the closeness of the object with cluster centre using Euclidean distance. After all the objects are distributed, recalculate new k cluster centres. The process is repeated until there is no change in k cluster centres.

The steps of K-Means Algorithm are given below:

- Randomly partition the dataset into k.
  - For each data point in the dataset: $x$  Calculate the distance from the data point to each cluster. If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.
  - Repeat the above step until a complete pass through all the data points’ results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
4. The choice of initial partition can greatly affect the final clusters that result, in terms of intercluster and intra-cluster distances and cohesion.

### 3.2. GENETIC ALGORITHM:

It is based on Darwin's theory of natural selection and 'survival of the fittest'. It is one important search method used for feature selection. It represents feasible solutions to the particular problem that evolves over successive iterations (generations) through a process of competition and controlled variation. The process of selection is to select the chromosomes in the population which satisfy the associated fitness to form new ones in the competition process. The genetic operators such as crossover and mutation are used to create the new chromosomes. The classical model of GAs consists of three operations:

- Evaluation of individual fitness.
- Formation of a gene pool (intermediate population) through selection mechanism.
- Recombination through crossover and mutation operators.

The selection mechanism produces a new population  $P(t)$  with copies of chromosomes in  $P(t-1)$ . The number of copies received for each chromosome depends on its fitness; chromosomes with higher fitness usually have a greater chance of contributing copies to  $P(t)$ . Then, the crossover and mutation operators are applied to  $P(t)$ . Crossover takes two individuals called parents and produces two new individuals called the offspring by swapping parts of the parents. In its simplest form, the operator works by exchanging substrings after a randomly selected crossover point. The crossover operator is not usually applied to all pairs of chromosomes in the new population. A random choice is made, where the likelihood of crossover being applied depends on probability defined by a crossover rate. Mutation serves to prevent premature loss of population diversity by randomly sampling new points in the search space. Mutation is applied by flipping one or more random bits in the bit string with a probability equal to the mutation rate. Termination may be triggered by reaching a maximum number of generations or by finding an acceptable solution by some criterion.

### 3.3. SUPPORT VECTOR MACHINE:

It is a classifier that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. The SVM method provides an optimally separating hyper plane in the sense that the margin between two groups is maximized. "Support Vectors" are defined as subset of data instances used to define the hyper plane. The distance between the hyper plane and the nearest support vector is called as margin. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. There are two types of SVMs,

- Linear SVM is used to separate the data points using a linear decision boundary and
- Non-linear SVM separates the data points using a nonlinear decision boundary. Traditional SVM training algorithms require quadratic programming (QP) package. Solving a quadratic programming problem is slow and requires a lot of memory as well as in-depth knowledge of numerical analysis. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering and is often used in industrial applications.

#### 4. DATA SOURCE

Pima Indian Diabetes contains female patients with at least 21 years old. It is used to diagnose the presence of diabetes in pregnant women. There are 768 records, out of which 268 cases in class ‘positive test for diabetes’ and 500 cases for “negative test for diabetes” with 376 records contain missing values. It contains 8 numerical attributes as input and one output variable. The attribute information present in the dataset is as follows:

1. *Number of times pregnant*
2. *Plasma glucose concentration a 2 hours in an oral glucose tolerance test*
3. *Diastolic blood pressure (mm Hg)*
4. *Triceps skin fold thickness (mm)*
5. *2-Hour serum insulin (mu U/ml)*
6. *Body mass index (weight in kg/(height in m)^2)*
7. *Diabetes pedigree function*
8. *Age (years)*
9. *Class variable (0 – Absence of disease / 1 – Presence of disease)*

TABLE 1: ATTRIBUTES OF DIABETES DATASET

Attribute	Description
Age	Person age
Family Background	Whether anybody from his family suffering from disease
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Glucose tolerance test is a medical test in which glucose is given and blood samples taken afterward to determine how quickly it is cleared from the blood
Diastolic blood pressure	Blood Pressure
Triceps skin fold thickness	A value used to estimate body fat, measured on the right arm halfway between the olecranon process of the elbow and the acromial process of the scapula. Normal thickness in males is 12 mm; in females, 23 mm.
2-Hour serum insulin	Injection of insulin in body
No. of times pregnant	Count of no. of times pregnant
BMI	Body Mass of a person
Class Variable	class value 1 is interpreted as tested positive for diabetes and class value 0 is interpreted as “tested negative for diabetes”

**5. PROPOSED METHOD:**

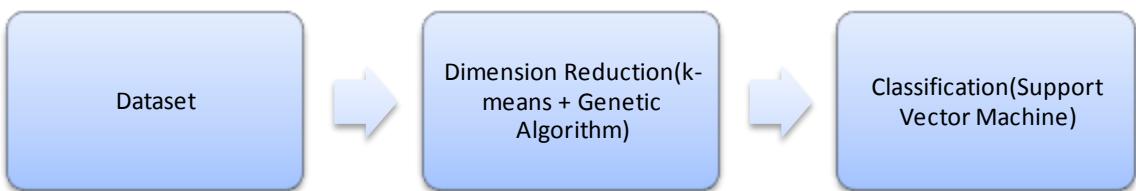


FIGURE1: BLOCK DIAGRAM OF PROPOSED METHOD

The working principle of proposed system shown in Fig.1 comprises of 3 steps:

1. Data cleaning is done by replacing the missing values with mean (as the dataset lies under a normal distribution curve).
2. The cleaned datasets is clustered using K-Means to remove outliers, inconsistent and noisy data and the reduced data is used for selecting the optimal features with genetic algorithm.
3. The reduced dataset is classified using SVM classifier to achieve better accuracy compared to existing methods in literature. In order to increase the reliability of classifier performance 10-fold cross-validation method.

The entries in the confusion matrix have the following meaning in the context of our study:

*A is the number of correct predictions that a instance is negative*

*B is the number of incorrect predictions that an instance is positive*

*C is the number of incorrect of predictions that an instance negative and*

*D is the number of correct predictions that an instance is positive.*

TABLE 2.CONFUSION MATRIX REPRESENTATIONS

		PREDICTIVE	
		NEGATIVE	POSITIVE
ACTUAL	NEGATIVE	A	B
	POSITIVE	C	D

The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC \text{ (accuracy)} = A+D/A+B+C+D$$

To visualize the performance of supervised machine learning algorithms, confusion matrix is used. The four classification performance indices present in confusion matrix . The evaluation metrics used are:

Sensitivity, Specificity, Positively Predicted and Negatively Predicted value. Sensitivity and Specificity is the proportion of actual positives and negatives which are correctly identified as such. The Positive and Negative predictive values are the proportions of positive and negative results, which are predicted.

*Sensitivity:  $A/(A+C) \times 100$*

*Specificity:  $D/(D+B) \times 100$*

*Positive Predictive Value:  $A/(A+B) \times 100$*

*Negative Predictive Value:  $D/(D+C) \times 100$*

## **6. EXPERIMENTAL RESULTS:**

After replacing the missing values by mean, outliers, noisy and inconsistent samples / cases were removed using simple K-Means clustering algorithm. Then, GA was used as a feature selection tool and its output was fed to SVM using 10-fold cross validation technique for classification. . In order to accomplish this research and to measure and investigate the performance on the selected methods namely k-means algorithm, genetic algorithm and SVM we use the experiment procedures by Tanagra and R Tool.

In very first experimental part we are using Tanagra known as powerful data mining tool, a text file is chosen that has been taken from U.C.I repository. This text file contains 768 observations which have information about the attributes taken for research proceeding. The data is taken from U.C.I repository namely Pima Indian diabetes dataset.

TABLE 4.1 PERFORMANCE MEASURING USING DIABETES DATASET USING TANAGRA

CLASSIFIER	ACCURACY	ERROR RATE	TIME(MILI SECONDS)
K-MEANS ALGORITHM + GENETIC ALGORITHM + S.V.M	76.44%	23.55%	11813MS

Then using R Tool data mining tool following are the results shown:

TABLE 4.2 PERFORMANCE MEASURING IN TRAINING AND TEST DATA SETUSING R TOOL

CLASSIFIER	ACCURACY	ERROR RATE
K-MEANS ALGORITHM + GENETIC ALGORITHM + S.V.M	75.39 %	24.61%

## **CONCLUSION:**

The main goal of this research is to analyze the risk factor of diabetes and providing efficient health care

services thought there are different data mining techniques that can be used for the identification of diabetes. The data source for this research is collected from UCI repository. In this research three techniques are applied to predict the diabetes disease that is k-means algorithm, genetic algorithm and the S.V.M (Support Vector Machine). Firstly by using Tanagra tool we have found the accuracy of 76.44% and then by using R tool we found the accuracy of 75.39%. If we compare with the average accuracy Tanagra tool is best as compared to R tool for diabetes disease prediction.

However there is always a chance of improvement in the research we use only three techniques but if we increase a number of techniques we could get more accurate results for the prediction diabetes disease. This research will serve as a training tool to doctors, nurses and medical students to diagnose patients with diabetes disease. It can also provide decision support to assist doctors to make better clinical decisions. It is based on the attributes used in diabetes data set. This research will help to predict the diabetes disease patient in future.

#### REFERENCES:

- [1] Azra Ramezankhani, Omid Poumik, Jamal Shahrabi, Davood Khalili, Fereidoun Azizi, Farzad Hadaegh,"*Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study*" diabetes research and clinical practice 105 ( 2014 ),pp.391– 398, 2014.
- [2] C. M. Velu, K. R. Kashwan,"*Visual Data Mining Techniques for Classification of Diabetic Patients*" 3rd IEEE International Advance Computing Conference (IACC),pp.1070-1075, 2012.
- [3] Emirhan Gülcin YÖldÜÖma , Adem Karahocaa , Tamer Uçara,"*Dosage planning for diabetespatients using data mining methods*",2011.
- [4] International Diabetes Federation, *Diabetes Atlas*, 3rd ed. Brussels, Belgium: International Diabetes Federation, 2007.
- [5] Jianchao Han, Juan C. Rodriguez, Mohsen Beheshti,"*Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner*"2008 Second International Conference on Future Generation Communication and Networking,pp.96-99,2008.
- [6] K Selvakuberan,D Kayathiri, B Harini Dr M Indra Devi,"*An efficient feature selection method for classification in Health care Systems using Machine Leaming Techniques*",pp.223-226,2011.
- [7] Mai Shouman, Tim Turner, Rob Stocker,"*Using data mining techniques in heart disease diagnosis and treatment*" 2012 Japan-Egypt Conference on Electronics, Communications and Computers,pp.173-177,2012.
- [8] Marios Skevofilakas,Konstantia Zarkogianni, Basil G. Karamanos, Konstantina S.Nikita,Senior Member, IEEE,"*A hybrid Decision Support System for the Risk Assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus*", 32nd Annual International Conference of the IEEE EMBS Buenos Aires, Argentina, pp. 6713-6716, 2010.
- [9] Marios Skevofilakas, Stavroula G. Mougiakakou, Konstantia Zarkogianni, Erika Aslanoglou,Sotiris A. Pavlopoulos, Andriani Vazeou, and Christos S. Bartocas, Konstantina S. Nikita,"*A Communication and Information Technology Infrastructure for Real Time Monitoring and Management of the Diabetes Patients*", Proceedings of the 29th Annual International Conference of the IEEE EMBS Cité Internationale,pp.3685-3688, 2007.
- [10] Nahla H. Barakat, Andrew P. Bradley, Senior Member, IEEE and Mohamed Nabil H. Barakat,"*Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus*" IEEE Transactions on Information technology in biomedicine, Vol.14, No.4, July 2010.
- [11] P. Kasemthaweesab, W. Kurutach,"*Association Analysis of diabetes mellitus (DM) with complication states based on association rules*", 7th IEEE Conference on Industrial Electronics and Applications (ICIEA) ,pp. 1453 - 1457,2011.
- [12] Sarojini Balakrishnan, Ramaraj Narayanaswamy, Nickolas Savarimuthu, Rita Samikannu,"*SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases*" 2008 IEEE International Conference on Systems, Man and Cybernetics (SMC 2008) ,pp.2628-2633,2008.
- [13] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui,"*Application of data mining: Diabetes health care in young and old patients*", Journal of King saud university- Computer and information sciences, Vol.25,pp.127-136,2012.

- [14] Akira Hara and Takumi Ichimura, "Data Mining by Soft Computing Methods for The Coronary Heart Disease Database" December, Fourth International Workshop on computational Intelligence and Applications, IEEE SMC Hiroshima chapter, Hiroshima University, Japan, pp. 132-138, 2008.
- [15] Arianna Dagliati, Lucia Sacchi, Carlo Cerra, Paola Loporati, Pasquale De Cata, Luca Chiovato, John.H. Holmes and Riccardo Bellazzi, "Temporal Data Mining and Process Mining Techniques to Identify Cardiovascular Risk- Associated Clinical Pathways in Type 2 Diabetes Patients" Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference, pp- 240 – 243, 2014.

# An Approach for Weather Nowcasting from Doppler Weather Radar images using Thresholding

Author: Rhythm Naswa (M.Tech Student), Er. Navdeep Kanwal (Asst. Professor)

## Abstract

*Among the recent technology trends in the domain of Weather Nowcasting in India, Doppler weather Radars installed at various cities under IMD has proved to be a very significant tool. In IMD, Radars are used for the detection of various severe weather systems like, thunderstorms, hailstorm and tracking of cyclone storms. Images acquired from radar are interpreted for prediction of severe weather useful in the fields of aviation, agriculture, disaster management etc. to prevent the loss of life and property. The radar images have been studied for cloud detection based on the form and shape of clouds or the reflectivity of clouds, but the basic necessity lies in identifying the areas currently affected or tend to be affected by the clouds in near future. An approach has been discussed in this paper where the clouds are automatically detected from images acquired from DWR Patiala and the information about affected districts along with other parameters is transmitted to concerned offices. All this is proposed to be done without any human intervention and in time many folds lesser than the conventional method used by meteorological offices.*

## 1. Introduction

### 1.1 Doppler Weather Radar and Application of its Image Products

Frequency shift principle proposed by Christian J Doppler in 1853 applies to electromagnetic radiation from radar thus named as Doppler Weather Radar. In this case, the radar is stationary but target is moving. If the target is moving towards the radar the frequency is increased, if it is moving away the frequency is reduced. In IMD, Radars are used for the detection of various severe weather systems<sup>[7]</sup> like, thunderstorms, hailstorm and tracking of cyclone storms. They are also useful in estimation of rain and hailstorm warning. Various meteorological, aviation and hydrological products that are generated from Doppler weather radar data are very useful to the forecasters. Radars also help in estimating the storm's intensity, location and in forecasting its future track for the safe navigation of aircrafts and ships.

Severe weather is any type of violent or extreme weather event that represents a hazard to public safety and welfare, including tornadoes, flash floods, lightning, damaging windstorms, and hailstorms. The societal costs of severe weather events are very high, involving injuries, lost lives, and property damage. Accurate severe weather forecasts can help reduce the loss of life and also economic loss.

<sup>[3]</sup>Data processing would take up where signal processing leaves off. The data processing algorithms stored in the computer of the DWR takes the base parameter estimates (Z, V and W) and further processes them thereby generating a lot of ready to use DWR products/images for the radar users and forecaster in near real time. Various Radar Products<sup>[7]</sup> and their applications are as given below:

- Reflectivity Product (MAXZ) can be used to detect convective clouds based on the moisture content and height of cloud, thus indicating whether a weather phenomenon would occur or not.
- Rainfall Products (SRI, RAIN1, RAINN, PAC) help in detecting the amount of rainfall in various areas. PAC product is updated once in 24hrs proving summarized 24 hr rain that occurred in respective areas.

- Aviation Products (SHEAR, WIND) are mainly used by pilots to be aware of high turbulence areas in order to avoid airplane crashes or any other damage.

### 1.2 Image construction in Doppler Weather Radar and Color configurations

The processing of radar data generally involves two steps<sup>[3]</sup> : The first step, being signal processing, is the extraction of raw radar parameters like echo strength i.e. reflectivity or Doppler velocity from the signals coming out of the radar receiver. The second step, data processing or product generation (done by RPG-Radar Product Generator), involves the further processing of raw radar parameters in order to obtain information which is useful for meteorological purposes. In general, these two steps are done by different computers, signal processing done at the radar site, while product generation can be done everywhere the data are sent to.

The RPG is a high power computer that runs a variety of algorithms and software that generate more than 70 meteorological products available to users. It also facilitates on-line mass storage and archiving of data. It serves as the command center for the entire system. The RPG processes digital raw data and generates the Base and Derived Products, by clutter filtering and providing other functions.<sup>[13]</sup> It performs some data quality checks on the raw data and then creates radar images and products.

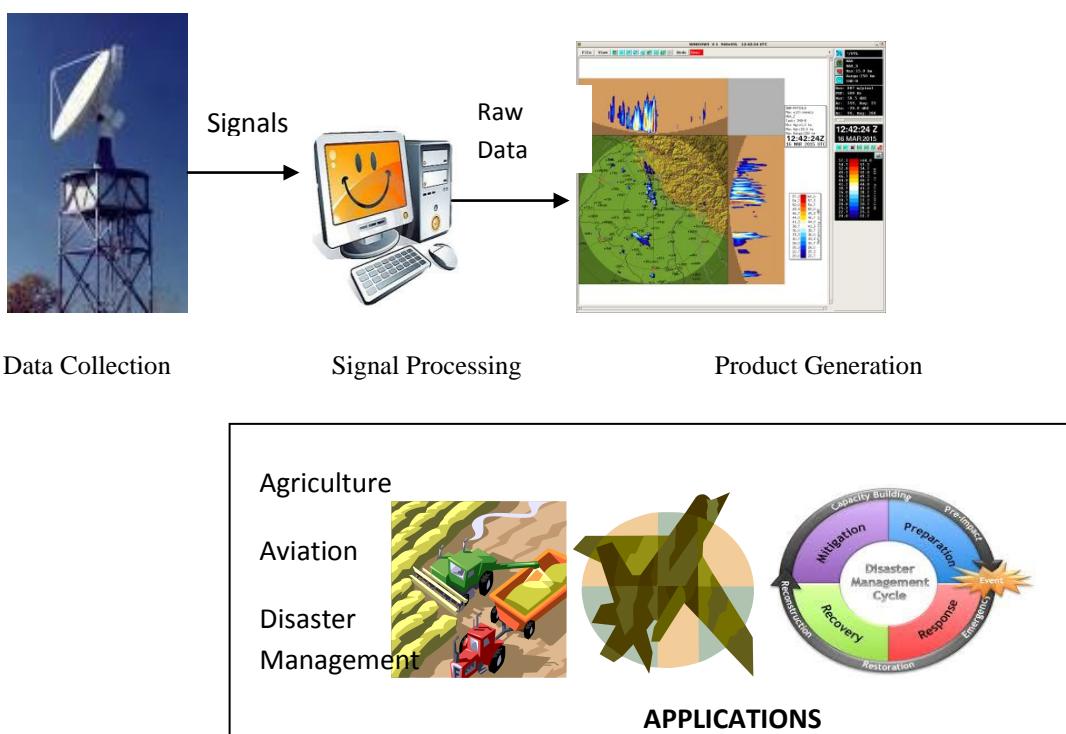


Fig.1 Radar Product Generation

## 2. Literature Survey

The field of Meteorology has gained importance over a period of time where weather affects various other major daily life domains such as Agriculture, Aviation, Disaster Management during cyclones, National budget and a lot more. Weather has been identified as a cause of 25% aviation accidents and contributed to fatalities. Severe weather warnings in certain cases can be quite efficient to prevent loss of life and property. Doppler Weather Radar is a very effective tool used for nowcasting and issuing warnings based on DWR products. Most of the researches focus on numerical weather forecasts and much work are not done to explore abundant radar products/images. There are a few researches done to detect the convective clouds which will form the basis of this project work on images of DWR Patiala.

Zhiying Lu, et al. in this paper, the author takes WSR 88D radar's reflectivity products as input image which is then converted, cut down and filtered (on the basis of pixels above certain threshold reflectivity) by matching the color scale in the image. Morphological filtering is then done by dilation, erosion and ellipse operator to find hailstone clouds. Further cloud extraction is done by edge detection Canny algorithm. Other features of hail clouds such as density, hook shape and high reflectivity core have been used to differentiate hailstone clouds from rainstorm clouds and to know the landfall point of hailstones. This work of author aims at helping the related department by informing to prevent loss due to hail phenomena. The author uses hook shape to predict the landfall point of hailstones. In case if hook cannot be found, author uses the criteria for pixel above 60 dBZ <sup>[4]</sup>. Hough Transform has been used to detect hook echo or detecting the pixel with more than 60dBZ reflectivity.

Hui Li, et al., in this research paper, the author considers high reflectivity to be the basis to define severity of weather to reroute the flights for safety. The image is segmented based on color feature by edge detection in the reflectivity image. Segmentation method for color image used is edge detection using Sobel operator. In the experiment, echoes were segmented by edge detection and then the boundaries of interested areas were extracted. Since, radar echoes for worse weather are corresponded to higher reflectivity, thus color is taken as the reasonable feature to segment images. Further, the fitting assessment with ellipse and polygon has been done to compare the results based on three indicators to assess fitting: accuracy, deviation and fitting. The author concluded polygon fitting method to be better than ellipse and the accuracy gradually decreases with the increasing length-width ratio of clouds <sup>[8]</sup>.

Ouarda Raaf, et al., in this paper, the researcher has made use of bidimensional Fourier analysis to clearly distinguish two main families of rainfall echoes in the images collected by the radars. These echoes represent either stratiform or cumuliform cells. The Fourier spectra obtained for the convective cells, are much wider and nearly ten times more intense than those characterizing the stratiform cells. Since the spectral features of radar echoes respectively arising from convective and stratiform cells, are found to be different, they can be efficiently used by radar operators to detect storms producing intense rains. However the researcher suggests further investigation as the segmentation method based only on Fourier spectra is not sufficient to distinguish between the two categories of clouds when their morphology is too close <sup>[10]</sup>.

Goswami, et al., in this paper, infrared satellite images are taken as input to detect the convective clouds by using k-means clustering algorithm (with Euclidean, Manhattan and Mahalanobis distances) for segmentation. Detection of clouds form IR images is based on the brightness depending on temperature difference at different areas. The high gray level values mean colder regions which in turn would mean clouds. The cluster with highest centroid value, comprised of pixels having very high values, was selected. The cloud track prediction techniques have been discussed to predict the path of movement of cloud cells. The track is predicted on the basis of CoM (Centre of Mass) by comparing displacement between two observations. CoM calculation for displacement of cloud clusters is done by the author using 3 models Difference Based Technique, Mean Path Adjustment Technique and AMPA Technique and the results are compared <sup>[12]</sup>.

D.Krezki, et al., the author present the need for storm tracking using Doppler radar images due to fast changing nature of storms. A spatio-temporal relaxation tracking algorithm <sup>[16]</sup> is proposed which stored the image in a pyramid structure and then split-merge based on threshold is done. The storm track would consist of temporal as well as spatial consistency properties. Various storm properties such as average intensity, storm size, similar velocity variance,

convexity and storm shape and orientation have been put under consideration to evaluate results and determine storm tracks.

Qingyong Li, et al., the author proposed a hybrid thresholding technique on ground based images acquires from sky imager applications to identify the cloud cover and cloud type etc. The author described the unimodal and bimodal images with respect to type of clouds and how HYTA can be better than fixed or adaptive techniques separately for both unimodal and bimodal images. The normalization of B/R ratio is done for improved visual contrast and noise robustness. Then, the hybrid thresholding is performed to distinguish clouds from sky. Identifying from the standard deviation of input ratio image, whether unimodal or bimodal, the fixed and minimum cross entropy(MCE) thresholding is applied appropriately.

Urban Meis, et al., the author discussed the radar image acquisition and interpretation for automotive traffic applications. The reflectivity of massive objects to find their relative speeds, distance etc. can help in obstacle warning, collision avoidance etc. Along with the identification of vehicles, different classes (cycles, cars, trucks) can be distinguished. Object detection, Road Course Prediction, Road Departure Warning etc. have been discussed by the author based on acquired radar images.

P. Kumar , et al., the author discussed a pre hail detection algorithm where the radar images are used for detecting potential convective cloud. This detection of area of interest and point of interest is done by manual clicking. Then the extraction of information is done to put into the algorithm [23] to detect the direction and speed of clouds. This forms the basis for formulating an approach where no manual intervention is involved from acquired image analysis till warning issuance.

### 3. Problem Formulation

Using Doppler radar images for image processing to identify the weather affected areas has not been a much explored area of research. Auto detection of convective clouds on the basis of radar reflectivity product through segmentation and feature extraction has been studied by researchers but in a limited scope.

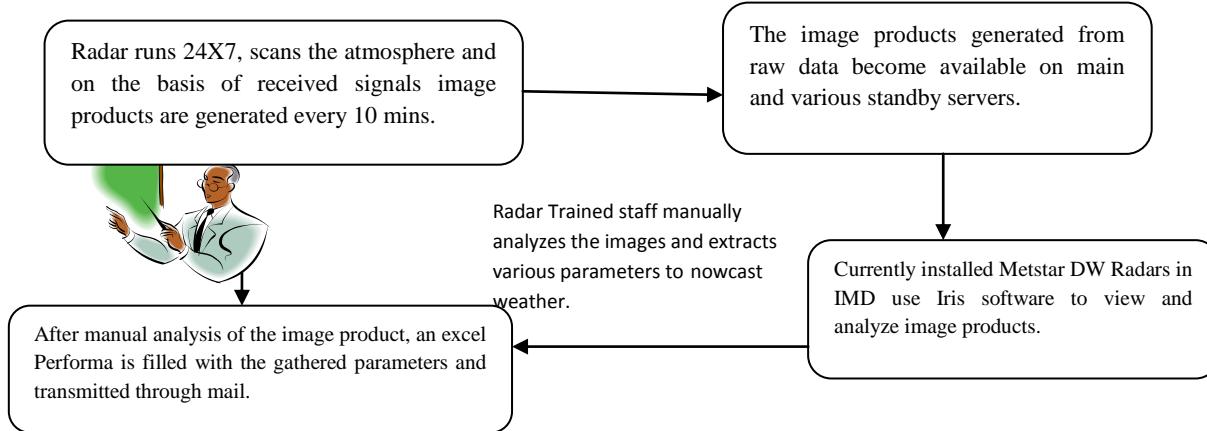


Fig.2 Current Process of Weather Analysis and Bulletin Generation at DWRs

It may be noted that this process of analyzing image for various parameters, creating a weather bulletin and issuing the same takes no less than 8-10 minutes with 80% accuracy. Doppler Weather Radar stations under India Meteorological Department do not use any automatic weather bulletin creation and issuance system based on digital image processing techniques.

It has been found that there is no automatic system for issuance of severe weather warnings to the affected area which causes a need of constant human monitoring to issue warnings. The radar image products, although freely accessible to all through internet, serve no purpose for the people who are not MET- trained. Therefore arises a need to auto detect convective clouds and the affected areas which could be issued warnings automatically in a common man readable form. Creating an automatic cloud detection system based on image processing and issuance of weather bulletins of existing accuracy in less than 2 minutes of product generation would improve the conventional system. Constant manual monitoring during hours of weather phenomenon can be reduced along with effective transmission to various related departments (Agricultural universities heads, Local Press, Head Met. Centre) in no time thus making a proactive warning approach.

#### 4. Proposed Approach and Methodology

The “approach for weather nowcasting from Doppler weather radar images using thresholding” proposes following outcomes

1. To detect the presence of rain bearing cloud patches from MAXZ product of DWR Patiala.
2. Classify the areas based on severity of impact of clouds (based on color scale).
3. Extract the names of the districts affected by convective clouds, their distance and direction from DWR Patiala.
4. To issue warning to the districts having severe weather conditions.

##### 4.1 Methodology

The methodology followed to achieve the above mentioned objectives can be summarized with the following points:

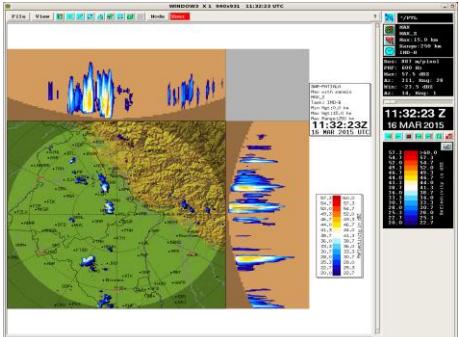
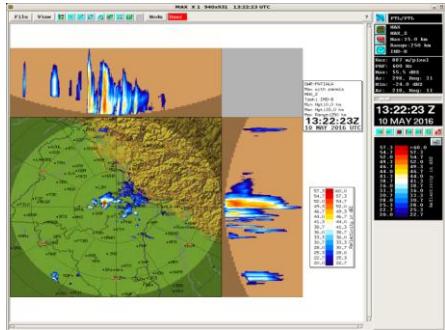
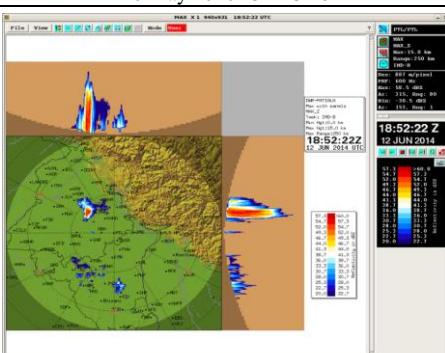
1. Image Acquisition: From Patiala, India DWR imageries of radar reflectivity was collected at 10 minutes interval for 10 elevations from  $0.2^{\circ}$  to  $21^{\circ}$  for a day with severe weather event. Automatic detection of cloud coordinates, reflectivity, time and direction is being aimed.
2. Feature Extraction: The radar images have been analyzed to extract the pixels with specific color (feature used) i.e. the pixels which fulfill the threshold range of colors from the color scale. Color has been used as the feature for thresholding in radar images as it represents the cloud reflectivity (which can be said as the water content in cloud).
3. Region Classification: Once, the required colors indicating possibility of severe weather have been identified in the radar image, the next step is to categorize them on the basis of their impact. These identified regions can be classified according to radar standards to categorize impacts <sup>[12]</sup> of weather. For example,
  - red: 54 dBZ (extremely heavy precipitation, with the possibility of hail)
  - orange: 49 dBZ (rain and thunderstorm)
  - yellow: 44 dBZ (clouds in developing state, may cause rain and thunderstorm)
4. Extract Region of Interest Information: Information such as the names of districts affected in and around the region of interest, its reflectivity value(in dBZ), distance of CB clouds from DWR Patiala and the direction of clouds with respect to Patiala is extracted and written to an excel file forming a Nowcast Bulletin.

5. Transmit Severe Weather Bulletins: Information extracted by automatically analyzing radar images for regions likely to be affected by severe weather is of social importance when it is broadcasted to the concerned departments for suitable preventive and proactive measures to be taken.

## 5. Result Discussion

Doppler weather radar images data set obtained from DWR Patiala have been analyzed and interpreted through image processing approach formulated to automatically detect Cumulonimbus clouds that are high in reflectivity and cause severe weather. The approach working on MAXZ radar product images for different dates are being discussed below:

TABLE 1: Interpretation of Results with new approach

Input Image	Conventional Result	New Approach
 16 March 2015 1132 UTC	Severe weather patches manually observed around Patiala, Nabha, Khanna, in the North-West Direction from Patiala. Warning Preparation and transmission takes 8-10 minutes	The algorithm identifies the cloud patches 24 km in North West direction from DWR and affected districts to be Patiala, Nabha, Khanna. Possibility of Heavy Rain and Thunderstorm warning issued. Total run time is less than 1 minute.
 10 May 2016 1322 UTC	Severe weather patches manually observed around Nabha, Khanna, in the West Direction from Patiala.	32 Km in West North West direction from DWR affecting Nabha Khanna and the adjoining areas. Possibility of Heavy Rain and Thunderstorm warning issued. Less than 1 minute execution time.
 12 June 2014 1852 UTC	Severe weather patches manually observed around Ludhiana, in the West Direction from Patiala.	71 Km in North West direction from DWR affecting Ludhiana, Nawanshahar and the adjoining areas. Greater Possibility of Hail, Heavy Rain and Thunderstorm warning issued. More than 2 minute execution time.

It is observed that the new approach is about 3 times faster than the conventional method that involves manual interpretation and information extraction from DWR images. The interpretation done by this auto cloud detection approach is with 90% accurate. The results also show an increase in run time for the algorithm with the increase in size of the cloud patch in radar image. As pixel by pixel method is used to check the threshold (color value), there is increase in execution time with bigger cloud patches as can be seen in the results of 12 June 2014 1852 UTC image.

The information extracted from interpretation of DWR images through newly proposed approach is automatically written to a excel file forming a ready to transmit severe Weather Nowcast bulletin. A snapshot of the same is shown below.

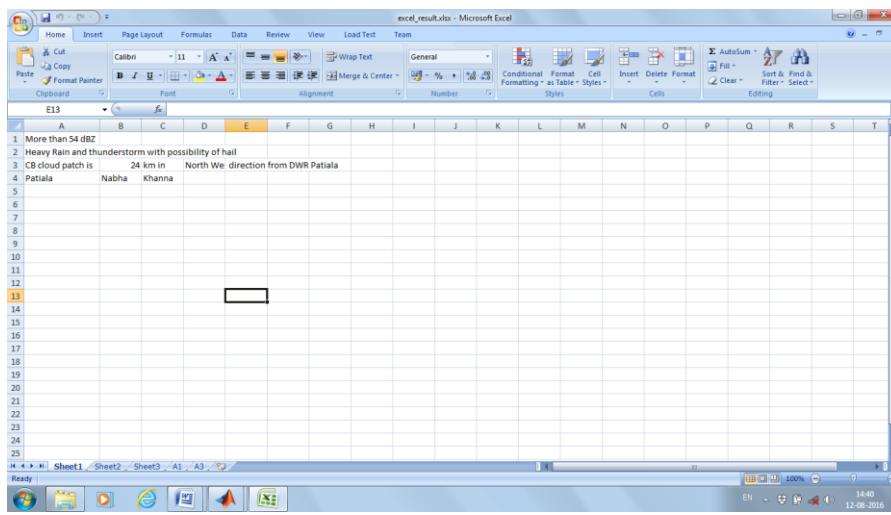


Fig. 3 Snapshot of Weather Nowcast Bulletin created through new approach

## 6. Conclusion and Future Scope

Doppler Weather Radar is an effective tool for weather forecasters for the purpose of nowcasting and has been stated better than satellites for the purpose of short term forecasting due to high resolution. Heavy Rains or thunderstorms caused by convective clouds in Plain Areas are considered severe weather and can be seen from maximum reflectivity in the MAXZ display. After the generation of digital image radar product with the inbuilt algorithms and parameter checks, manual analysis and generation of severe weather bulletins is done in DWR stations under India Meteorological Department. This work has been carried out on image products (Reflectivity product-MAXZ) of Doppler Weather Radar installed at Patiala (the only one in Punjab and Haryana).

A system of Rain Clouds Auto-detection from image products of DWR Patiala can facilitate severe weather warnings to be issued to affected districts of Punjab, Haryana and Himachal Pradesh. This approach successfully detects and classifies the areas/districts where clouds would be in developing or mature state and there exist the possibility of Rain/Thunderstorms/Hail based on the reflectivity value. Along with this, the name of affected districts, location of convective clouds (distance and direction w.r.t. to Patiala) has been automatically identified for issuance of severe weather warnings without any human intervention.

The future orientation concerns with the detection of direction of movement of convective clouds (along with other parameters such as height of clouds) so that warnings can be issued with the lead time of 15-30 minutes. Also, the work can be taken to a higher level by studying the increase in core of clouds to detect development or dissipation of the clouds. Finding out the possibilities for process application on compressed images like JPEG and 3D rendered images and improvement in the current work can be under consideration for future work.

#### REFERENCES

- [1] Shapiro and Stockman,"Image Segmentation for Computer Vision", March 2000.1
- [2] "Introduction to Image Processing Chapter-10 Segmentation", Dept. of Information and Computing Sciences, University of Utrecht2
- [3] Cuneyt Gecer, "Training Course on Weather Radar Systems, Module D: Radar Products and Operational Applications", World Meteorological Organisation (WMO), Turkish State Meteorological Service (TSMS), 2005. 11
- [4] Zhiying Lu, Yufeng Yang, "A Study of Hailstone Detection System Based on Radar Echo Reflectivity Image", 2006.5
- [5] Andre Aichert, "Feature Extraction Techniques", 2008.3
- [6] Vaisala Inc, "IRIS/RDA Utilities Manual", 2010. 15
- [7] "WSR 98D/S Doppler Weather Radar", System Description Manual, METSTAR Training Material.4
- [8] Hui Li, Xiaohao Xu, and Xinglong Wang, "Research on Boundary Extraction and fitting Analysis for Convective Clouds Based on Rerouting Strategy", 2011.6
- [9] Ashraf A. Aly, Safaai Bin Deris , and Nazar Zaki, "Research Review of Digital Image Segmentation Techniques", 2011. 10
- [10] Ouarda Raaf, Abd El Hamid Adane, "Pattern recognition filtering and bidimensional FFT-based detection of storms in meteorological radar images", 2012.7
- [11] Barnali Goswami, Gupinath Bhandari, "Convective Cloud Detection and Tracking from series of Infrared images", 2012.8
- [12] NWFC India Meteorological Department, "Guide Book on Nowcast-Challenges, Development and Opportunities in Nowcasting", 2012.9
- [13] Lockheed Marteen, "NEXRAD-WSR-98D Doppler Weather Radar", Ocean, Radar & Sensor Systems, Syracuse, New York 13221-4840. 12
- [14] M. Jogendra, GVS RajKumar, and R Vijay Kumar Reddy, "Review on image Segmentation Techniques", 2014. 13
- [15] Shilpa Kamdi and R K Krishna, "Image Segmentation and Region Growing Algorithms", IJCTEE Vol2, Issue1. 14
- [16] D. Krezeski, R. E. Mercer, J. L. Bamn, P. Joe, and H. Zhang, "Strom Tracking in Doppler Radar Images", IEEE.
- [17] Rohan Kandwal, Ashok Kumar, and Sanjay Bhargava, "Review:Existing Image Segmentation Techniques", 2014.
- [18] Anil Z Chitade, and S.K. Katiyar, "Color Based Segmentation using K-means Clustering".
- [19] Diljeet Singh, Navdeep Kanwal, "An Approach to Steganography using Local Binary Pattern on CIELAB based K-Means Clustering", 2015.
- [20] LIANG-KAI-HUANG and MAO-JIUN-J.WANG, "Image Thresholding by Minimizing the measures of Fuzziness", 1995.
- [21] Urban Meis and Robert Schneider, "Radar Image acquisition and interpretation for automotive applications", 2003 IEEE.
- [22] Qingyong Li, Weitao Lu and Jun Yang, "A Hybrid Thresholding Algorithm for Cloud Detection on Ground-Based Color Images", 2011.
- [23] P. Kumar and Deba Prasad Pati, "Radar imageries information extraction and its use in pre-hail estimation algorithm", Mausam, 66, 4(October 2015), 695-712.

# Digital Rights Management in India: A Study of Legal Mechanism to Curb Digital Piracy

Amjad Ali

UGC-JRF, Department of Law, Punjabi University, Patiala.

E-mail: amjadalithind@gmail.com

*Abstract- The intellectual property (IP) regime grants the exclusive right in the form of Copyright to the creators of digital works to use and distribute those works for their financial benefits. But with the advancement in digital technology, the copyright realm is facing the challenge of ‘digital piracy’ i.e. unauthorized use of and access to copyrighted works. To tackle it, the IP community has advanced a mechanism called Digital Rights Management (DRM). India, in order to bring its copyright law in conformity with international standards and to prevent the revenue loss to original content creators caused by wide spread evil of digital piracy, has introduced in 2012, inter alia, three new provisions to the Copyright Act, 1957 under its Section 2 (xa), Section 65A and 65B to specifically deal with Digital Rights Management. The research paper is an attempt to analyse and evaluate the efficacy of these new provisions and suggests some concrete submissions to remove various lacunae and to strengthen mechanism of the digital rights management vis-a-vis the rights of consumers.*

**Keywords:** Digital Piracy, Digital Rights Management, RMIs, TPMs.

## 1. INTRODUCTION

The economic realm postulates that everyone is entitled to earn financial incentives out of his works. Such right will be vague and meaningless unless there is some mechanism to enforce its application. The Intellectual Property (IP) regime has also recognized and granted the creator of literary, dramatic, musical and artistic works, cinematograph films and sound recordings,[1] a bundle of rights including, *inter alia*, rights of reproduction, communication to the public, adaptation and translation of the work, in the form of Copyright.[2] These copyrights shall have no meaning if any unauthorized person uses those works without acknowledging the financial rights of the original creator. Therefore, like all other rights, the law of copyright demands protection against its infringement. Following the principle, States have taken initiatives to safeguard the rights of copyright holders of digital contents, *i.e.* the data produced and stored in digital format which includes software, applications, text, graphics, images, audio, video etc. from unauthorized use, access and distribution. Consumer’s demands to get these copyrighted works in digital formats have resulted into the emergence of the evil of ‘digital piracy’ which some authors have described as “an illegal act of copying digital goods without explicit permission from and compensation to the copyright holder.”[3]

The digital world has raised the voice against the expanding roots of digital piracy and the need to curb it. The global IP regime, which is the outcome of Trade Related Aspect of Intellectual Property Rights (TRIPS) Agreement, has answered the problem with catena of international instruments including, *inter alia*, WIPO Copyright Treaty, 1996 (WCT) and WIPO Performances and Phonograms Treaty, 1996 (WPPT) also known as ‘internet treaties’. It

has formulated technologies like Rights Management Information (RMI) and Technological Protection Measures (TPM) enveloped under the shield of Digital Rights Management (DRM) to manage the ownership and handling of rights of digital works.

## 2. DIGITAL RIGHTS MANAGEMENT: MEANING, SCOPE AND SIGNIFICANCE

The emergence of personal computer during early 90s and availability of cheap and affordable computer resources and services including CD, DVD, USB, Internet, eBooks, P2P sharing escalated the enigma around digital piracy. It has resulted into pecuniary loss to the legitimate creators and royalty holders when someone other than them copies and resells their digital works without any authorization and paying costs. The RMI and TPM as tools of DRM were introduced during 1994 for the first time as an elixir to control the use, access and handling of digital works.[4]

The term ‘Rights Management Information’ (RMI) is legally defined to be meant as “the information which identifies the work, the author of the work, the owner of any right in the work, or information about the terms and conditions of use of the work, and any numbers or codes that represent such information, when any of these items of information is attached to a copy of a work or appears in connection with the communication of a work to the public.” [5]. The purpose of RMI is of an electronic watermark or a label of identification. It embeds the information about the owner as well as terms and conditions to digital works. Though easy to change, reproduce, alter and put into free distribution without the owner’s consent, the technique helps the owner to trace the illegal use of these works, when there is circumvention of the embedded information. Even if there is no circumvention at all, the unauthorized person can be held guilty for non-compliance with other terms and conditions, if he makes unauthorized use of such work. It is equally important for consumers as it ensures the quality and authenticity of work.

The RMI metadata *i.e.* the data which describes other data so embedded provides the recognition to authors, performers etc. of original works. They can affix their names, addresses, identification marks, barcodes, ISBN, ISSN, copyright notices, dates, disclaimers etc. to ensure their identity and authenticity as to the original piece, either in written words or in machine readable form. A foremost example of it can be seen while playing music where on display screen of media player one can easily find the name of the artist, album, composer, lyricist, publisher, release date etc. appended to the music commonly known as MP3 Tagging. [6] To sum up it can be said that in copyright terms, RMI frequently serves as a means of compliance with *the moral right of attribution*, in that it identifies the author and performer of a work. The WIPO Internet Treaties protect all such RMI: information about works, phonograms and performances, as well as the identification of authors, phonogram producers, performers or other rights owners. [7]

On the other hand, the Technological Protection Measures (TPM) or simply Technological Measures as defined under Article 6.3 of the INFOSOC Directive means “any technology, device or component that, in the normal course of its operation, is designed to prevent or restrict acts, in respect of works or other subject-matter, which are not authorised by the right-holder of any copyright or any right related to copyright as provided for by law.” [8] The

TPM technology is used to protect the digital contents with the help of technological measures like encryption of software, passwords, access codes, restrictive permissions etc. TPMs are of two types: access control TPMs and copy protection TPMs. [9] Access control TPMs allow the owner to control access by way of password protections, file permissions and encryption whereas the copy protection TPMs are designed to control activities such as reproduction of copyrighted material. One of the main differences between the two is that an access control TPM will block access generally, while a copy protection TPM will operate at the point where there is an attempt to do an act protected by the copyright such illegal copying, transferring etc. However, it is still possible that TPM can be circumvented through different computer programs or electronic devices.

So, there is a fundamental difference in the mode of operation of these two technologies as TPM is designed to curtail access or copying, while RMI technology does not curtail access or copying *per se*, but rather provide an environment in which various types of use, including copying, are only practically possible in compliance with the terms set by the right holders. [10]

### 3. LEGISLATIVE MECHANISM RELATING TO DRMs IN INDIA

Earlier this was the view that private sphere of consumers shall be kept out of the ambit of the Copyright laws. Individuals were, therefore, free to make any use of their legally obtained copyrighted copies. But now transformation of society from Maine's conception of contract to collective status has blurred the line between the private and public domains. With invention of digital reproduction techniques, the 'reprography' of digital contents becomes comparatively easy and cheap. [11] To save financial loss to the original creator's, copyright laws have been formulated and applied by the countries under the control of World Intellectual Property Organisation (WIPO), established in 1967 as one of the specialized agencies of United Nations. Since then WIPO has become an effective international organization in the development of a balanced and effective international IP system. As of present WIPO administers 26 treaties to regulate the IP regime surrounding Trademarks, Patents, Copyrights, Geographical Indications etc. at international level. Out of these, the WIPO Copyright Treaty, 1996 and the WIPO Performances and Phonograms Treaty, 1996, specifically emphasize on the protection of copyrighted contents in digital environment. India, though not a signatory to these two 'internet treaties', has applied, to the extent which it considers necessary and desirable in relation to the Indian context, to its Copyright Act, 1957 due to constitutional mandate under Article 51. [12] As a result, it has amended its Copyright Act, 1957 in 2012 and inserted three new provisions, *inter alia*, under sections 2 (xa), Section 65A and 65B dealing with Digital Rights Management.

#### A. Section 2 (xa)

Section 2 (xa) provides the definition of "rights management information (RMI)." The provision, in the form of interpretation clause, is inserted to clarify the extent of applicability of DRM provisions under Section 65B and is inclusive in nature. It defines RMI as an information including the identification of work or performance, name of the author or performer, name and address of the owner of rights, terms and conditions regarding the use of the rights and includes any number or code that represents the information mentioned before. Some common examples

of these can be found in the form of digital file properties, EULA (End User License Agreement), use of words like ‘for home use only’, ‘for use on company specific devices’ like the *Fairplay* system of Apple, [13] etc. So RMI as a part of DRM, *de facto*, could be seen as the imposition of unilateral contractual terms and conditions.

**B. Section 65A**

Section 65A has been introduced for use of technological measures to protect owner’s rights on digital copyrighted work. Under sub-section 1, it provides punishment for circumvention of an effective technological measure applied for the purpose of protecting any of the rights conferred by this Act. The offence is made punishable with imprisonment which may extend to two years and shall also be liable to pay fine. However, sub-section 2 lays down the following exceptions under which circumvention of TPM is permitted, where a person:

- a) doing anything referred to therein for a purpose not expressly prohibited by this Act: Provided that any person facilitating circumvention by another person of a technological measure for such a purpose shall maintain a complete record of such other person including his name, address and all relevant particulars necessary to identify him and the purpose for which he has been facilitated; or
- b) doing anything necessary to conduct encryption research using a lawfully obtained encrypted copy; or
- c) conducting any lawful investigation; or
- d) doing anything necessary for the purpose of testing the security of a computer system or a computer network with the authorisation of its owner; or
- e) operator; or
- f) doing anything necessary to circumvent technological measures intended for identification or surveillance of a user; or
- g) taking measures necessary in the interest of national security.

This Proviso under Section 65A (2) has maintained a balanced approach between the rights of copyright holders as well as of consumers. Its spirit corresponds with language of Article 11 of WCT and Article 18 of WPPT. [14]. The exceptions provided under sub-section 2 are to minimize the impact on public access to copyrighted works. The circumvention of TPM is permitted for purposes which are not expressly prohibited by the Act, for research, legal investigation, testing with authorization of owner or operator, securing individual’s privacy or national security. Section 65A does not exclude the right of Fair Dealing, guaranteed under section 52 of the Copyright Act, 1957 unlike that of the stringent US counterpart Digital Millennium Copyright Act, 1998 (DMCA). [15].

**C. Section 65B**

Section 65B prevents the removal of RMI embedded to the digital contents. It provides protection to the content owners by prohibiting removal of information containing in those works without authority or purchasing modification rights. The language of Section is punitive in nature and prescribes that any person, if found guilty of removal or alteration of RMI or distribution, import for distribution, broadcasting, communicating to the public, copies of any work or performance knowing that RMI has been removed or altered without authority, then he or she

shall be punishable with imprisonment up to two years and shall also be liable to pay fine. The language of Section 65B conforms to Article 12 of WCT and Article 19 of WPPT.

Apart from criminal remedies, the proviso attached to the language of section 65B enables the owner of copyright to avail civil remedies under Chapter XII of the Act. [16] Chapter XII containing provisions from section 54 to section 62, both inclusive, entitles the owner of copyright with remedies by way of injunction, damages, accounts and otherwise as are or may be conferred by law for the infringement of a right, protection of separate rights when there are more than one owner, authors special rights, right of owner against persons possessing or dealing with infringing copies, restriction on remedies in the case of works of architecture, remedy in case of groundless threat of legal proceedings, owner of copyright to be party to the proceeding and jurisdiction of court over matters arising under this Chapter.

#### 4. DRMs *vis-à-vis* CONSUMER RIGHTS

Whereas the DRM mechanism tries to provide the content owners with control over their works to prevent digital piracy, the consumers, on the other hand, are left with comparative less rights as a bonafide purchaser of copyrighted works. The digital content owners are now able to affix excessive restrictions in the form RMI or TPM which intervene with consumer's right to full enjoyment of digital contents they purchase with their hard earned money. Sometimes they are not even able to play the same CD in all their devices, thanks to TPM restrictions. The phenomena have led to an era of war between DRM lobbies versus consumer organisations on various points:

##### *A. Excessive Cover of DRMs violates Right of Fair Use*

It is possible for a content owner to circumscribe the digital use of his works with the help of TPM and RMI with tools like passwords, codes, read-only permissions, copy protection, terms of use etc. Traditionally the consumers have always been free to read, listen or view the copyrighted creations for learning or enjoyment. The race for accumulation of more and more money and hunger of fame has made the content owners to invent restrictive DRM methods to put monopoly control over their works. The excessive use and cover of DRM methods proves to be a disaster for fair use by bonafide consumers. There is no definition of the term 'private use' or 'fair use' given under any of the international instruments or the Copyright Act, 1957 which further entitles the content managers to go to any extent while using the DRM techniques to prevent digital piracy. What comes out of this scenario, are the aggressive restrictions on the fair use right of end-consumers, where they are unable to fully enjoy the works of their favorite authors, performers etc.

##### *B. DVD Region Codes*

Sometimes the use of digital contents is restricted to only a region. For example, a movie DVD released in USA cannot be played in India. Similarly, a game released in India may be restricted only for use in Indian subcontinent only. All this is possible with technology of Region Codes. Not only playback of its contents, Region Codes are also capable to control aspects of a released work including price, release date etc. The technology minimises the rights

of consumers as they cannot play a music CD/DVD/BD at home they had bought during foreign tours, causing trauma and wastage of money, all hail to Region Codes technology.

#### C. Disclosure of Information

The content owner, sometimes, does not find it necessary to disclose the restrictions embedded to the digital works. For example, if I buy and play a movie DVD in my home DVD player, it may play normal. But as soon as I insert it into my Laptop's DVD drive, it stops working. If there is anyone to blame here, then it is that content owner who has not mentioned the clause like 'for playback on DVD systems only', or 'not for use in computer' etc. on the outer-cover, thus concealing information and causing monetary as well as mental loss to his consumer. The Consumer Protection Act, 1986 does not specifically talk about the protection of Consumers from DRM. However, under Section 6(b) of the Act, consumers are given the right to receive information about the quality, quantity, potency, purity, standard and price of goods or services. Hence, interpreting the statute, content owners selling DRM / TPM encrypted material in India may be obliged to disclose the same to the consumer.

#### D. Tracking infringes Right to Privacy

Some of the TPM technologies are able to copy certain codes and programs into computer files. These codes or programs are meant to send tracking information to the content owner about the use of his works. The tracking report, sometimes, also collects personal information about the users including his routine, internet interests, working hour's etc. thus leading to the infringement of right to privacy. A well-known case is of Sony RootKit which was included on Sony audio CDs and was designed to prevent music on the CD being transferred to a computer and then burned to another CD. This TPM installed a Rootkit which left the user's computer vulnerable to attacks by malware or spyware. The users were usually unaware of that the Rootkit had been installed and therefore their machine was vulnerable. Legal action was taken against Sony and they recalled all the CDs that included the Rootkit. [17] The functions like 'Do Not Track' and 'Ad-Blocker Plus' are developed to protect user's privacy.

#### E. Exaggerated Use of TRIPS Mandate

Article 61 of the Trade Related Aspects of Intellectual Property Rights (TRIPS) agreement provides that "members shall provide for criminal procedures and penalties to be applied at least in cases of wilful trademark counterfeiting or copyright piracy on a commercial scale. Remedies available shall include imprisonment and/or monetary fines sufficient to provide a deterrent, consistently with the level of penalties applied for crimes of a corresponding gravity. In appropriate cases, remedies available shall also include the seizure, forfeiture and destruction of the infringing goods and of any materials and implements the predominant use of which has been in the commission of the offence. Members may provide for criminal procedures and penalties to be applied in other cases of infringement of IP rights, in particular where they are committed wilfully and on a commercial scale."

The provision prescribes the punishment in case of *piracy on commercial sale*. However, the practical approach to the provision has failed to recognize the difference between commercial and non-commercial copying for consumer's personal use. Thus a consumer making a backup copy of recently purchased 'rare collection CD' might

be held guilty for digital piracy. Indian legal regime does not follow this stringent view and by way of section 52 of the Copyright Act, 1957 it authorizes ‘fair dealing’ with copyrighted works, where making backup copies, *inter alia*, is kept out of the ambit of definition of infringement of Copyright.

## 5. MAJOR FINDINGS

One has to understand that at the root of every innovation rests the evil to diminish its value. For digital world this evil is in the form of digital piracy. The present study finds the following about use of DRM techniques to put a bar on digital piracy:

- Digital Piracy exists at a vast scale throughout the world.
- The unauthorized use of piracy techniques causes severe monetary loss to the copyright holders.
- The DRM system is formulated to tackle the problem.
- Techniques like RMI and TPM are used for strengthening Digital Rights Management.
- The copyright regime follows a balanced approach both in favor of copyright holders as well as consumers.
- Actual enforcement of Copyright laws is still not made as required.
- Complicated application of digital technologies is a big setback to stop digital piracy.
- Excessive and adverse use of DRM techniques violates consumer’s rights.

## 6. SUBMISSIONS

Following are the submissions to be considered for desired results:

- Digital piracy can be curbed only after proper enforcement of legal provisions.
- Punishments should be made more stringent to deter the pirates.
- Strict software programming can be used while giving proper authorization to the consumers so that it does not intervene with their enjoyment of their purchased works.
- The DRM techniques should not be used for tracking personal information of consumers.
- Moral teachings should be imparted to youngsters about the problem and its effects to the society at large.
- Awareness programs to make people attentive about piracy and pirated copies and the consequences of its use should be conducted.
- The Digital world should respect the right to fair use of their consumers.

## 7. CONCLUSION

The Copyright (Amendment) Act, 2012 introduced three provisions, *inter alia*, to bring a balance between the owners of copyrighted digital contents and the consumers. These provisions prove to be effective, but not with such gravity for which these were introduced. The conditions prevailing in Indian society as well as digital world did not allow the Indian parliament to follow the stringent approach as provided under ‘Internet Treaties.’ Accordingly,

it has not ignored the rights of consumers while granting the digital rights to copyright owners in the form of RMI and TPM. DRM technologies are invented to save the economic aspects for content owners. The use of these technologies must not yield to satisfy lust for money and accumulation of inventive resources to a handful of people. Preferably the public domain should remain paramount for law while granting rights, either to its individuals or a particular group of individuals, here in this case the group of copyright holders. No right can come in the way of nation's commitments towards its citizens for providing equal rights. One such right demands the desirable use and handling of purchased works by the citizens for their own purposes. The copyright world in association with digital environment should not be provided with that extent of monopoly rights where they can sell their works without assigning the full access rights to their clients. Restrictions by way of RMI and DRM must be used in 'minimal version' *i.e.* for identification, labelling and protection. It should not go beyond the criteria set by law and infringe the rights of its clients *i.e.* right to privacy and fair use. So it will be for the good of society if these DRM technologies are used as rights and not as 'restrictions' in extensive and adverse forms.

**References:**

1. Section 13 of the Copyright Act, 1957.
2. Section 14 of the Copyright Act, 1957.
3. Gennaro F. Vito and Jeffrey R. Maahs, *Criminology*. Burlington, USA: Jones & Bartlett Publishers, 2015, p. 296.
4. M. Myska. (2009). The True Story of DRM. *Masaryk University Journal of Law and Technology*. [Online] 3(2). p. 267. Available: <https://journals.muni.cz/mujlt/article/view/2540>
5. Article 12.2 of the WIPO Copyright Treaty, 1996.
6. MP3 Tag FAQ: How does an MP3 tag work? [Online]. Available: <http://mp3.about.com/od/digitalmusicfaq/f/ID3-Tag-Faq.htm>
7. The WIPO Treaties (2003): Protection of Rights Management Information [Online]. Available: <http://www.ifpi.org/content/library/wipo-treaties-rights-management-information.pdf>
8. Directive 2001/29/EC of The European Parliament and of The Council on the harmonisation of certain aspects of copyright and related rights in the information society. (2001) Official Journal of the European Communities. [Online]. Available: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:167:0010:0019:EN:PDF>
9. I. H. Bhat (2013). Technological Protection Measures Under Copyright. *International Journal of Emerging Trends & Technology in Computer Science* 2(2), pp. 320-321.
10. Review of TPM exceptions, Report of House of Representatives, Standing Committee on Legal and Constitutional Affairs, Australia (2006). [Online]. Available: [http://www.aph.gov.au/Parliamentary\\_Business/Committees/House\\_of\\_representatives\\_Committees?url=laca/protection/report.htm](http://www.aph.gov.au/Parliamentary_Business/Committees/House_of_representatives_Committees?url=laca/protection/report.htm)
11. Section 2 (x) of the Copyright Act, 1957 defines reprography as "*the making of copies of a work, by photocopying or similar means.*"
12. Though India is not a member to WCT and WPPT, it has showed respect towards international comity by adopting the fundamentals laid down under these two international conventions. Moreover, India has adopted and rectified other international conventions including Berne Convention of 1886, the Universal Copyright Convention of 1951, the Rome Convention of 1961 and the Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPS) which patently deal with Copyright law.
13. The FairPlay technology is a Digital Rights Management (DRM) technique created by Apple, Inc., based on technology created by the company Veridisc. FairPlay is built into the QuickTime multimedia software and used by the iPhone, iPod, iPad, Apple TV, iTunes, and iTunes Store and the App Store. [Online]. Available: <https://en.wikipedia.org/wiki/FairPlay>
14. Article 11 of WCT and Article 18 of WPPT provides:  
*"Contracting Parties shall provide adequate legal protection and effective legal remedies against the circumvention of effective technological measures that are used by authors/performers/producers of phonograms in connection with the exercise of their rights under this Treaty or the Berne Convention and that restrict acts, in respect of their works/performances/phonograms, which are not authorized by the authors/performers/producers of phonograms concerned or permitted by law."*

15. Arnab Naskar and Shubhangi Gupta. (2012). Digital Rights Management: A Pandora's Box Trying to Wipe off the Rights of Consumers. *Indian Journal of International Property Law* [Online] 5. pp.45-66. Available: <http://www.liofindia.org/in/journals/INJIPLaw/2012/5.html>
16. Proviso to Section 65B of the Copyright Act, 1957 lays down:  
*"Provided that if the rights management information has been tampered with in any work, the owner of copyright in such work may also avail of civil remedies provided under Chapter XII against the persons indulging in such acts."*
17. Technological Protection Measures (TPMs), [Online]. Available: [http://www.flinders.edu.au/library/copyright/technological\\_protections.cfm](http://www.flinders.edu.au/library/copyright/technological_protections.cfm)

# ***Linear and non-linear equalizer for chromatic dispersion compensation in coherent optical receiver***

## ***A comparison using different QAM modulation formats***

GURPREET KAUR<sup>1\*</sup>, GURMEET KAUR<sup>2</sup>  
Corresponding author\*: [dhammu.gurpreet836@gmail.com](mailto:dhammu.gurpreet836@gmail.com)

**Abstract—**Optical fibers are now widely used in communication systems, mainly because they offer much faster data transmission rates. But by increasing transmission rates over longer distances, the data is affected by nonlinear inter-symbol interference caused by dispersion phenomena. Efficient and low-cost dispersion compensation is necessary to achieve useful transmission distances in optical communication systems at higher bit rates. Adaptive equalizers can be used to compensate the effects caused by nonlinear intersymbol inference, i.e., restoring the original signal as transmitted. This paper evaluates the performance of single layer feed forward equalizer using least mean square algorithm and multilayer perceptron equalizer based on artificial neural networks using back error propagation algorithm for chromatic dispersion compensation at bit rate of 12.5 Gb/s. This paper compares these two adaptation techniques in coherent receiver based on analogue electronic dispersion equalizers by simulation and experiment.

**Keywords—** Artificial Neural Network (ANN), Bit Error Rate (BER), Electronic Dispersion Compensation (EDC), Feed-Forward Equalizer (FFE), Sign-Sign Least Mean Square (SS-LMS), Multilayer Perceptron with Back Error Propagation (MLP-BP).

### I. INTRODUCTION

Optical Fiber impairments such as chromatic dispersion (CD) severely influence the performance of high speed optical fiber communication systems [1, 2]. Although many optical fiber communication systems use dispersion compensation fibers (DCFs) to compensate the chromatic dispersion, but this increases the complexity and cost of the communication systems. Generally, an equalizer for the compensation of chromatic dispersion performs like an inverse filter, which is placed in the receiving side of the communication link. An equalizer transfer function is the inverse of the associated link transfer function [3], and it reduces the errors between the desired and estimated signals. The traditional equalization techniques based on the finite impulse response (FIR) provide a significant performance improvement for a different number of communication channels [4]. Now a days adaptive channel equalizers are significant in communication systems [4]. Digital coherent receivers allow equalization for linear transmission impairments in the electrical domain [5, 6], and have become the most promising alternative approach to dispersion compensation fibers. While coherent detection was experimentally confirmed as early as 1979, its use in actual systems has been delayed by the additional complexity, due to the need to track the phase and polarization of the incoming signal [7]. With the introduction of more advanced modulation formats i.e., QAM modulation format in coherent system the channel capacity and spectral efficiency increases. But this increases the constellation complexity. This increase in constellation complexity brings with it increased nonlinear impairments. For the Compensation of non-linear impairments multilayer perceptron equalizer based on artificial neural networks can be used. ANN equalizer has a number of advantages when used for a time-varying environment, including the adaptability, a nonlinear decision boundary and parallel processing capabilities [3, 8]. The remainder of this paper is organized as follows. Second section presents the theory of equalization for nonlinearities produced by optical fiber using single layer feed forward equalizer, third section describes the ANN equalization, whereas in section fourth the operating principle of

MLP is outlined. Fifth section presents algorithms used for training the equalizer i.e., sign-sign least mean square algorithm for linear equalizer and back error propagation algorithm for the non-linear equalizer. A section sixth provides the comparison between algorithms using experimental results. The conclusion is presented in section seventh.

## II. SINGLE LAYER FEED FORWARD EQUALIZER

A feed-forward equalizer is the simplest type of linear equalizer and its output is produced by summing the present and past values of the received signal which linearly weighted by the equalizer coefficients. The basic structure of an adaptive equalizer is shown in Fig. 1, where the subscript  $k$  is used to denote a discrete time index. Note that in Fig. 1 there is a single input  $y_k$  into the equalizer at any time instant. The adaptive equalizer has  $N$  delay elements,  $N + 1$  taps, and  $N + 1$  tunable complex multipliers, called filter weights or filter coefficients. These weights are updated continuously by using an adaptive algorithm.

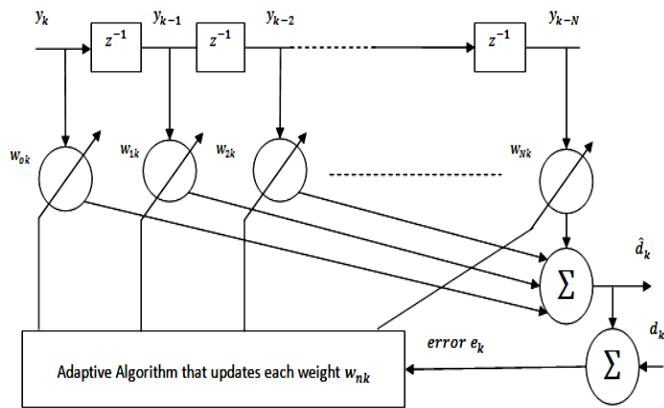


Fig. 1. A basic linear equalizer during training [9]

The adaptive algorithm uses an error signal to minimize a cost function which is mean square error (MSE) between the desired signal and the output signal of the equalizer based on the classical equalization theory [9], [10]. The MSE is denoted by  $[e \ k \ e^* \ k ]$ , and a known training sequence must be periodically transmitted when a replica of the transmitted signal is required at the output of the equalizer. The equalizer coefficients are computed by the sign-sign least mean square (SS-LMS) method because normalized LMS gives better performance than standard LMS algorithm [11].

## III. THE ANN EQUALIZATION

ANN is a mathematical model of massively connected parallel-distributed processors, made up of simple building blocks called neurons, which are based on the principle and behavior of the biological neural networks. The ANN information learning capabilities made it possible to solve complex problems and applications such as nonlinear system identification, motor control, and pattern recognition. Moreover, due to the large parallel interconnection between different layers of neural network and the nonlinear processing characteristics, adaptive ANN has the ability to perform nonlinear mapping between a set of inputs and an output space, therefore, the ANN is now mainly used for equalization [12-15], finance [16], control system [17], statistical modeling [18], and engineering applications [19]. Because of these capabilities of artificial neural networks in efficiently showing arbitrary nonlinearities, there has been recent interest in employing them in adaptive equalization for optical communication channels [20-22]. In this case, the linear adaptive filter is replaced by an artificial neural network. Different artificial neural network architectures such as multilayer perceptron, radial basis functions, and recurrent neural networks have all been offered in the literature for channel equalization [23]. Among all these structures, the most commonly and widely-used is the multilayer

perceptron structure. The popularity of MLP-based equalizers is due in part to their computational simplicity, stability, finite parameterization, and smaller structure size for a particular problem as compared to other structures.

#### IV. MULTILAYER PERCEPTRON

A multilayer perceptron consists of several hidden layers of neurons that are able to perform complex, nonlinear mappings between the input and output layer. The hidden layers provide the capability to use the nonlinear function to create intricately-curved partitions of space with complex nonlinear decision boundaries [24]. Furthermore, it has been presented that only three layers are needed by the MLP to generate these boundaries [25]. The basic element of the multilayer perceptron is the neuron.

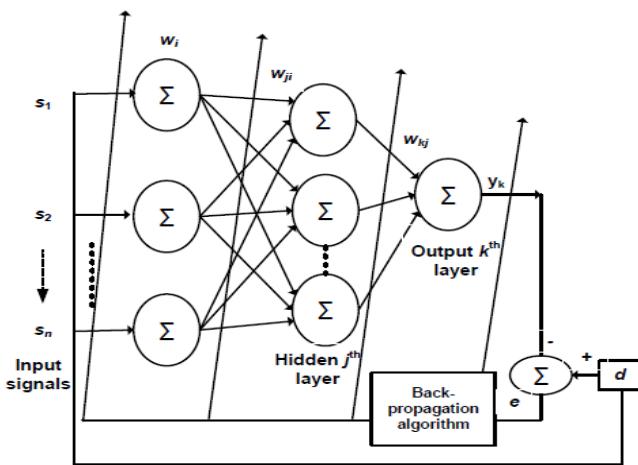


Fig. 2. *MLP Network [4]*

Fig. 2 shows a four layers MLP, in the figure  $s$  denote the inputs  $[s_1, s_2, \dots, s_n]$  to the network input layer, and  $y_k$  denotes the output of the final layer. The layers linking weights between the input to the first hidden layer is  $w_i$ , first to second hidden layer is  $w_{ji}$  and the second hidden layer to the output layer is  $w_{kj}$ . Therefore, the output signal at the final output layer of the MLP can be expressed as [4]:

Where  $P_1, P_2, P_n$  is the number of neurons in the MLP network layers,  $b_i, b_j, b_k$  is the threshold to the neurons of each specific layer, and  $\varphi^+$  is the nonlinear activation function. The most popular activation functions are sigmoid and the hyperbolic tangents as these are differentiable functions. In this study, the activation function of the hidden layer,  $\varphi^+$ , is a hyperbolic tangent function and the output layer function,  $\varphi^+$ , is a linear function. The MLP did not receive much consideration in applications until the introduction of the BP algorithm [26].

## V. ALGORITHMS USED FOR TRAINING

This section presents the basic type of algorithms used for updating the coefficients of linear and nonlinear equalizer.

#### A. Sign-Sign Least Mean Square Algorithm

The filter coefficients are computed by the sign-sign least mean square (SS-LMS) method in linear equalizer, because it demonstrates the simplicity and robustness needed for realization in very high speed circuits [11].

Step 1: Initialize the weights and set tolerable mean square error value.

Step 2: Present the input vectors,  $x(n)$  and desired output  $u(n)$ .

Step 3: Calculate the actual output  $y(n)$  from the input vector sets.

Step 4: Adapt weight based on the following equation.

Step 5: Go to step 2.

where  $w(n)$  is updated weights,  $w(n - 1)$  is previous weights,  $e(n)$  is error signal,  $u(n)$  is actual input signal and  $\mu$  is convergence parameter whose value varies from 0 to 2,  $\alpha$  is leakage factor whose value varies from 0 to 1.

### B. Back Error Propagation Algorithm

A two layer feed-forward network can overcome some of the restrictions associated with the single layer network. But it did not solve the problem of adjusting the ANN weights from input to hidden layers. A solution to this problem was presented in 1985 by Parker, and by Rumelhart in 1986 [28]. The solution is based on the concept that the hidden layer neuron errors are determined by back-propagating the errors of the neurons of the output layer. That's why this method is often called back-propagation learning rule [28]. The main advantage of the BP algorithm is that its hardware circuit can be easily realized [28]. The BP supervised learning algorithm is one of the most popular training algorithm used for adaptive the weights in multilayer networks [28], to minimize the cost function  $E(n)$ :

In the BP algorithm, the weights are updated by applying gradient descent on the cost function; this is in order to reach a minimum value of the cost function. The weights are updated according to:

Where  $w_{ij}$  is the weight from the hidden unit  $i$  to the unit  $j$ , and  $\eta$  is the learning rate parameter. When  $\eta$  is very small, the algorithm will take more time to converge, in contrast when it is very large; the system may produce oscillations and causes instability [25].

The BP algorithm procedures can be summarized as [25]:

Step 1: Initialize the weights and thresholds to small random numbers.

Step 2: Present the input vectors,  $x(n)$  and desired output  $d(n)$ .

Step 3: Calculate the actual output  $y(n)$  from the input vector sets.

Step 4: Adapt weight based on equation (3).

Step 5: Go to step 2.

## VI. EXPERIMENTAL SIMULATIONS AND RESULTS

Simulations were made for different QAM modulation formats, with no channel encoders, at fiber length of 50km. Simulation parameters used for this study have been summarized in Table 1.

<b>Table 1: Optical Channel Parameters</b>	
<b>Parameters</b>	<b>Values</b>
Bit rate (B)	12.5Gbps
Wavelength ( $\lambda$ )	1550nm
Speed of Light (c)	$3 \times 10^8$ m/s
Chromatic Dispersion (D)	17ps/nm-km
Fiber Length (L)	50km
Extinction Ratio ( $r_{ex}$ )	6.6dB
Leakage Factor ( $\alpha$ ) in LMS-SS	0.1
Convergence Parameter ( $\mu$ ) in LMS-SS	0.8
Learning Rate ( $\eta$ ) in MP-BP	1

The results obtained with linear and nonlinear equalizer using Sign-Sign Least Mean Square and Back error propagation algorithm respectively by performing various experiments, have been summarized in Fig. 3 and Fig. 4.

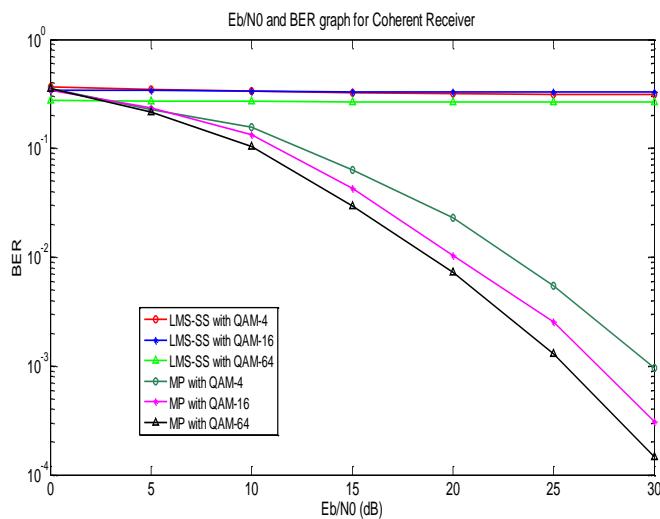


Fig. 3. shows the Bit error rate (BER) versus  $E_b/N_o$  ratio using Sign-Sign Least Mean Square Algorithm and Multilayer Perceptron with Back Error Propagation algorithm for different modulation formats .

Fig. 3 shows the Bit error rate (BER) versus  $E_b/N_o$  ratio using Sign-Sign Least Mean Square Algorithm and Multilayer Perceptron with Back Error Propagation algorithm for different modulation formats at fiber length of 50km and typical dispersion value of 17ps/nm-km. When the  $E_b/N_o$  ratio varies between 0 to 30dB the percentage change in BER is 15.228 for QAM-4 with LMS-SS, 99.58 for QAM-4 with MP-BP, 4.18 for QAM-16 with LMS-SS, 99.89 for QAM-16 with MP-BP, 2.84 for QAM-64 with LMS-SS, 99.95 for QAM-64 with MP-BP. At  $E_b/N_o$  ratio 30dB the value of BER is  $3.084 \times 10^{-1}$ ,  $3.278 \times 10^{-1}$ ,  $2.663 \times 10^{-1}$ ,

$1.421 \times 10^{-3}$ ,  $3.412 \times 10^{-4}$ , and  $1.502 \times 10^{-4}$  for QAM-4 with LMS-SS, QAM-16 with LMS-SS, QAM-64 with LMS-SS, QAM-16 with MP-BP, QAM-4 with MP-BP, QAM-64 with MP-BP respectively.

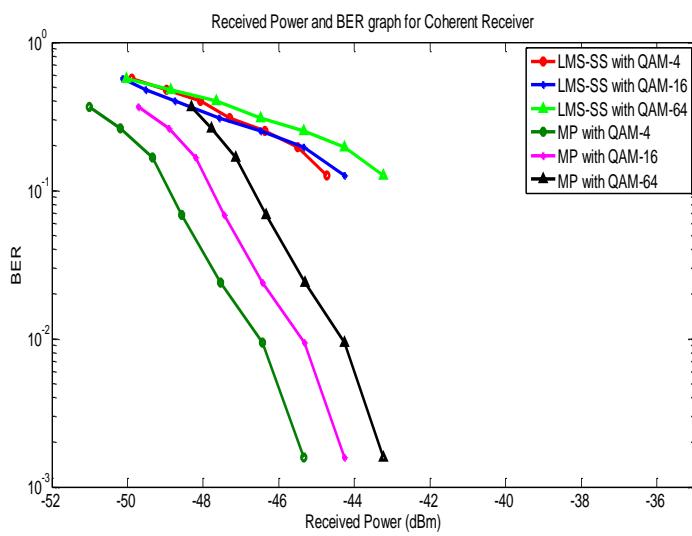


Fig. 4. Bit error rate (BER) versus Received Power (dBm) using Sign-Sign Least Mean Square Algorithm and Multilayer Perceptron with Back Error Propagation algorithm for different modulation formats.

Fig. 4 shows the Bit error rate (BER) versus Received Power (dBm) using Sign-Sign Least Mean Square Algorithm and Multilayer Perceptron with Back Error Propagation algorithm for different modulation formats at fiber length of 50km and typical dispersion value of 17ps/nm-km. With the percentage change in BER of 78.8, the increase in received power should be 10.142 for LMS-SS with QAM-4, 11.69 for LMS-SS with QAM-16 and 13.57 for LMS-SS with QAM-64 respectively. With the percentage change in BER of 99.58, the increase in received power should be 11.37 for MP-BP with QAM-4, 11.035 for MP-BP with QAM-16 and 10.552 for MP-BP with QAM-64 respectively. At BER  $1.234 \times 10^{-1}$  the percentage change in received power should be 3.375 when the modulation format changes from QAM 4 to QAM 64 for LMS-SS. At BER  $1.421 \times 10^{-3}$  the percentage change in received power should be 4.653 when the moduation format changes from QAM 4 to QAM 64 for MP-BP.

**Table 2: Measured BER at  $E_b/N_o = 20$  dB and Typical dispersion value of 17ps/nm-km**

Type of Modulation Format	SS-LMS	MP-BP
QAM-4	0.3309	0.0179
QAM-16	0.3201	0.0098
QAM-64	0.2680	0.0071

It has been shown in Table 2 that there is improvement in value of Bit Error Rate by using Multilayer Perceptron with Back Error Propagation algorithm as compared system with Single layer Feed Forward Equalizer using Sign-Sign Least Mean Square algorithm. Moreover it has been concluded from this table that the BER improvement in case of QAM-64 is more as compared to other types of QAM formats used for equalization. The measured value of Bit Error Rate is  $2.68 \times 10^{-1}$  and  $7.1 \times 10^{-3}$  for SS-LMS and MP-BP technique respectively at 20dB  $E_b/N_o$  ratio.

## VII. CONCLUSION

It has been found from this study that nonlinear equalizers are preferred over their linear counterpart because the linear equalizers do not perform well enough on channels which have deep spectral nulls in the pass-band. In an effort to compensate for the distortion, the linear equalizer places too much gain in the vicinity of the spectral nulls, thereby enhancing the noise present in these frequencies. Linear equalizers view equalization as an inverse problem while non-linear equalizers view equalization as a pattern classification problem where equalizer classifies the input signal vector into discrete classes based on transmitted data.

By using different form of channel equalization techniques to mitigate the effects of the interference in communication system it has been observed that MLP equalizer is a feed-forward network trained using BP algorithm, it performed better than the linear equalizer, but it has a drawback of slow convergence rate, depending upon the number of nodes and layers.

Another conclusion from this study is that capacity and spectral efficiency increases by using more advanced modulation format i.e., QAM-64. Means QAM-64 provides better results as compared to QAM-4. But by using more point QAM technique the constellation complexity increases. QAM-64 provides the balance between capacity and constellation complexity. Therefore in this study QAM-64 has been used.

## REFERENCES

- [1] P.S. Henry, IEEE J. Quantum Electron. 21 (1985) 1862.
- [2] G.P. Agrawal, Fiber-Optic Communication Systems, 3rd ed., John Wiley & Sons, Inc., New York, 2002.
- [3] S. Haykin, "Adaptive filter theory, ". NJ, USA: Englewood Cliff, Prentice Hall, 1991.
- [4] D. R. Guha, "Artificial neural network based channel equalization," Master Thesis, Department of Electronics and Communication Engineering, National Institute of Technology, Rourkela, 2011
- [5] J.G. Proakis, Digital Communications, 5th ed., McGraw-Hill Companies, Inc., Boston, 2008.
- [6] H. Bulow, F. Buchali, A. Klekamp, J. Lightwave Technol. 26 (2008) 158.
- [7] M.G. Taylor, IEEE Photonics Technol. Lett. 16 (2004) 674.
- [8] K. Burse, R. N. Yadav, and S. C. Shrivastava, "Complex channel equalization using polynomial neuron model," in *International Symposium on Information Technology, ITSim*, pp. 1-5, 2008.
- [9] T.S. Rappaport, Wireless Communications-Principles and Practice. 2nd Edition, Prentice Hall, 1996.
- [10] S.U.H. Qureshi, Adaptive Equalization. *Proceeding of IEEE*, **37**, 1340-1387, 1985.
- [11] A. Shoval, et al., Comparison of DC Offset Effects in Four LMS Adaptive Algorithms. *IEEE Transactions on Circuits and Systems-II*, **42**, 176-185, vol. 42, no. 3, pp. 176–185, 1995.
- [12] A. N. Ramesh, et al., "Artificial intelligence in medicine," *Annals of The Royal College of Surgeons of England*, vol. 86, pp. 334-338, 2004.
- [13] H. B. Burke, "Evaluating artificial neural networks for medical applications," Presented at the *international Conference on Neural Networks, Houston, TX, USA*, vol. 4, pp. 2494-2495, 1997.
- [14] S. C. B. Lo, et al., "Application of artificial neural networks to medical image pattern recognition: detection of clustered microcalcifications on mammograms and lung cancer on chest radiographs" *Journal of VLSI Signal Processing*, vol. 18, pp. 263-274, 1998.
- [15] A. M. Nogueira, et al., "Using neural networks to classify Internet users," Presented at the *Advanced Industrial Conference on Telecommunications/Service Assurance with Partial and Intermittent Resources*, pp. 183-188, 2005.
- [16] S. Walczak, "An empirical analysis of data requirements for financial forecasting with neural networks," *Journal of Management Information Systems*, vol. 17, pp. 203-222, 2001.
- [17] F. Yue and T. Chai, "Neural-network-based nonlinear adaptive dynamical decoupling control," *IEEE Transactions on Neural Networks* vol. 18, pp. 921-925, 2007.
- [18] G. Dorffner, "Neural networks for time series processing," *IEEE Transactions on Neural Network World*, vol. 6, pp. 447-468, 1996.
- [19] A. Patnaik, et al., "Applications of neural networks in wireless communications," *Antennas and Propagation Magazine, IEEE*, vol. 46, pp. 130-137, 2004.
- [20] S. Siu, et al., "Decision feedback equalization using neural network structures and performance comparison with standard architecture", IEE Proceeding 137, 221–225, 1990.
- [21] S. Chen, et al., "Adaptive equalization of finite nonlinear channels using multilayer perceptrons", *Signal Process*. 20, 107–119, 1990.
- [22] G.J. Gibson et al., The application of nonlinear structures to the reconstruction of binary signals, *IEEE Transactions Signal Processing* 39, 1877–1884, 1991.
- [23] B. Mulgrew, Applying radial basis functions, *IEEE ASSP Mag*. 13, 50–65, 1996.
- [24] A. Wieland, R. Leighton, "Geometric analysis of neural network capabilities", *1st International Conference on Neural Networks*, June 1987, pp. 385–393.
- [25] S. Haykin, Neural Networks: A Comprehensive Foundation, Macmillan Co., New York, 1994.
- [26] D.E. Rumelhart, J.L. McClelland, "Parallel Distribution Processing: Explorations in the Microstructure of Cognition", vol. 1, MIT Press, Cambridge, MA, 1986.
- [27] B. Krose and P. S. Smagt., Introduction to neural networks, 8th ed., The University of Amsterdam, 1996.

# Pre-requisites of Big Data and Hadoop

Jagbir Singh, Student, Rakesh Singh, Assistant professor (Department of Computer Engineering,  
Engineering Wing, Punjabi University, Patiala)

**Abstract-**Data is increasing at a speed that is even difficult to capture in today's world as a result of regular uploading of data to the internet. Social networking websites are the main sources for the collection of this huge amount of data. This collection of online data happens at a great pace. "Big Data" is a collection of data sets that is extremely large in numbers highlighting its size; it varies from one another i.e. different varieties of data; and velocity of collection of data i.e. its speed. This extreme amount of data is difficult to handle by traditional software packages. Hence it becomes necessary to introduce a new and efficient method to handle this huge data and then came into existence is Hadoop. Hadoop is an open source project that enables separation of processing of different datasets using simple programming model along with high rate of fault tolerance. Hadoop consists of two main parts: HDFS for storing large amount of data and MapReduce for analysis of data. It becomes easy to handle even petabytes or zettabytes of data with the help of this Java based software i.e. Hadoop.

## I. INTRODUCTION

Big data is a collection of large datasets that contains enormous amount of data. This data can't be handled by traditional software packages as it exceeds their capability limits. The reason why big data is generated depends on the dependency of the whole world on the Internet. Many great searching websites like Google, Yahoo along with most common social networking sites like facebook, instagram, whatsapp, viber etc allowed sharing of data between the users of their websites that eventually result in appearance of large amount of data.

Big data is used to show the data that is really huge in amount (Size); that is appeared at a very large speed (Velocity) and different formats (Variety) that makes traditional software packages like relational databases difficult to captured, analyzed, processed and stored that particular data. As cleared from its name, "Big Data" is generally "Big" in number or size; it means it usually exist in zettabytes or terabytes in memory exceeding the capability limit of traditional software to handle this huge data.

### A. WHY BIG DATA?

Most of the people are using internet for the sharing of data with their friends, mates and known's. Facebook, Twitter, Instagram and LinkedIn are the most commonly used communication websites now a day on the internet. According to a survey, facebook, which is the most commonly used communication website developed by Mark Zuckerberg, is having 1.393 billion monthly active users; 890 million daily active users. 350 million photos are regularly uploading on facebook by its users and 4.75 billion items are regularly shared on it at a daily basis . Its like button is pressed 2.7 billion times a day. Facebook is responsible for storing around 300 Petabytes of data and on an average it is responsible for accepting 600 Terabytes of data at daily basis [1]. The storage of this data is increasing regularly and the worst part is, this data is never deleted but the upcoming data is consistently adding to the existing data. It means when new petabytes of data is getting added to the existing terabytes of data; it becomes zettabytes of data missing none of the information from the internet.

### B. VARIOUS APPLICATIONS OF BIG DATA

Big Data is a term given to this regularly increasing data with no deletion of previously stored data on the internet. Big data comprises three types of data. These are Traditional enterprise data; machine generated or sensor data and the social data. The enterprise data includes the data from CRM i.e. Customer Relationship Management systems, ERP i.e. Enterprise Resource Planning, Web store Transactions etc. The Machine generated data includes CDR i.e. Call Detail Records, smart meters. The social Data includes feedback from the customers, social media platforms etc[2]. Big data is having many applications in today's world from hospital

systems to large-scale analytics. In the case of health care industry, with the help of combination of all the information of the patient from his past diseases; the treatment can be done by the doctor at a comfortable level. Germany had done collection of big data in FIFA World Cup 2014 with the help of a tool named Match Insights that was prepared by a German Software Company named SAP. This tool was able to analyze data from the video with the help of cameras from on-field. The speed of the player and the position were produced with this tool along with much other important information. And eventually, by the assistance of the collected result the coach was able to give the instructions to the mobile phones of the players. The same kinds of applications of big data also exist in many other games like basketball and tennis [3]. Moreover, Big Data is also used in many other areas including military decision making, advance natural disasters warning, surveillance and controlling crowd etc.

## II. CHARACTERISTICS OF BIG DATA

Big Data is characterized with the help of three basic following V's. This Consists of the following:-

- a. Volume: Big data comprises collection of large amount of data that is generally huge in numbers like zettabytes, petabytes or terabytes of data. The existing data is not get deleted from the internet but the new data remains on adding along with the existence of the already existed data making it large after every session. Hence, Existing petabytes of data cross the limit of Petabytes and become Zettabytes after the addition of some terabytes into it.
- b. Velocity: Big data is collecting to the internet at a very fast pace. Here, Speed always matter. For example; there exists some hard real time software where response is required within split of seconds; so if the data is getting late, some disaster could be happened possibly. Other dimension of velocity is, for how much time the data will get stored? Will it be permanent or not in case of storage? Other dimension specifies the speed at which the data is stored and processed. There exist many cases where five minute response time is very much more than the required.
- c. Variety: Big data is different in formats. It is not only structured data that was the only available option in relational databases. Instead, it is a semi-structured or unstructured data that comes in many different forms. It could be text files, audio clips or video clips, data generated from the e-mails, blogs etc.
- d. Value: Value is very important aspect of big data to be considered. In relational databases; group results were produced as an outcome; but big data gives more personalized result as an outcome. Big data is having huge potential value. With the help only the required amount of data is produced with the help of some specified value.
- e. Veracity: Veracity originally describes the provenance of the data. It shows whether the data produced have come from the reliable sources. It represents the noise, biases etc. There are no chances that we'll get 100% of accurate data i.e. free from dirty data when we are dealing with bulk amount of data.
- f. Variability: It specifies whether the data produced is in consistent form or not? We should have knowledge of the fact that the data so produced is either useful or it is dirty data?
- g. Virability: It is the rate at which data spreads. It is the transformation of data by different events..

## III. HADOOP

Hadoop is an open source software package developed by Doug Cutting. It is responsible for handling distributed data sets having large amount of data. It is developed using programming language Java. In hadoop, clusters of commodity machines are generated and the total work is coordinated between the clusters. Hadoop is scalable, open source, fault-tolerant virtual grid operating systems architecture for data storage and processing. Hadoop allows applications based on MapReduce to run on large clusters of commodity hardware. Hadoop is developed to parallelize data processing across computing nodes to speed computations and hide latency. With hadoop, the analytics can be done on the whole of the dataset rather than small part of data. Hadoop consists of two main parts including HDFS and MapReduce platform.

a) HDFS

It is a file system responsible for storing large amount of data by saving that on commodity clusters. It generally have high throughput, high fault tolerance along with low cost on hardware. It works by breaking incoming data into small parts called “blocks”. HDFS is responsible for storing three copies of each distinct file by copying it on three different servers. Usually the size of each block in data node is 64 MB. It is based on Master/Slave architecture. HDFS contains the following three nodes: Name Node, Data Node and Edge Node.

i) *NAME NODE*: Name node acts as a master node. It is centralized node. It is responsible for maintaining metadata, records and location of the blocks in the data node along with all the other information whether it is related to addition of new data or the modification of existing data nodes.

ii) *DATA NODE*: There exist more than one data nodes. Data nodes act as slave nodes. Data nodes contain blocks having a large amount of data inside them. It also acts as a framework for running different jobs.

iii) *EDGE NODE*: It is also known as HDFS clients. It behaves like a connecting link both between Data node and the Name node. Usually, it is one in number; but it can be more than one depending on the performance needs.

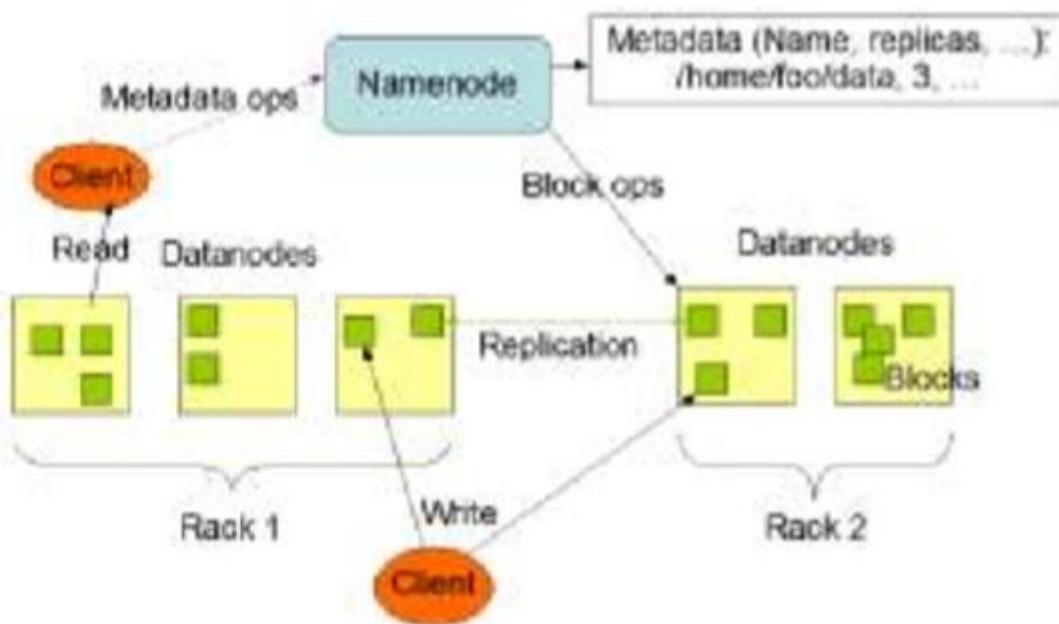


Fig 1: HDFS Architecture [12].

b) MAPREDUCE:

MapReduce is, developed by Google, like a programming model for analyzing large datasets. Mapreduce system distributes incoming data to multiple chunks, then a map task is assigned that can process data parallel. It consists of two different phases: Map and Reduce. The set of key and values are used both at Map and Reduce level for transformation and production of data depending on these pairs. The “Map” Components actually distribute the programming problems or task across a large number of systems and handle the placement in a way that manages the load and handles recovery from the failures. After distributed computation is completed, another function “reduce” aggregates all the elements back together to generate the final result. MapReduce consist of a Job Tracker and the different Task trackers. Job tracker works as a master and assigns different tasks to its task trackers. The task trackers, on the other hand, act as slaves that handle the tasks given by the job tracker. Task trackers remained in a consistent communication with the job tracker to update themselves for

future jobs. On the other hand, if any task node don't behave properly or become unresponsive; then it is the duty of the job tracker to assign task to some other task tracker.

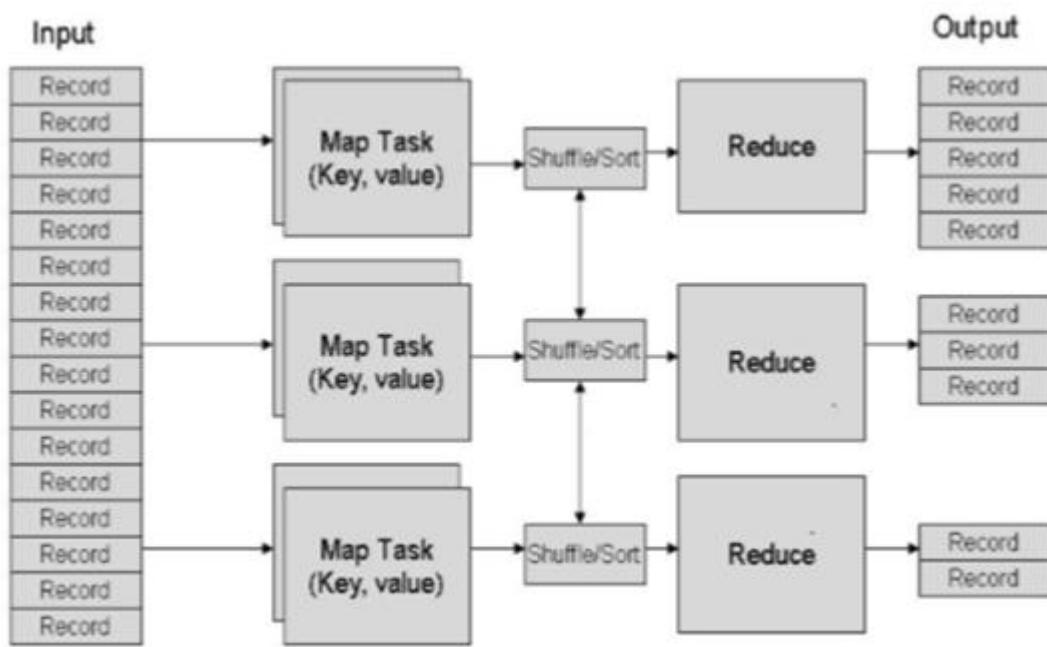
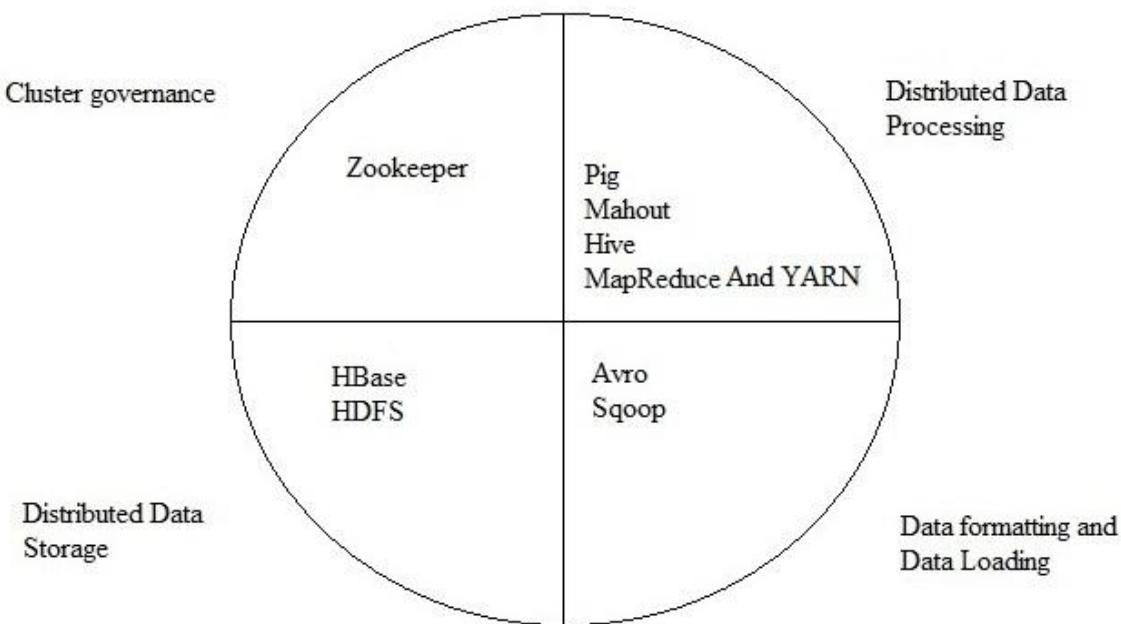


Fig 2: Mapreduce Architecture and working

#### IV. SOME COMPONENTS OF HADOOP DISTRIBUTED FILE SYSTEM:

- 1) HBases: it is Distributed Column oriented Database system in contrast to HDFS which is a file system. It is open source, non relational, distributed database system written in programming language called Java. It is based on BigTable of Google. It acts as both input and the output for the MapReduce.
- 2) Avro: Avro is responsible for bringing data interoperability among multiple hadoop components. It believes that data produced by one component should be consumed by some other component. It also supports many programming languages like Java and python etc.
- 3) Pig: Pig is big data analytical platform that also ensures processing of this data. With the help of pig, data processing jobs becomes very easy.
- 4) Hive: Hive is a data ware housing framework. SQL queries can be processed under hive to process and analyze the data
- 5) Sqoop: Sqoop is a command line interface tool that is used for the transformation of data from relational components like oracle, SQL into hadoop environment.
- 6) Mahout: Mahout is a machine learning library that is also used in data mining. It consists of four parts as: collective filtering, categorization, clustering and mining of parallel frequent patterns.
- 7) Zookeeper: it is a centralized service that provides distributed coordination and synchronization. It will be useful in hadoop if any particular node's failure come to know in advance so that the necessary steps can be

taken.



8) Fig 3: Different functionality of HDFS Components

### Conclusion

In this paper, I have tried to cover basic concepts related to big data. Big data is applicable to many areas including healthcare, sports, financial industry, IT industry and entertainment industry. If big data is handled carefully, it will give useful results. Moreover, the needed information regarding Hadoop is summarized along with the Hadoop components and its future scope. We can write different MapReduce Programs for big data analysis.

### REFERENCES

- [1] Craig Smith. (2015, March 19). By the Numbers: 200+Amazing Facebook User Statistics (February 2015). [Online].Available: <http://expandedramblings.com/index.php/by-the-numbers- 17-amazing-facebook-stats>
- [2] Jean Pierre Dijcks, “Oracle: Big Data for the Enterprise,” in Oracle White Paper, 2013© Oracle Corporation.
- [3] Steven Norton. (2014, July 10). Germany’s 12th man at The World Cup: Big Data. [Online]. Available: <http://blogs.wsj.com/cio/2014/07/10/germany-s-12th-man- at-the-world-cup-big-data/>
- [4] Shilpa Manjit Kaur, “Big Data and Methodology”- A review”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.
- [5] Andrew Pavlo, “A Comparison of Approaches to Large-Scale Data Analysis”, SIGMOD, 2009.
- [6] Apache Hadoop: <http://Hadoop.apache.org>
- [7] Dean, J. and Ghemawat, S., “MapReduce: a flexible data processing tool”, ACM 2010.
- [8] DeWitt & Stonebraker, “MapReduce: A major step backwards”, 2008. [9]Hadoop Distributed File System, <http://hadoop.apache.org/hdfs>
- [10] HadoopTutorial: <http://developer.yahoo.com/hadoop/tutorial/module1.html>
- [11] J. Dean and S. Ghemawat, “Data Processing on Large Cluster”, OSDI ’04, pages 137–150, 2004
- [12] J. Dean and S. Ghemawat,“MapReduce: Simplified Data Processing on Large Clusters”, p.10, (2004
- [13] Greenplum Analytics Workbench ,visit us at [www.greenplum.com](http://www.greenplum.com) .

# Review on Big Data and Hadoop

Simranjeet Kaur, Student And Navroz Kaur Kahlon, Assistant Professor (Department of Computer Engineering, Engineering Wing, Punjabi University, Patiala)

**Abstract-**The term "Big Data" refers to the data which is really "Big" in size. In today's world, data is generating each and every second with very high pace that makes the traditional software incapable to manage such data as it is both huge in capacity and generate at fast speed. Hadoop is an open standard architecture that is actually a platform for handling this huge data. It contains HDFS for saving this bulk amount of data and MapReduce to deal with the speed of the data. One of the world's largest social networking site facebook's like button is clicked 2.7 billion times a day across the world generating ample amount of data within small period of time. This data can be handled with the big data concepts.

## I. INTRODUCTION

### A. Need of Big Data:

It has become a major problem in today's world to deal with large amount of data. Consider a file of large size (say 1 TB) on our machine. The problem that the user will face with this bulk of data could be:-

- It requires ample amount of time in order to access those files
- There would exist problem of performing analytical operations on that file.

Approximately in between 1950 to 2002, it was data warehousing in traditional systems that was responsible to deal with the data. The demerits of existing traditional system included:-

- It was unable to handle structured or semi-structured data i.e. images or videos etc.
- It was unable to process very large files like PDFs, Excel files etc.
- Inability to process semi-structured data like XML, log files and Sensor Data.

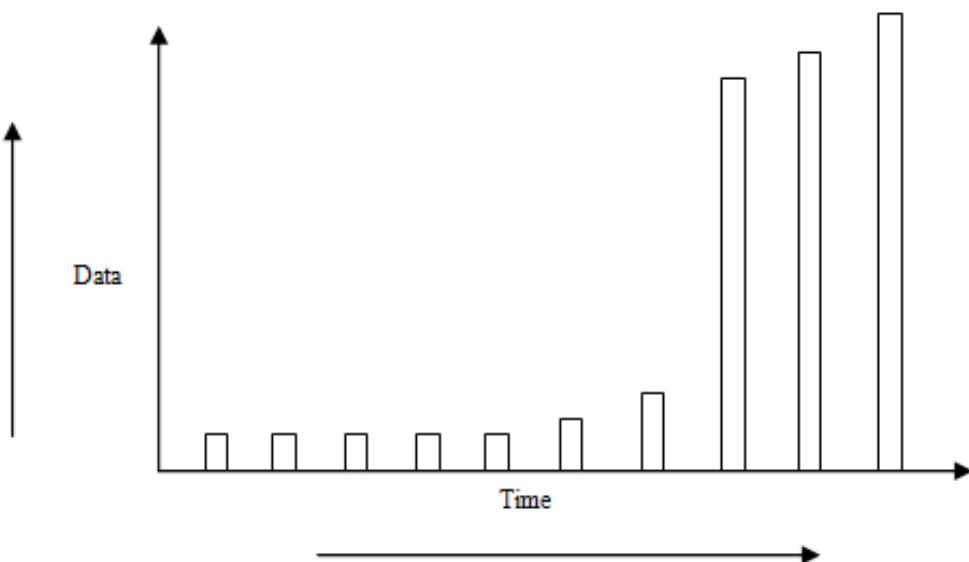


Fig 1: Data Growth Rate (with respect to time)

- It didn't respond to enormous amount of data as well.
- It was having fixed Schema.
- Cost issue was critical that couldn't be ignored.

b) *History of Big Data*

In 2002, it was Google that imagined these limitations of existing database systems of saving huge files of web addresses originated on its servers which could be faced in the future as the amount of data was increasing abruptly. Around 90% of data have come into the universe in just last two years. So Google started preparing a white paper (Called Google Distributed File System) for dealing with intense amount of data. But the problem here appeared of saving these files manually. Google wanted such kind of framework or platform with the help of which these files could be saved automatically. At that time, one of the research student named Doug cutting(inventor of Lucene Search Engine) had a look on this paper. He started performing implementation on it. Now, Yahoo had a look on this student's work and it hired Doug cutting. After successful implementation of the project within two years, yahoo submitted the full-fledged software (i.e. Hadoop) to apache foundation.

In the mean time, Google worked preparing another white paper called "Google Mapreduce algorithm". Again, Doug cutting had look on it and he, with the help of yahoo, start doing implementation on. Eventually, Yahoo gave the implemented project to the apache foundation.

## II. Big Data

Big Data can actually be said to the collection of data i.e. datasets which is typically not possible to handle for the traditional database system. Generally, the size of big data starts from terabyte or zettabyte. As its name describes, the big data is "Big" is size; but there is still no formal definition of big data. Big data is responsible to manage huge volume of data with right time and with the right speed.

IBM gave the definition of big data in its own way:

Big data is the data that is generally too big in size and that generates too fast and it is made up of following four characteristics, generally called four V's:-

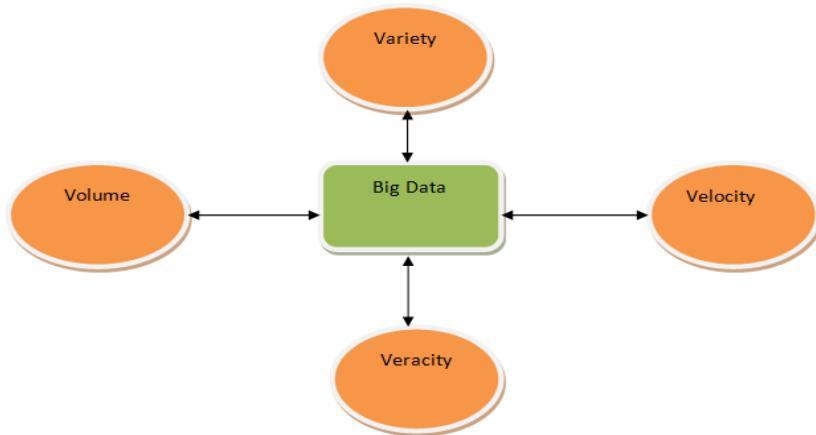


Fig 2: Characteristics of Big Data

Here, the first V called "Volume" signifies the amount of data to be stored. It may be terabytes or in zeta bytes. It is generally huge in number. The Second V called "Velocity" describes the speed at which the data processed. In handling huge amount of data, generally fast processing is required on that data. The 3<sup>rd</sup> V is called as "Variety"

which is used to describe various types of data i.e. Structured, Unstructured and semi-structured data. According to Accenture's survey done on different organizations, who used the term big data at least once, It was the 3<sup>rd</sup> V i.e. "Variety" that in actual force them to use the term big data; Because this was the main pitfall of traditional software system as those were unable to handle different kind of data. Data can be structured, semi-structured or unstructured data. The last V is called as "Veracity" used to describe how much accurate the results are? It means the usability of result to the user.

### 1) Types of Big Data

The Big data is of following types:

- a) **Structured Data:** The structured data is the data that have some defined length and format. Examples of structured data include numbers, dates and groups of words and numbers called strings. It is usually stored in the form of table having well defined rows and columns. Data is stored in the database which can be easily accessed with the help of queries in SQL.
- b) **Unstructured Data:** Unstructured data is lying everywhere in the universe. Its not at all in structured form as structured data. It is not having any specified length or format. Images , videos and large files i.e. PDFs could be the example of unstructured data. Data from different social media websites i.e. facebook, linkedin, yahoo are also the type of unstructured data.
- c) **Semi- Structured Data:** Semi-structured data is neither structured nor unstructured. It doesn't having any fixed schema but may be self defined having simple label/value pairs. Example includes XML, SWIFT, EDI etc.

## III. HADOOP

Hadoop is framework managed by Apache that is derived from both Mapreduce and Big Table.Hadoop is a scalable, open source, fault-tolerant virtual grid operating systems architecture for data storage and processing. Hadoop allows applications based on MapReduce to run on large clusters of commodity hardware. Hadoop is developed to parallelize data processing across computing nodes to speed computations and hide latency. With hadoop, the analytics can be done on the whole of the dataset rather than small part of data.

Two major components of Hadoop are: a massively scalable distributed file system that can support petabytes of data and a massively scalable MapReduce engine that computes results in batch. So, HADOOP= HDFS+ MAPREDUCE.

It helps:-

- Save the file with the help of Hadoop Distributed File System i.e. HDFS
- Perform analytics on top of it with the help of Google Map Reduce Implementation Algoirthm.

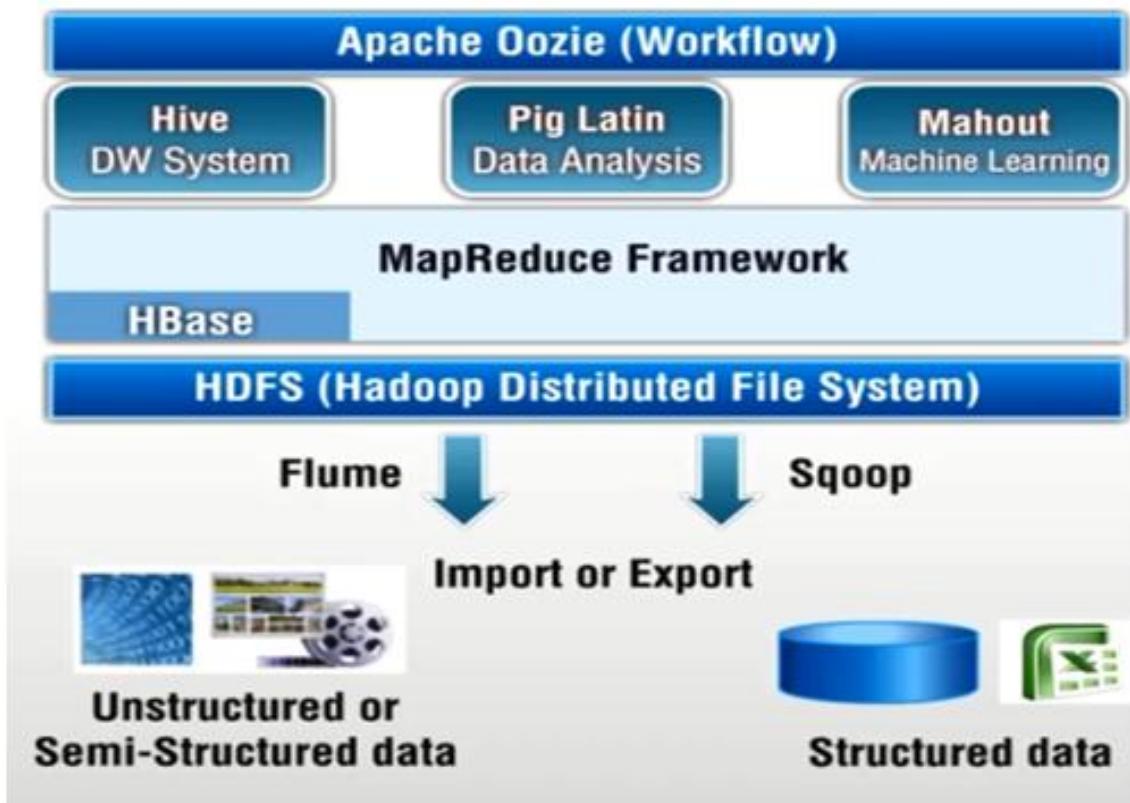


Fig 3: Hadoop Ecosystem [7]

- a) *Hive*: Hive is for querying and management of Large distributed datasets. Usually the implementation language used is Java. Hive Query Language is another Query language which is used for the structured data
- b) *Pig*: Pig was initially developed by Yahoo research around 2006 to make hadoop more approachable and understandable it for non-developers. Pig is script based, interactive supporting pig Latin which is used to express data flows. It works in two modes: Local Mode and Hadoop.
- c) *Mahout*: Mahout deals with Machine learning i.e. Artificial Intelligence.
- d) *Apache Oozie*: It is workflow scheduler system that helps handle hadoop jobs.
- e) *Flume*: Flume is fault tolerant, distributed service for moving large amounts of log data. It is designed to deal with unstructured or semi-structured data. It may be pushing the data to the commodity machines of the distributed file system or having that data back.
  - a) *Sqoop*: It is important to load and unload data in bulk between relational database servers and hadoop. So, it is designed to deal with structured data. It deals with exporting the structured data to the commodity machines of Relational database servers or importing the same data back.

#### IV. MapReduce and Hadoop Architecture

The Architecture of MapReduce and HDFS consists of Master Node and the Slave nodes. Master node is responsible for handling the functioning of the slave nodes. Slave nodes could be more than one; on the contrary, there could be only one master node for all slave nodes to maintain them.

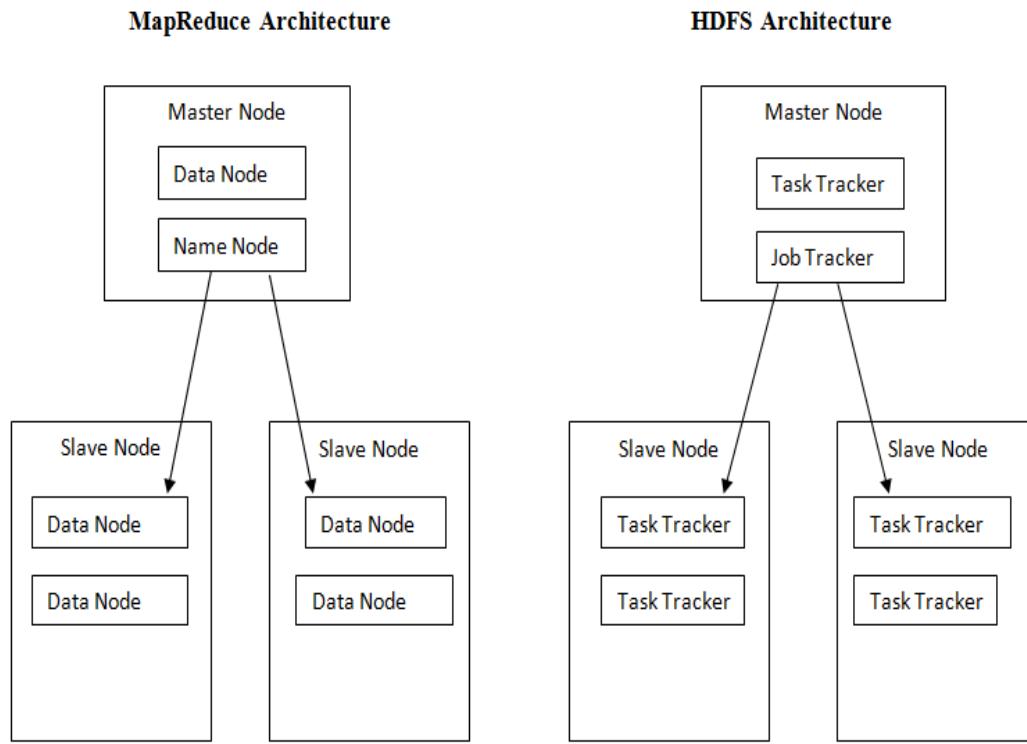


Fig 3: Map Reduce and HDFS Architecture

- a) *Name Node:* HDFS works by breaking large files into smaller pieces called blocks. The blocks are stored on data nodes, and it is the responsibility of the NameNode to know what blocks on which datanodes make up the complete file. The NameNode also acts as a “traffic cop,” managing all access to the files, including reads, writes, creates, deletes, and replication of data blocks on the data nodes. Data nodes are not very smart but the NameNode is. Data nodes kept on asking NameNode whether there is any function for them to do. The data Nodes also communicate with themselves.
- b) *Data Node:* Data nodes are not smart but they are resilient. Within the HDFS cluster, data blocks are replicated across multiple data nodes and access is managed by the NameNode. The replication mechanism is designed for optimal efficiency when all the nodes of the cluster are collected into a rack. In fact, the NameNode uses the “rack Id” in order to maintain the records of the Data Nodes.
- c) *Task Tracker:* TaskTracker runs on all DataNodes. TaskTracker get the Mapper and Reducer tasks for execution by the TaskTrackers with the help of library available. TaskTracker remained in consistent communication with the JobTracker to update him for the progress.

- d) **Job Tracker:** JobTracker acts as a master which creates and runs all the jobs. It is the duty of the JobTracker to allocates different jobs to the slaves i.e. TaskTrackers. When any TaskTracker become unresponsive, it becomes the duty of the TaskTracker to assign the same task to the other Node.

Hadoop solves the problems of existing system as:-

- It helps save the entire files
- Map reduce framework does analytics on top of the entire set of data sets
- Takes very less time to do the analytics
- Scale out architecture .
- toggle with the amount of time it takes or processing

## V MapReduce

MapReduce is developed by Google in order to execute set of functions in batch mode on very large amount of data. It supports parallel and distributed computing. The MapReduce consists of two functions i.e. “Map” and “Reduce”. The “Map” Components actually distribute the programming problems or task across a large number of systems and handle the placement in a way that manages the load and handles recovery from the failures. After distributed computation is completed, another function “reduce” aggregates all the elements back together to generate the final result. An example of MapReduce may be to document that how many workers are working in the different departments of Punjabi University.

It's a programming model for processing large scale datasets in computer clusters.

a) *Map Reduce Components:*

1. Name Node: it manages HDFS metadata, doesn't deals with files directly at all.
2. Data Node: it contains blocks of HDFS and are replicated
3. Job Tracker-schedules, allocates and monitors job execution on slaves - Task trackers
4. Task Tracker- runs Map Reduce operations.

b) *MapReduce Techniques:*

1. Prepare the Map() input
2. Run the user-provided Map() code
3. “Shuffle” Map output to the Reduce processors
4. Run the user-provided Reduce() code
5. Produce the final output

## VI. STATISTICS

### Statistics

- On a very common social networking site i.e. *facebook*, the like button is pressed around 2.7 billion times a day from the internet social lovers all around the globe presenting the love for these sites by them.

- It is found that normally a person spends around 2.5 hours a day on like, comment, poke, tweet and doing similar operations on social media.

In fact, social media have become the number 1 source of advertisement by the leading organizations that further embodies a bulk of data by publishing hundreds and thousands of advertisements everyday. According to McKinsey, Big Data refers to datasets whose size are beyond the ability of typical databases software tools to record, store, manage and analyze. According to Gartner Group, the “Big” in Big Data also refers to several other characteristics of big data source which includes increased velocity and increased variety also which leads to extra complexity as well.

Manyika have predicted that Only America will face shortage number of people with analytical skills and shortfall of millions data savvy manager with knowledge of dealing with big data in order to make effective decisions.

With the reason of smart phone users surfing internet and social media websites becomes the main reason for increasing the data making the big data an important point to be considered. Let's have a look at some interesting facts:

- Facebook is having 600 million active users which spend more than 9.3 billions hours a month.
- You tube is having 490 million visitors worldwide who spend approximately 2.9 billion hours each month and it upload 24 hours of video every minute, forming the bulk of data on the internet.

## CONCLUSION

With this fast technology friendly environment, data is constantly generating at an extreme velocity. It is becoming so difficult for traditional system to handle such kind of data due to lack of scalability facing challenges. Moreover, the data so generated can be in any form like it can be structured, unstructured or semi-structured. The result of what came is generation of big data which works with four characteristics as Volume, velocity, variety, veracity that makes easy to handle this extreme data overcoming all the pitfalls of traditional system. Big data concept make the production of platform that further helps in handling the unstructured or semi-structured data.

## REFERENCES

1. Statistics of Facebook concerning Big Data. Available: <http://www.wired.com/insights/2014/03/facebook-decade-big-data/>
2. Facebook and Big Data Co-relation. Available:<https://www.linkedin.com/pulse/20140716060957-64875646-facebook-and-big-data-no-big-brother>
3. “*Big Data Analysis using Apache Hadoop*” by Shankar Ganesh Manikandan and Siddarth Ravi, dept. Of Information Technology, Dhanalakshmi college of Engineering Tambaram, Chennai, India
4. “*Big Data for Dummies*” by Judith Hurwitz, Alan Nugent, Dr. Fern Halper, Marcia Kaufman
5. A “*Survey of Big Data in Social media using data mining techniques*” by Sheela Gole and Bharat Tidke, Dept. Of Computer Engg., Flora Institute of Technology, Pune, Maharashtra, India
6. A “*Big Data Modeling Methodology for Apache Cassandra*” by Shiyong Lu, Wayne State university
7. An online “*Big Data and hadoop concepts*” video by Syed Rizvi

# OS SPECIFIC HOST BASED DoS ATTACKS

Er. Harleen Kaur\*

\*Student, Department of Computer Engineering

Punjabi University, Patiala

(kharleen644@gmail.com)

Er. Harmandeep Singh

Assistant Professor, Department of Computer Engineering

Punjabi University, Patiala

(harmanjhajj@yahoo.co.in)

**Abstract-** The most hardest and common problem in today's Internet is the Denial of Service (DoS) attack. These attacks deprived off the resources and the services of the user connected over Internet within a short period of time and with very ease. The effect of the attack is depend upon the attacker's will, so it is easy for an attacker to organize an attack with destructive results. These attacks have severe impact on the operating system (OS) of a system. So the operating system must protect itself from the security gaps. In client-server communication, a DoS attack may be lanced by sending many useless requests to the server. These useless requests results in depriving off server's resources and stopping it from responding legitimate client requests. In this paper, the approach used to prevent these attacks and some OS specific host based attacks are discussed.

## I. INTRODUCTION

Due to an increasingly vulnerable to unauthorized attempt in today's security changing environment, the protection of a legitimate user or an organization's security system has become a major concern. With the global accessibility to the Internet, the internet can become a threat to an organization or user's security system because it is easy to compromise its security controls by internal as well as external attacks. Denial of service (DoS) attack is one of the security threat amongst various threats and is most destructive according to the security experts. Hence, the purpose of this section is to strengthen the awareness of DoS by examining the effectiveness and efficiency of the security system. DoS attack is the most dangerous attack that targets the system's availability of data or resource to prevent normal access whenever it is desired [1]. DoS attacks are responsible for the rapid increase in the foreshadowing for the internet's productivity and the profitability [2]. The services such as e-mail, web or internet connectivity that a legitimate user normally expect to have, are deprived off by straightforward events that are characterized as attacks. The explicit events include - event to flood a network (preventing network traffic), event to prevent particular user from accessing services, event to interrupt particular service to a specific user or system. DoS attack stops the legitimate user from accessing various services by interrupting the internet by consuming network's bandwidth. The definition of "DoS" by Computer Emergency Readiness Team (CERT) is as follows: [3]

- Precise use of bounded resources such as network bandwidth, memory of the system, data structure etc.
- Damage network data for example shutdown web services.

An attack to make an access or to damage system is not the aim of DoS but to block them so badly that they cannot be used for any useful purpose is the main aim of the DoS. Thus, system must protect its operating system from security gaps like stack overflow, memory-access violations, launching of programs with excessive privileges and many more. This paper presents the different categories of DoS attacks and the defense approaches against these attacks.

## II. HISTORY

With the rapid increase in the dependency of the Internet over the past few years, the weakness as well as attacks over the Internet also increases. The first major DoS attack marked its presence in August 1999 which brought down the University of Minnesota's network for three days [4]. After this, the era of DoS begun. Approximately six months later, the attack by a Canadian teenager on February, 2000 on several major websites made headlines. The high profile distributed DoS attack on February 7 hit the most popular website on the Web - "Yahoo". According to Media Metrix, a company that specializes in online traffic measurement, Yahoo had more unique visitors in January than any other site. The attackers next day morning knocked out Buy.com, an online store celebrating its initial discount on available stock for 2 hours 30 minutes. Within next nine hours, the attacks knocked out Amazon.com (10th highest number of unique visitors), eBay.com (15th highest) and CNN.com (41st highest). ZDNet (19th highest), E\*TRADE and Excite (11th highest) also became victim on the following day [5]. These attacks are increasing in numbers because of lack of thorough defending techniques. In 2000 over a period of 3 weeks, the frequency of the DoS attack was found to be 12,000 attacks. DoS attacks after the virus infections are the second most widely detected obvious attack types in computer networks. According to CERT, the number of reported Internet security incidents has jumped from six in 1988 to 82,094 in 2002, and to 137,529 in 2003 [6]. Due to the excessive increase in the number of security incidents, CERT has decided not to publish the number of incidents reported since 2004. Thus, DoS attack is a most destructive method of blocking service from its intended users.

Year	Percentage of Observed DoS Attack
1999	30
2000	27
2001	36
2002	40
2003	42
2004	39
2005	32

Table-1: Percentage of CSI/FBI Cyber security Survey Responders who observed a DoS Attack during 1999–2005 [6]

## III. TECHNIQUES of DoS ATTACKS

The attacks that crash the service and that flood the service are the two general phases of DoS attacks [7]. It is impossible to specify all the existing techniques, therefore in this section we describe some DoS techniques. The following are the techniques into which different DoS attacks are classified:

Network based DoS Attacks

Host based DoS Attacks

Distributed DoS Attacks

*a) Network based DoS Attacks*

The attacks which flood the network and target the packets in order to decrease the bandwidth for the user are categorized under network based attacks. The following are the techniques under network based DoS: [8]

*a.1) TCP SYN flood*

The network protocols which absorb resources to maintain states are overworked by DoS attacks and TCP SYN flooding is one of them [9]. In this type of attack, the multiple requests are sent for connection establishment. The multiple requests are sent because the client does not respond to the SYN-ACK message send by the server at the middle stage (i.e. half open connection exists and TCP will wait forever for that packet). It causes high bandwidth & less disk space. The TCP SYN flood attack is one of the type of protocol exploit attacks [10].

*a.2) ICMP Smurf flood*

Smurf attacks are one of the most devastating DoS attacks. The working of ICMP (Internet Control Message Protocol) is generally used to determine whether a computer in the Internet is responding or not and to check that an ICMP echo request packet is sent to a computer. In a smurf attack, attacker forge ICMP echo requests having the victim's address as the source address and the broadcast address of the remote networks as the destination address.

*a.3) UDP flooding*

In UDP flooding, the attacker send a large amount of UDP (User Datagram Protocol) packets towards the computer connected to the Internet. As the victim network can handle only some extent of the higher traffic volume delivered by an intermediate network, so the flooding traffic can consume the victim's connection resources. With the UDP flooding, pure flooding can be done with any type of packets.

*a.4) Intermittent flood*

In such attacks, the flooding of packets by the attacking hosts in a burst to congest and disrupt existing connections. Since all the disrupted connections will wait for a specific period of time (known as retransmission time out [RTO]) to retransmit lost packets, the attackers further flood the next RTO to disrupt connections by sending more and more flood packets. Giurgiu's study showed that a burst of 800 requests can bring down a web server for 200 seconds, and thereby the average flooding rate could be as low as 4 requests per second [11].

*b) Host based DoS Attacks*

Host based DoS attacks aim at attacking computers. Either a vulnerability in the operating system, application software [8] or in the configuration of the host are targeted. Crashers are a form of host based DoS that are designed to crash the host system, the system needs to be restarted. Crasher introduces vulnerability in the host's operating system. It can be done by exploiting the implementation of network protocols by various operating systems.

*c) Distributed DoS Attacks*

It is generated by a large number of hosts which can be the amplifiers or reflectors that can be seeded on remote hosts and wait for the command to 'attack' a victim. It floods the victim with as many packets in order to overwhelm the victim.

#### IV. DEFENSE MECHANISM

As there are a large number of DoS attacks. So to make an effective defense measures, one must know the category of the attack in which it is classified. Also it is highly recommended to determine where the defense has to be deployed. There are many defense mechanisms that count on the ability to differentiate between attack and legitimate flows [12]. The following are the four categories of defense against DoS attacks: [6]

Attack prevention

Attack detection

Attack source identification

Attack reaction

*Attack Prevention*

The attack prevention aims to stop the DoS attack at the first place before they can reach the victim. This step assumes that the source address of the traffic of attack is spoofed since intruders need spoofed traffic to hide real source of the traffic of attack. The variety of packet filtering schemes are accorded by this step. These packet filters are used to make sure only non-spoofed (valid) can pass through. This greatly reduces the chance of occurrence of DoS attacks. The following are some attack prevention approaches: [13]

*Ingress filtering:* The traffic coming into the local network of a user is known as ingress filtering. It is an approach to set up a router such that the incoming packets with illegal source address is not allowed to enter in network.

*Egress filtering:* The traffic leaving the local network of a user is known as egress filtering and it makes sure that only allocated or assigned space of IP address leaves the network.

*Router based packet filtering:* The approach which is efficient to filtering out large part of spoofed IP addresses and is used to analyses the IP traceback is known as router based packet filtering. It prevents the attack packets from reaching their targets. It extends the ingress filtering to the core of the internet [13].

*Attack Detection*

The attack detection aims to detect or to defend the DoS attacks when they occur and is an important function to direct any further action. By detecting the attack, a host user or a network can protect themselves against being a victim of a DoS attack as well as being a source of network attack. The coverage of the DoS attack that is what actual part can be detected is the measure of performance of attack detection. The following are some attack detection approaches: [13]

*Anomaly based detection:* It depends on detecting the abnormal behavior with respect to the normal standards. The detection systems have been developed in order to check the pass out signs of DoS attacks.

*Misuse detection:* The well-defined patterns of known exploits can be identified and then the occurrence of such patterns can be looked up, the approach used here is known as misuse detection.

*Attack Source Identification*

An ideal approach to block the attack traffic after the detection of the attack as its source is known as attack source identification. But it is not easy to track the IP traffic as its source due to the two aspects of the IP protocol. The two aspects are – first, the content with which IP source addresses can be duplicated and the second where the routers know the next hop for forwarding a packet that is stateless IP routing.

*Attack Reaction*

The approach which minimizes the loss caused by DoS attacks is known as attack reaction and is used when an attack is underway. The techniques discussed above that detects and traceback aims to shorten the time needed to locate the attack sources and to detect the attack.

## V. OS SPECIFIC HOST BASED ATTACKS

The denial of service attacks such as runaway processes, memory-access violations, the launching of program with excessive privileges etc., these are some attacks from which the OS must protect itself [14]. The tangible security policies, its measures and architectures vary between operating system [15]. The attacker not only identifies the active host's IP address and its respective port numbers but also the active host's operating system(OS), then attacker can launch the attacks that produce one or two packet "kills" [16]. The attacks which disable a system with minimal effort are comes under the category of OS specific DoS attacks. These attacks take the advantage of the way the protocols implemented on the operating system. The following are some OS specific DoS attacks:

### 1. Ping of Death

The maximum size of the IP packet which is allowable is 65,535 bytes (approximately 20 bytes). An ICMP echo request is an 8 byte IP packet including pseudo header. Therefore, 65,507 bytes (subtracting 8 and 20 from above total) is the maximum allowable data size. A packet size larger than 65,507 bytes included by a user is allowed by ping attacks. An oversized ICMP packet can start an event which has adverse system reactions such as crashing and rebooting. An attempt to penetrate a network by sending a ICMP echo requests in a continuous sequence over a high-bandwidth connection to a lower-bandwidth connection of a target host.

IP header (20 bytes)	ICMP header (8 bytes)	ICMP data (65.510 bytes)	Unfragmented Packet (original)
-------------------------	--------------------------	-----------------------------	-----------------------------------

Figure 1: Oversized ICMP data

The following example shows how to ping a particular victim system:-

The two operating systems are used for the implementation of ping of death:

System 1: Kali Linux (attacker's say) Version 3.4.2 – OS type 32 bit

System 2: Microsoft Windows XP Professional (victim's say) Version 5.1.2600 Service Pack 3 Build 2600

*Screenshots:*

*Step 1:* hping3 192.168.75.130, attacker ping the victim system by this command, where hping is a command-line oriented TCP/IP packet assembler/analyzer and 192.168.75.130 is victim's IP address.

The above command ping all the web services of the victim. In order to stop this ping attack press Ctrl+C.

*Step 2:* ping yahoo.com -t -l 65500, where -t = It pings the specified host until stopped.. -l = It send buffer size and 65500 is an ICMP data.

As the user on windows XP is a victim now, so whenever the victim tries to ping the services from any website etc., he will not get access to get that services. The victim is deprived off from the internet services.

The following are the screenshots of the respective system of attacker's and the victim's:

Attacker's:



The screenshot shows a Kali Linux desktop environment. The top bar includes icons for Applications, Places, system status, date/time (Thu Aug 18, 11:07 AM), and user (root). A terminal window is open, displaying the command hping3 192.168.75.130 and its output.

```
root@kali:~# hping3 192.168.75.130
HPING 192.168.75.130 (eth0 192.168.75.130): NO FLAGS are set, 40 headers + 0 data bytes
```

Victim's:

## 2. Teardrop

The teardrop attacks exploits the reassembly of disintegrated IP packets. One of the fields of IP packet is the fragment offset field, which tells the offset or the starting position, of the data present in a fragmented packet respective to the data of the original fragmented packet. Teardrop attack is an approach that exploit system's design weaknesses [17].

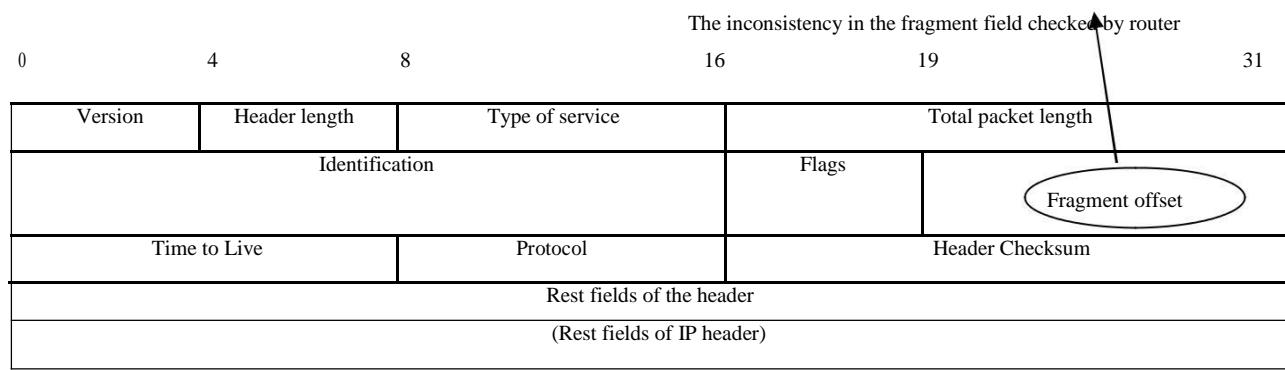


Figure 2: IP header

### 3. WinNuke

The DoS attack which targets any computer on the Internet running windows is known as WinNuke. The attacker sends a TCP segment generally to NetBIOS port 139 the urgent (URG) flag set to well established connection. This causes a NetBIOS fragment to overlap, which results in crash of running Windows [18].

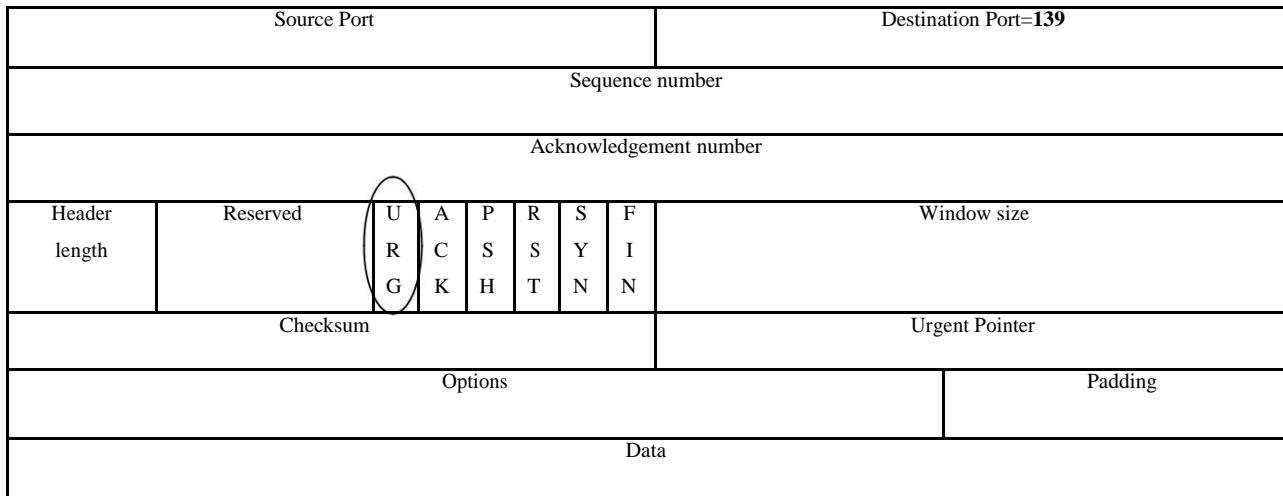


Figure 3: TCP header

## VI. CONCLUSION

The users are given more control over computer resources with every new version of operating systems. With the global accessibility of the Internet, the users over internet also increases resulting in increase in the bandwidth of user's system. Furthermore, the level of security knowledge in users connected over internet is decreasing day by day. As a result of this, the attacks are becoming more easy and popular. The attackers exploit gaps in systems to disallow access of target services. Thus, the power of attackers are growing rapidly. In this paper, we have discussed the classification of DoS attacks and the existing defense mechanisms in the Internet. The more reliable internet's network infrastructure is needed which authenticates the source traffic in order to fight against these attacks. The power of attackers automatically decreases with an increase in more secure computer systems over the internet.

## References

- [1] J. Atoum and O. Faisal, "Distributed Black Box and Graveyards Defense Strategies against Distributed Denial of Services," in *Computer Engineering and Applications (ICCEA), 2010 Second International Conference*, 2010.
- [2] D. A. ASOSHEH and N. RAMEZANI, "A Comprehensive Taxonomy of DDoS Attacks and Defense Mechanism Applying in a Smart Classification," *WSEAS TRANSACTIONS on COMPUTERS*, vol. 7, no. 4, pp. 281-290, 2008.
- [3] B. B. Gupta, M. Misra and R. C. Joshi, "An ISP level solution to combat DDoS attacking using combined statistical based approach," *arXiv preprint arXiv*, p. 1203.2400, 2012.
- [4] D. Karig and R. Lee, "Remote denial of service attacks and countermeasures," Princeton University Department of Electrical Engineering Technical Report CE-L2001-002, 2001.

- [5] L. Garber, "Denial-of-service Attacks rip the Internet," *IEEE Computer*, vol. 4, no. 33, pp. 12-17, 2000.
- [6] T. PENG, C. LECKIE and K. RAMAMOHANARAO, "Survey of Network-Based Defense Mechanisms Countering the DoS and DDoS Problems," *ACM Computing Surveys*, vol. 39, April 2007.
- [7] S. Rao and S. Rao, "Denial of Service attacks and mitigation techniques: Real time implementation with detailed analysis," The SANS Institute, 2011.
- [8] Q. Gu and P. Liu, "Denial of service attacks," in *Handbook of Computer Networks: Distributed Networks, Network Planning, Control, Management, and New Trends and Applications*, 2007.
- [9] W. Jian, "A possible LAST\_ACK DoS and fix," 8 April 2000.  
[Online]. Available:  
<http://lkml.iu.edu/hypermail/linux/kernel/0004.1/0105.html>. [Accessed 8 2016].
- [10] S. M. Specht and R. L. Lee, "Distributed Denial of Service: Taxonomies of Attacks, Tools and Countermeasures," in *ISCA 17th International Conference on Parallel and Distributed Computing Systems*, The Canterbury Hotel, San Francisco, California, USA, 2004.
- [11] M. Guirguis, A. Bestavros, I. Matta and Y. Zhang, "Reduction of Quality (RoQ) Attacks on Internet End-Systems," in *INFOCOM*, New York, 2005.
- [12] M. Geva, . A. Herzberg and Y. Gev, "Bandwidth Distributed Denial of Service: Attacks and Defenses," in *IEEE Security and Privacy*, 2014.
- [13] C. Douligeris and A. Mitrokotsa, "DDOS attacks and defense mechanisms: classification and state-of-the-art," *Science direct*, vol. 44, pp. 643-666, 2004.
- [14] "Security," [Online]. Available:  
[https://www.cs.uic.edu/~jbell/CourseNotes/OperatingSystems/15\\_Security.html](https://www.cs.uic.edu/~jbell/CourseNotes/OperatingSystems/15_Security.html). [Accessed 28 7 2016].
- [15] J. Liedtke, N. Islam and T. Jaeger, "Preventing denial-of-service attacks on a  $\mu$ -kernel for WebOSes," in *Operating Systems, 1997., The Sixth Workshop on Hot Topics in*, Cape Cod, MA, 1997.
- [16] "Juniper Networks," [Online]. Available:  
[http://www.juniper.net/documentation/en\\_US/junos12.1x46/topics/concept/denial-of-service-os-attack-overview.html](http://www.juniper.net/documentation/en_US/junos12.1x46/topics/concept/denial-of-service-os-attack-overview.html). [Accessed 8 2016].
- [17] R. K. C. Chang, "Defending against Flooding-Based Distributed Denial-of-Service Attacks: A Tutorial," *IEEE Communication Magazine*, vol. 40, no. 10, pp. 42-51, October 2002.
- [18] "Juniper Networks," [Online]. Available:  
[http://www.juniper.net/documentation/en\\_US/junos12.1x46/topics/concept/denial-of-service-os-winnuke-attack-understanding.html](http://www.juniper.net/documentation/en_US/junos12.1x46/topics/concept/denial-of-service-os-winnuke-attack-understanding.html). [Accessed 8 2016].

# Detection of phishing websites using SVM technique based on Data Mining

1Leena, 2 Er.Amrit Kaur

1 pursuing M.Tech, Computer Scholar

2Assistant Professor, Department of Computer Engineering, Punjabi university, Patiala, Punjab

Email-id: [leena.lalwani13@gmail.com](mailto:leena.lalwani13@gmail.com)<sup>1</sup>, [amrit.tiet@gmail.com](mailto:amrit.tiet@gmail.com)<sup>2</sup>

**Abstract:** Phishing is a major issue as to get stealing of confidential information. Various methods of phishing the confidential information are emails, phishing websites etc. In our work, we propose a phishing detecting tool to overcome the stealing of confidential information over phishing websites. For this purpose, features of websites play a very important role to detect phishing websites. Websites features like IP address, URL features, Length of URL, foreign anchors, null anchors and certificates of the websites are critically analyzed because to know the possibility of website to be phishing website. To get results for experiments of our tool, data set of phish-tank is used, which is a website to store the information of phishing and harmful websites. In this work, SVM classifier is used to collect the features and classify them. This tool is classifying 97% URL correctly and is a very effective tool to recognize the phishing websites.

**Keywords:** Phishing Detection, Websites, legitimate, feature extraction.

## I. INTRODUCTION

### A. *Introduction to Data Mining*

Data mining is an iterative procedure inside which advancement is characterized by revelation, through either programmed or manual strategies. Data mining is most helpful in an exploratory investigation situation in which there are no foreordained thoughts about what will constitute an "interesting" result. Data mining is the quest for new, significant, and nontrivial data in extensive volumes of data. Data mining is a procedure of finding different models, synopses, and inferred values from a given gathering of information. Practical the two essential objectives of data mining have a tendency to be prediction and description. Prediction involves utilizing a few variables or fields as a part of the data set to foresee obscure or future estimations of different variables of interest. Portrayal, then again, concentrates on discovering designs portraying the information that can be deciphered by people.

Data mining techniques:

- i). **Prescient data mining**, which delivers the model of the framework depicted by the given information set, or
- ii). **Descriptive data mining**, which delivers new, nontrivial information based on the accessible information set.

On the prescient end of the range, the objective of data mining is to create a model, communicated as an executable code, which can be utilized to perform arrangement, expectation, estimation, or other comparative assignments. On the descriptive end of the range, the objective is to pick up a comprehension of the dissected framework by revealing

examples and connections in substantial information sets. The relative significance of expectation and depiction for specific information mining applications can fluctuate significantly.

#### *B. Introduction to Data Mining*

Phishing remains a significant issue that undermines the security and protection of online clients regardless of endeavors to avert such assaults subsequent to 1995 [1]. Phishing assaults ordinarily include an assailant taking on the appearance of an honest to goodness online element to take private data from the clueless casualties. The adequacy of phishing relies on the capacity of the foe to befuddle the casualty. Phishing sites regularly show content comparable or indistinguishable to their honest to goodness partners and again produce the format and 'look-and-feel' of these websites. These assaults frequently begin with deceitful messages sent to a gathering of online clients to draw them to the fake sites. In fruitful assaults casualties click on the email hyperlinks which connection to the phishing sites, and after that casualties reveal their private qualifications, (for example, passwords or Master card data) or download malware.

Alongside an expansion in the quantity of potential targets, three central point [4] including unawareness of danger, unawareness of strategy, and criminal's specialized modernity have been used by culprits to exploit. They not just utilize fake email messages and sites to draw clients into disclosing their own data, additionally progressively utilize vindictive codes that particularly target client account data. Moreover, phishers today have a substantial fishing supply container of instruments accessible to them. Some phishing assaults for the most part utilize compound traps and the unawareness of the assault that turn out to be increasingly troublesome for clients [4].

Despite the fact that phishers are currently utilizing a few procedures in making phishing sites to trick and appeal clients, they all utilize an arrangement of common components to make phishing sites on the grounds that, without those elements they lose the upside of trickiness. This helps us to separate amongst genuine and phishing sites in light of the elements extricated from the visited sites.

Generally, two methodologies are used in identifying phishing sites. The first one depends on blacklist [2], in which the asked for URL is contrasted and those in that rundown. The drawback of this methodology is that the blacklist for the most part can't cover all phishing sites; subsequent to, inside seconds, another false site is relied upon to be dispatched. The second approach is known as heuristic-based strategy, where a few components are gathered from the site to group it as either phishy or genuine. As opposed to the blacklist technique, a heuristic-based arrangement can perceive naturally made phishing sites. The precision of the heuristic-construct technique depends with respect to picking an arrangement of discriminative elements that may help in recognizing the site class [3]. The route in which the elements are prepared likewise assumes a broad part in ordering sites precisely.

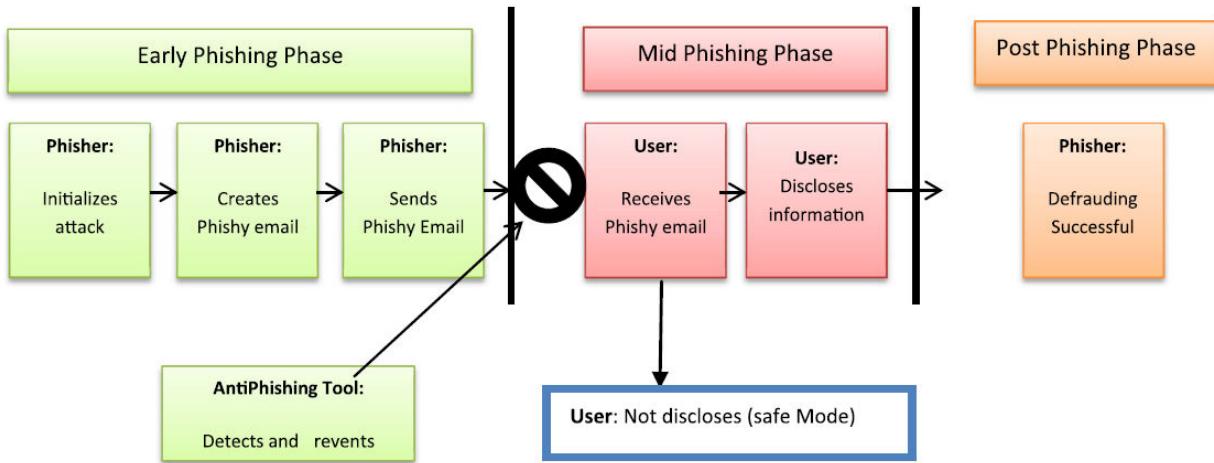


Fig: 1: Phishing Life Cycle. [5]

## II . Literature survey

Aydin et al. [6] propose a more secure structure for identifying phishing sites with high precision in less time. The recognition of phishing can be accomplished by either expanding client mindfulness or utilizing programming based discovery. In this study, they separated more URL elements and broke down subset based component determination techniques which have not been utilized already with the end goal of phishing sites identification in view of URL.

Priyan et al. [7] propose a TSO automatically answer for Desktop phishing assault. It has a capacity to discovery and counteractive action of the assault. This is likewise able to keep the assault by programmers to change the passwords of the framework and don't permit to them to oppose clients to change their PC settings.

R. et al. [8] talk about essential qualities that distinguish illegitimate sites from unique destinations are examined and an execution of C4.5 calculation is utilized for characterizing illegitimate sites furthermore plans to enhance the execution by consolidating with boosting calculations.

Li et al. [9] propose a novel methodology in view of least encasing ball support vector machine (SVM) to phishing Website location. They first play out an investigation of the topology structure of site as indicated by the DOM tree and utilize the Web crawler to remove topological elements of the site. At that point, the component vectors are distinguished by SVM classifier. Contrasted and the general SVM, this strategy has moderately high exactness of distinguishing, and supplements the dis-favorable position of moderate rate of meeting on substantial scale information

Gautham et al. [10] examined the attributes of legitimate and phishing site pages inside and out, and in view of this investigation; they proposed heuristics to concentrate 15 parameters from such site pages. Before applying heuristics to the website pages, they utilized two preparatory screening modules as a part of this framework, the initially approve website identifier to check site pages against a private white-list kept up by the client, and the Login Form

Finder to order pages as original when there are no login frames present. These modules lessen pointless calculation in the framework and moreover diminishing the rate of false positives without trading off on the false negatives.

Hamid et al. [11] propose a methodology for email conceived phishing location in view of profiling and bunching strategies. They figure the profiling issue as a grouping issue utilizing different elements present as a part of the phishing messages as highlight vectors and produce profiles taking into account bunching expectations. These forecasts are further used to create complete profiles of the email messages. They analyzed the execution of the proposed approach against the Modified Global K implies (MGK implies) approach. The outcomes demonstrate that the proposed methodology is proficient when contrasted with the benchmark approach.

### III. System module

This section explains the proposed phishing detection approach in which features of URL of the WebPages are fed and after some pre-processing, feature extraction has been carried out for legitimate as well as phished data. The main work in this is to explore machine learning algorithm that use a wide range of features .The computer simulation results demonstrate that the classifiers can get data from the URLs, page links of webpages. The content parameters of web pages contributed the most to lessen the fault rates, and then the other significant ones were the URLs content. The system has been tested for more than 25000.

URLs and proved varying efficiency depending upon number and type of features used for final classification. This allowed to detect whether a webpage is malicious or it is not in a very short interval of time. The light weight classifiers are very speedy but are slightly less correct, whereas heavyweight classifiers are more precise but takes more time. This extension enabled the SVM classifier to be ‘middleweight’ with very high accuracy and yet less prediction time along with its easiness in implementation. We have used a number of features basis on which phishing detection system has been developed in MATLAB software tool. The brief description of the steps followed in feature extraction phase and phases of classification are explained below.

#### **Step1) Data Collection and pre-processing**

In this work we have used phishing dataset from Phishtank websites (total 28887 URL's) which includes the latest phishing websites that are reported as to be phished and similarly legitimate websites (total 994) has been collected which are proved to be well known and most clicked websites. First we have collected the dataset in separate excel sheets which are read in matlab workspace and further processing has been passed on the import of URL's so as to acquire seven types of features about each URL. The features are briefed in second step.

#### **Step 2) Feature extraction**

**URL:** A URL (uniform resource locator) is used to locate the resources. However, in many cases, URL is often used as a synonym for URI. [14] The structure of URL is as follows:

< protocol > // < subdomain > . < primarydomain > . < TLD > / < pathdomain >

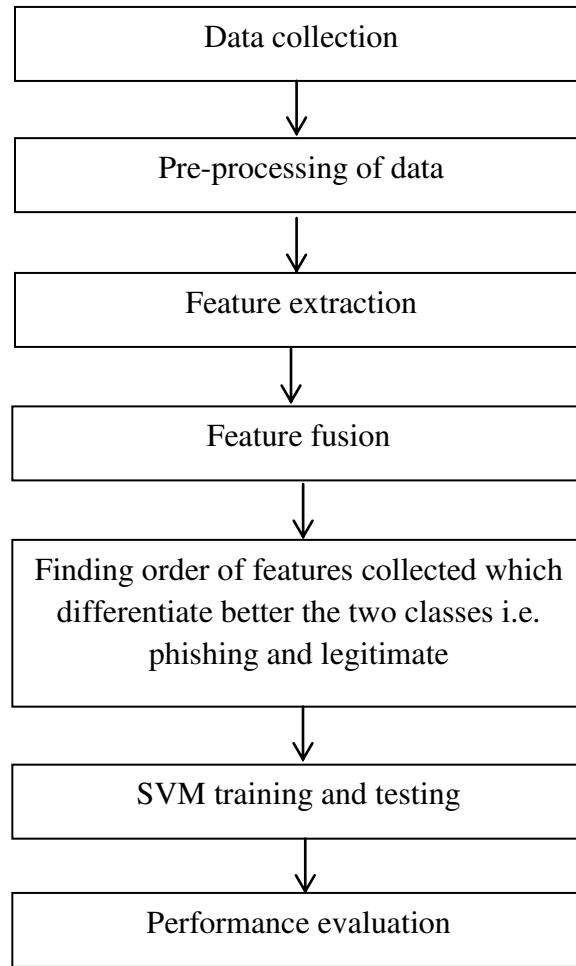


Fig. 2: Flowchart for the proposed scheme based on proposed work

## Step 2) Feature extraction

**URL:** A URL (uniform resource locator) is used to locate the resources. However, in many cases, URL is often used as a synonym for URI. [14] The structure of URL is as follows:

< protocol >: // < subdomain > . < primarydomain > . < TLD > / < pathdomain >

In our work we compare URL of the website with phishy website data set given by various phishing detecting websites.

**Length of URL:** Phishers hide the suspicious part of the URL to forward information that is provided by users or redirect the uploaded webpage to a doubtful domain. When the URL length is greater than 45 characters the URL can be considered phishy. [13]

Rule: If URL length < 45 → Legit

URL length P >45 and P<75 →Suspicious else →Phishy

**Similar websites domain names:** An attacker will generate a lot of URLs under a registered domain. The domain name contains: a main level domain (mld) and a public suffix (ps). Following figure is showing the similar phishing name as true website URL.

Type	Legitimate URL	Phishing URL
1	<a href="http://www.taobao.com/">http://www.taobao.com/</a>	<a href="http://taobao.689zx.cc/">http://taobao.689zx.cc/</a>
	<a href="http://www.icbc.com.cn/">http://www.icbc.com.cn/</a>	<a href="http://www.icbc.com-eis.tk">http://www.icbc.com-eis.tk</a>
2	<a href="http://store.apple.com/us">http://store.apple.com/us</a>	<a href="http://fimisland.us/wp-content/themes/twentytwelve/inc/iTunes/8e9722ea5ed5265fb47/">http://fimisland.us/wp-content/themes/twentytwelve/inc/iTunes/8e9722ea5ed5265fb47/</a>
3	<a href="http://voice.zjstv.com">http://voice.zjstv.com</a>	<a href="http://www.zvs15.cc/">http://www.zvs15.cc/</a>
4	<a href="https://itunesconnect.apple.com">https://itunesconnect.apple.com</a>	<a href="http://122.36.238.169/mnt/files/iTunes-Connect/06ebc5d3288bfcd9dee11abfeab8e8f/">http://122.36.238.169/mnt/files/iTunes-Connect/06ebc5d3288bfcd9dee11abfeab8e8f/</a>
5	<a href="http://runningman.cztv.com">http://runningman.cztv.com</a>	<a href="http://www.benpa0.com">http://www.benpa0.com</a> <a href="http://www.benp88.com">http://www.benp88.com</a> <a href="http://www.benpa8.com">http://www.benpa8.com</a>
6	<a href="https://login.alibaba.com/">https://login.alibaba.com/</a>	<a href="http://198.27.76.49/jd/login.jsp.htm">http://198.27.76.49/jd/login.jsp.htm</a> <a href="http://198.27.76.49/j/s/login.jsp.htm?amp=&amp;biz_type=&amp;crm_mtn=3597645053">http://198.27.76.49/j/s/login.jsp.htm?amp=&amp;biz_type=&amp;crm_mtn=3597645053</a>

Fig. 3: - The confusion phishing URL instances.

**Number of dots:** The dot in the URL represents the presence of the sub-domain in the URL. Some phishers may utilize the sub-domain to look the site address as the legit site hence causing the user to deceive to phish site. More the number of dots in URL more the sub-domain existing in URL to hide the web URL and look alike the legit site

**NULL Anchors:** These are the Anchor tags in the web page which are not pointing to anything. When clicked on such hyperlinks nothing happens or the links are redirected to the same existing page. After copying the source code of legit site the phisher may delete most of the links or replace with the link of the same page that mislead the customers. Hence, More the NULL anchors are in page more the page is likely to be a phishing site.

**URL of anchor:** An anchor is an element that can be defined as the `<a>` tag. This parameter is treated exactly as “Request URL” but for this parameter of URL, the links within the web page might refer to a domain distinct from the domain typed on the URL address bar. This parameter is a ternary parameter and treated exactly as “Request URL”.

**Fake HTTPs protocol/SSL Final:** The presence of HTTPs protocol every time sensitive information is being transferred reflects that the customer certainly connected with a legitimate website. However, phishers may use a fake HTTPs protocol so that the customers may be deceived, so checking that the HTTPs protocol is offered by a trusted issuer.

**Rule: use of https & trusted issuer & age ≥ 2 years → Legit**

**Using https & issuer is not trusted → Suspicious else → Phishy [15]**

**Step 3) Training and testing using SVM classifier**

### Support Vector Machines (SVM)

The foundations of Support Vector Machines (SVM) have been introduced by Vapnik et al [16] for binary classification. The simplest form, given data points represented as p-dimensional vectors, the SVM classifier tries to

find a hyperplane which separates these points into two-class data with maximal margin (maximizes the distance between the margin and the nearest data point of each class). The margin is defined as the distance of the nearest training point to the separating hyperplane [17]. There are many hyperplanes that might separate the data. The hyperplane to chose is the one that shows the largest separation. Figure below shows two-class data which can be separated by many liner classifiers, but only one is considered that maximize the margins (the green line) and the linear classifier is known as a maximum margin classifier.

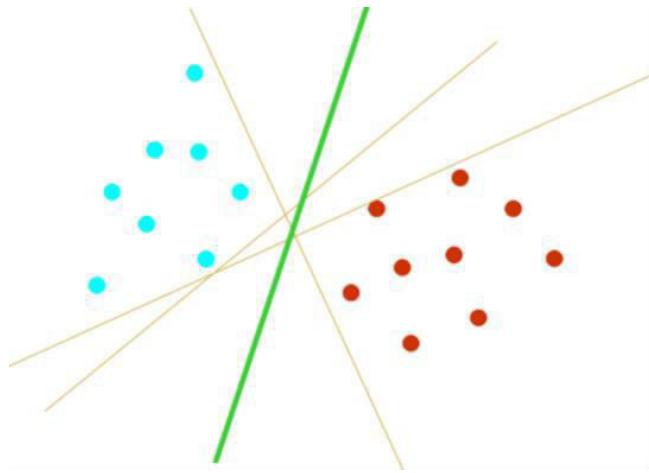


Figure: 4. optimal separating hyper plane [17]

#### IV. CONCLUSION

Phishing is a serious issue regarding the security and protection of personal information and data of individuals; phishers can steal confidential data to harm in financial and many other ways. For this purpose they use many ways like emails, phishing websites etc. To prevent them to attack the confidential data, we propose an anti-phishing websites tool which is used to detect the phishing websites and prevent them to steal our confidential information like usernames and passwords. In this work, we used various features of the websites, like IP address, URL features, Length of URL, number of dots in the URL, null anchor and foreign anchor, and certificates of the websites. All these features are very helpful to detect Phishing websites. For experimental results we gather phishing websites data from phish tank which is a data tank website to store the information about the phishing and harmful websites. Further classification has been done for collected feature set using SVM classifier. Experimental results show the effectiveness of proposed algorithm in which approx. 97% URL's are classified correctly.

#### REFERENCES

- [1] M. Langberg, "AOL acts to thwart hackers," Published in San Jose Mercury News, Sep. 1995.
- [2] Sanglerdsinlapachai, Rungsawang, "Using domain top-page similarity feature in machine learning-based web" Published in Third Int. Conf. Knowledge Discovery and Data Mining, 2010, pp. 187–190.

- [3] Sophie, Gustavo, Maryline, "Decisive heuristics to differentiate legitimate from phishing sites" Published in Proc. 2011 Conf. Network and Information Systems Security, 2011, pp. 1–9.
- [4] Jason Milletary, "Technical trends in Phishing attacks", [http://www.uscert.gov/sites/default/files/publications/phishing\\_trends0511.pdf](http://www.uscert.gov/sites/default/files/publications/phishing_trends0511.pdf), US-CERT.
- [5] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, "Phishing detection based Associative Classification data mining" Published in Expert Systems with Applications 41 (2014) 5948–5959
- [6] Mustafa AYDIN, Nazife BAYKAL, "Feature Extraction and Classification Phishing Websites Based on URL" Published in Communications and Network Security (CNS), 2015 IEEE Conference on Date of Conference: 28-30 Sept. 2015 Page(s): 769 - 770
- [7] M.K. Priyan, C. Gokul Nath, E. Vishnu Balan, Prof. K.P. Rama Prabha, Prof. R. Jeyanthi, "Desktop Phishing Attack Detection and Elimination using TSO Program" Published in 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 6 - 8 May 2015. Pp 198-201
- [8] Rakesh R, Kannan A, Muthuraj kumar S, Pandiya raju V, Sai Ramesh L, "Enhancing the Precision of Phishing Classification Accuracy using Reduced Feature Set and Boosting Algorithm" Published in 2014 Sixth International Conference on Advanced Computing (ICoAC) Date of Conference: 17-19 Dec. 2014 Page(s): 86 - 90
- [9] Yuancheng Li, Liqun Yang, Jie Ding, "A minimum enclosing ball-based support vector machine approach for detection of phishing websites" Published in Optik 127 (2016) 345–351
- [10] R. Gowtham, Ilango Krishnamurthi, "A comprehensive and efficacious architecture for detecting phishing webpages" Published in computers & security 40 (-2014) 23e37
- [11] Isredza Rahmi A. Hamid, Jemal H. Abawajy, "An approach for profiling phishing activities" Published in computers & security 45 (2014) 27e41
- [12] Kang Leng Chiew, Ee Hung Chang, San Nah Sze, Wei King Tiong, "Utilisation of website logo for phishing detection" Published in computers & security 54 (2015) 16e26
- [13] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, "Phishing detection based Associative Classification data mining" Published in Expert Systems with Applications 41 (2014) 5948–5959
- [14] Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen and Minh Hoang Nguyen, "Detecting Phishing Web sites: A Heuristic URL-Based Approach" Published in 2013 International Conference on Advanced Technologies for Communications (ATC 2013) Date of Conference: 16-18 Oct. 2013 Page(s): 597 - 602
- [15] Noman Mazher, Imran Ashraf, Ayesha Altaf, "Which web browser work best for detecting phishing" Published in Information & Communication Technologies (ICICT), 2013 5th International Conference on Date of Conference: 14-15 Dec. 2013 Page(s): 1 - 5
- [16] VAPNIK, V., GOLOWICH, S. E. & SMOLA, A., 1996. Support vector method for function approximation, regression estimation, and signal processing. Advances in Neural Information Processing Systems, NIPS, 281-287
- [17] GUNN, S. R. (1998)," Support vector machines for classification and regression." ISIS technical report

# Research Gaps and Multidisciplinary Research Trends of SOA Quality Attributes

Aradhana Negi<sup>1</sup>, Manmohit Singh<sup>2</sup>, Parminder Kaur<sup>3</sup>

<sup>1, 3</sup>Department of Computer Science, <sup>2</sup>Department of Computer Engineering and Technology  
Guru Nanak Dev University, Amritsar, Punjab, India

<sup>1</sup>aradhana.csersh@gndu.ac.in, <sup>2</sup>manmohitsingh@live.com, <sup>3</sup>parminder.dcse@gndu.ac.in

**Abstract—** The base architectural style of this service era is Service-Oriented Architecture (SOA), which leads to the development of an environment where heterogeneous distributed systems can communicate with each other by providing their application functionality as a service with features like high interoperability, loose coupling, low-cost, agility. It also promotes the rapid development and integration of legacy systems. The growing importance of this architectural style encourages the quality attribute concept for quality based service development and service management. As this technology is platform independent so industrialist and academicians are putting various efforts to grow the more interesting multidisciplinary solution to get the benefit of integration and interoperability feature of SOA. The main objective of this paper is to summarize the Research Gaps and discuss Multidisciplinary trends in SOA Quality Attributes.

## I. INTRODUCTION

The most intricate task during application development is to transmute the requirement specification into the application architecture. For productive application architecture, non-functional requirements must be regarded along with functional requirements of the system. Service-Oriented Architecture (SOA) based on Service Orientation paradigm also considers the quality attributes as critical and inseparable aspect for its successful implementation. Service Orientation is a popular designing strategy in Software Engineering that exposes system capabilities to the end-users or to other applications as a service [1]. An essential unit of SOA is the concept of service where a service is a goal-oriented reusable component that represent a business or operational task [2] and discoverable via networks. These Services simplify the design of the mobile system and can be cloned as well as made moveable like software agents. In the past ten years, voluminous research work has been published regarding the implementation of SOA.

The literature directly or indirectly emphasis on quality attributes along with the functional requirement for its success. “Quality attributes requirements such as those for performance, security, modifiability, reliability, and usability have a significant influence on the software architecture of a system.” [3] and improvement of the supplied quality attributes can attract further users [4]. Analysts, developers, architects, providers and consumers have the main role in developing, implementing and in the deployment of services under service-oriented computing. The objective of this paper is to address the Challenging concepts of SOA Quality Attributes and its Research trends. With this stated aim, current paper is organised in four sections. Section II introduces SOA Quality attributes. Section III discusses the Challenging concepts of SOA Quality Attributes. Section IV presents the Key Research trends of SOA Quality attributes. Last section V will conclude this paper.

## II. SOA QUALITY ATTRIBUTES

Quality attribute is the lineament of software system that affects its quality and describes the non-functional requirements of the system. It deals with various aspects of a system like architectural aspect, technical aspect, business-related aspects etc and enhances the beauty of system. With the value-oriented output of quality attributes, these are also considered by SOA to ensure the full realization of this architectural design. SOA has been extended beyond its demarcation [5] whereas initially it was just an asynchronous document-based message exchange approach

that was chiefly meant for the alignment of business applications and IT infrastructure but now this concept has been used for the larger set of distributed system with more expectations for quality attributes [6].

SOA quality attribute illustrates that how SOA affirm non-functional requirements of the service-oriented system. In order to achieve organization goals and to meet the customer needs, these quality attributes must be satisfied by the architecture of the software. “Achieving quality attributes in service-based systems is critical [...] because service-based systems lack central control and authority, have limited end-to-end visibility of services, are subject to unpredictable usage scenarios and support dynamic system composition.” [7].

The sum of quality attributes is known as Quality-of-Service (QoS). QoS are measured by quality metrics. Quality attribute requirements may vary from one business to another and from one customer to another customer. QoS can be evaluated using various criterions such as performance, reliability, security, availability, accessibility, maintainability etc. Till now many quality models have been proposed for quality attribute classification. A systematic mapping of quality models has been done by [8], they consider 47 different quality models from 65 papers. Here authors are considering ISO 25010 standards for describing SOA quality attributes, where Functional suitability, Portability, Usability, Performance Efficiency, Compatibility, Maintainability, Reliability and Security are eight characteristics with 31 other sub-characteristics.

### III. RESEARCH GAPS

It is widely admitted that SOA is a mature concept but still a big ground is left to cover. Various challenges are associated with Service foundation, Service composition, Service design and development [9]. Further efforts are required to address these challenges and ensure that SOA fulfils the desired quality level while implementation. Broadly the challenges colligated with quality attributes can be seen as 1) Strengthen the attributes, where quality is positively affected 2) Reduce the negative impact, where quality is degraded by other attributes and factors. Priorities of these challenges must be decided by the implementing organization. Here challenges are described as per ISO 25010 model (Functional suitability, Portability attributes are intentionally left because they are indirectly covered by Usability).

#### A. *Performance Efficiency*

It describes the performance goodness of a service and can be evaluated using commonly known metrics like throughput, latency, complexity, execution time, bandwidth etc. With increasing complexity of services, the performance of the service system deteriorated [10] [11]. The biggest challenge with performance attribute is to minimise the negative effect of other attributes over it. O’Brien found an open issue about the requirement of a performance model that can run in a high complex runtime environment and performance parameter for that model [11].

#### B. *Compatibility*

This comprises two sub characteristics: Co-existence and Interoperability. Co-existence concerns with the degree to which a service system can be integrated with other service and perform well under shared resources policies whereas interoperability deals with technical and operational freedom from platform dependency and facilitate users to invoke and compose services seamlessly. Efficient Service composition and orchestration are understudied areas for these attributes. Co-existence and Interoperability of mission-critical applications (healthcare, stock trading, and

air traffic control) [12], Semantic modelling of QoS category and need of matching algorithm between desired and supplied QoS [13] , interconnection of heterogeneous physical devices (using RFID technology, WSN) [14] [15] are significant challenge.

#### C. Usability

It is the measurement of user's experience about services. This is the only attribute which directly rendered the customer's satisfaction and ultimately enhances the reputation of service and service provider. The hotspots of usability are Quality of Experience (QoE) [16] and Quality of Protection (QoP), development of QoS-aware middleware, QoS negotiation protocols, QoS monitoring, Mapping global to local SLAs, Automatic QoS controller algorithms, Self-Managing system for QoS (at various resource system), Workload forecasting [17].

#### D. Security

Security is always a prime concern for a software system. Security is largely important for SOA because of its decentralized and loosely coupled nature where machines interact with each other. National Security Agency (NSA) consider Injection Flaws, XML Denial of Service issues, Insecure Communication, Information Leakage, Reply Attack Flaws, Insufficient Authentication, Inadequate Testing, Insecure Configuration, Insufficient logging as SOA web Service security Vulnerabilities [18]. Issues like dynamic trust relationship between service composing partners, Balanced Security mechanism [19] can advance SOA implementation. Nine factors, which complicate SOA security has been described by [20].

#### E. Reliability

Reliability is the actualization of required functionality of a service understated condition. The most challenging area under reliability is Optimized Fault Tolerance strategies, Data provenance during dynamic composition, SLA (Service-Level Agreement), Service monitoring, Self-Configuration, Self-Healing [17], Adaptive metric measurement for service monitoring [21], Automated Governance and Multi-Organizational Implementations [6].

#### F. Maintainability

Modularity, Reusability, Modifiability and Testability are the prime sub-characteristics of this attribute. The biggest challenge with SOA maintainability attribute is to update a service in a highly complex runtime environment.

### IV. MULTIDISCIPLINARY RESEARCH TRENDS

Service-Oriented Architecture follows additive approach and it is extended from its existing boundaries to take benefits of QoS. The valuable features of SOA make it possible to use this design patterns for multidisciplinary research, keeping this idea in mind, this part key out the major multidisciplinary research trends in brief.

#### A. Employing SOA In Healthcare Solutions

Decision support system for diseases diagnosis is considered as one of the major application area of SOA. The healthcare solution emphasises on Clinical decision support [22], Neuron based molecular communication [23]. Healthcare ecosystems using SOA like HL7 based healthcare information exchange system. [24]. Security attribute is of prime concern in this case so further research is being done for healthcare information security.

#### B. Developing Environmental-Based Applications

Another growing interest of SOA researcher is towards environmental issues and solutions. Flood forecast from numerical weather prediction [25], Remote sensing product generation system for geostationary operational satellite [26] are new trends in recent years.

#### *C. Quality Evaluation and Management*

To ensure the satisfactory use of SOA, research is continuously being done on quality enhancement. Monitoring frameworks, Quality parameter evaluation using Fuzzy model and Agents [27], Zigbee [28], Hybrid diagnosis [29] and Bayesian diagnosis [30], Fault configuration [31], SOA design defects on quality attributes, QoS-aware workflow in different environments (e.g. virtualization environment, real-time environment) are the new trends in SOA quality management. .

#### *D. SOA-IoT*

IoT (Internet of Things) is the vast network of communicable sensory heterogeneous devices. “SOA is considered a good approach to achieve interoperability between heterogeneous devices in a multitude way.” [32], [33]. In SOA-IoT, services are provided through serviced Internet [34], and “it allows applications to use heterogeneous objects as compatible services.” [35]. Further research will focus on designing of an SOA for IoT with quality attributes. Some other interest move towards SOA for Wireless sensor network, Network improvement strategies, Things-as-a-service oriented architecture.

#### *E. Self-Adaptive SOA System*

As SOA deals with Machine-to-Machine interaction with services and it is very important in dynamic networking to recover network crash, temporarily link failure etc. Self-adaptive, Context-aware and Semantic-based nature of service system will be able to handle such circumstances in a network. Although it is hard to implement Self-adaptive SOA system along with quality attributes but still researches are going on this path using new SOA approaches like [36] [37]. Further efficient strategies are required to make SOA Self-adaptive.

#### *F. Using SOA for Business Automation*

SOA’s very basic idea of business and IT infrastructure alignment is now stretched to real-time Industrial automation [38], chain-based business alliance formation [39], Business process execution [40], Financial sectors like e-banking [41] and e-resource planning.

#### *G. Cloud Computing and SOA*

The core architecture supporting Cloud environment is SOA. One of the future directions for SOA is Cloud Computing [42]. Cloud computing using the concept of everything-as-a-service, where it considers quality oriented service computing [43] for better performance. SOA is being extended towards private cloud computing as a service [44] which will give a new direction to cloud computing.

## V. CONCLUSION

This paper presents the challenging issues and research trends related with SOA Quality attributes. The unique integrating approach of SOA has been presently used in many application areas. Along with the vast opportunities of

growth, SOA quality attributes are facing some serious challenges which if not considered important, will definitely not only reduce the quality of services but also the popularity of this design pattern.

#### ACKNOWLEDGMENT

The Authors would like to express their gratitude to the University Grant Commission (UGC), Govt. of India, to give financial support as Fellowship under UGC-NET for Junior Research Fellowship (JRF) scheme, to do this research effort. The Authors are also thankful to the Department of Computer Science, Guru Nanak Dev University (Amritsar, Punjab) for providing infrastructure and facility for research work.

#### REFERENCE

- [1] Stojanović and A. Dahanayake, *Service-oriented software system engineering*. Hershey, PA: Idea Group Pub., 2005, pp. 1- 26.
- [2] G. Lewis, "Getting Started with Service-Oriented Architecture (SOA) terminology," unpublished.
- [3] Software Eng. Institute|Carnegie Mellon University. Glossary [Online]. Available: <http://www.sei.cmu.edu/architecture/start/glossary/>.
- [4] L. Eboli and G. Mazzulla, "Service Quality Attributes Affecting Customer Satisfaction for Bus Transit," *Journal of Public Transportation*, vol. 10, no. 3, pp. 21-34, 2007.
- [5] G. A. Lewis, "Is SOA Being Stretched Beyond its Limits?," *Advances in Computer Science*, vol. 2, no. 1, pp. 17-23, 2013.
- [6] G. A. Lewis, D. B. Smith, and K. Kontogiannis, "A Research Agenda for Service-Oriented Architecture (SOA): Maintenance and Evolution of Service-Oriented Systems," Technical Note, Mar. 2010.
- [7] M. Galster and P. Avgeriou, "Qualitative Analysis of the Impact of SOA Patterns on Quality Attributes," in *Proc. 12th International Conf. on Quality Software*, Xi'an, Shaanxi , 2012, pp. 27-29.
- [8] M. O. Hilari, X. Franch, and J. Marco, "Quality models for web services: A Systematic Mapping," *Information and Software Technology*, vol. 56, no. 10, pp. 1167-1182, 2014.
- [9] M. P. Papazoglou, P. Traverso, S. Dustdar, and F. Leymann, "Service-Oriented Computing: State of the Art and Research Challenges," *IEEE Computer* , vol. 40, no. 11, pp. 38 - 45, Nov. 2007.
- [10] M. Swientek, U. Bleimann, and P.S. Dowland, "Service-Oriented Architecture: Performance issues and Approaches," in *Proc. 7th international Network Conf.*, Plymouth, UK, 2008, pp. 261-296.
- [11] L. O'Brien, P. Merson, and L. Bass, "Quality Attributes for Service-Oriented Architecture", In *Proc. International Workshop on Systems Development in SOA Environments*, Minneapolis, MN, 2007, pp. 3-10.
- [12] Q. Z. Sheng, and et. al, "Web services composition: A decade's overview," *Information Sciences*, vol. 280, pp. 218-238, Oct. 2014.
- [13] S. Ran, "A model for web services discovery with QoS," *ACM SIGecom Exchanges*, vol. 4, no. 1, pp. 1-10, Spring 2003.
- [14] J. Bronsted, K. M. Hansen, and M. Ingstrup, "Service composition issues in pervasive computing," *IEEE Pervasive Comput*, vol. 9, no. 1, pp. 62-70, Jan-Mar 2010.
- [15] S. Kalasapur, M. Kumar, and B. A. Shirazi, "Dynamic service composition in pervasive computing," *IEEE Transaction Parallel Distr. Syst.*, vol. 18, no. 7, pp. 907-918, Jul. 2007.
- [16] B. Upadhyaya, Y. Zou, I. Keivanloo, and J. Ng, "Quality of Experience: User's Perception about Web Services," *IEEE Transactions On Services Computing*, vol. 8, no. 3, pp. 410-421, May/Jun. 2015.
- [17] D. Menasce, "QoS Challenges and Directions for large Distributed Systems," unpublished.
- [18] N. S. Agency (NSA), [Online]. Available: [https://www.nsa.gov/ia/\\_files/factsheets/soa\\_security\\_vulnerabilities\\_web.pdf](https://www.nsa.gov/ia/_files/factsheets/soa_security_vulnerabilities_web.pdf).
- [19] N. M. Josuttis, *SOA in Practice* , USA: O'Reilly Media, 2007. pp. 45-193.
- [20] Simplicable Business Guide, [Online]. Available: <http://arch.simplicable.com/arch/new/9-soa-security-challenges>.
- [21] M. H. Hasan, J. Jaafar, and M. F. Hassan, "Monitoring web services' quality of service: a literature review," *Artificial Intelligence Review*, vol. 42, pp. 835-850, Oct. 2012.
- [22] S. R. Loya, K. Kawamoto, C. Chatwin, and V. Huser, "Service Oriented Architecture for Clinical Decision Support: A Systematic Review and Future Directions," *Journal of Medical Systems*, vol. 38, no. 12, pp. 1-22, Oct. 2014.
- [23] J. Suzuki, S. Balasubramaniam, S. Pautot, V. D. P. Meza, and Y. Koucheryavy, "A Service-Oriented Architecture for Body Area NanoNetworks with Neuron-based Molecular Communication," *Mobile Networks and Applications*, vol. 19, no. 6, pp. 707-717, Nov. 2014.
- [24] N. K. Janjua, M. Hussain, M. Afzal, and H. F. Ahmad "Digital health care ecosystem: SOA compliant HL7 based health care information interchange," in *Proc. International Conf. on Digital Ecosystems and Technologies*, Istanbul, 2009, pp. 329 – 334.
- [25] H. Shi and e. al, "A service-oriented architecture for ensemble flood forecast from numerical weather prediction," *Journal of Hydrology*, vol. 527, pp. 933–942, Aug. 2015.
- [26] S. Kalluri and e. al, "A High Performance Remote Sensing Product Generation System based on a Service-Oriented Architecture for the Next Generation of Geostationary Operational Environmental Satellites," *Remote Sensing*, vol. 7, no. 8, pp. 10385-10399, Aug. 2015.
- [27] M. Thirumaran, P. Dhavachelven, S. Abarna, and G. Aranganayagi, "Architecture for Evaluating Web Service QoS Parameters using Agents," *International Journal of Computer Application*, vol. 10, no. 4, pp. 15-21, Nov. 2010.
- [28] P. C. Tseng, C. Y. Chen, W. S. Hwang, J. S. Pan, and B. Y. Liao, "QOS-Aware Residential Gateway supporting ZIGBEE-related services based on a Service-Oriented Architecture," *International Journal of Innovative Computing, Information and Control*, vol. 6, no. 6, pp. 2803-2816, Jun. 2010.
- [29] J. Zhang, Z. Huang, and K. J. Lin, "A Hybrid Diagnosis Approach for QoS Management in Service-Oriented Architecture," in *Proc. 19th International Conf. on Web Services*, Honolulu, HI, 2012, pp. 82-89.
- [30] J. Zhang, X. Zhang, and K. J. Lin, "An efficient Bayesian diagnosis for QoS management in service-oriented architecture," in *Proc. International Conf. on Service-Oriented Computing and Applications*, Irvine, CA , 2011, pp. 1-8.
- [31] S. Srivastava and A. Sharma, "An Approach for QoS Based Fault Reconfiguration in Service Oriented Architecture," in *Proc. Information Systems and Computer Networks*, Mathura, India, 2015, pp. 766-773.
- [32] D. Miorandi, S. Sicari, F. D. Pellegrini, and I. Chlamtac, "Internet of Things: Vision, application and research challenges," *AD HOC Network*, vol. 10, no. 7, pp. 1497-1516, Sept. 2012.

- [33] L. Xu, "Enterprise System: State-of-the-art and future trends," *IEEE Transaction Industrial Informatics*, vol. 7, no. 4, pp. 630-640, Nov. 2011.
- [34] L. Atzori, A. Lera, and G. Morabito, "The internet of things: A survey," *Computer Network.*, vol. 54, no. 15, pp. 2787-2805, Oct. 2010.
- [35] D. Uckelmann, M. Harrison, and F. Michahelles, "An Architectural approach towards the future internet of things," in *Architecture the Internet of Things*, New York, USA, Spinger, 2011, pp. 1-24.
- [36] B. Perez and D. Correal, "MENTA: A model-driven architecture to enable self-adaptive SOA systems," in *Proc. 6th Colombian Computing Congress*, Manizales, 2011, pp. 1-6.
- [37] S. Cherif, R. B. Djemaa, and I. Amous, "ReMoSSA: Reference Model for Specification of Self-adaptive Service-Oriented-Architecture," in *New Trends in Databases and Information Systems*, New York, USA, Springer, 2014, pp. 121-128.
- [38] T. Cucinotta and e. al, "A real-Time Service-Oriented Architecture for Industrial Automation," *IEEE Transaction on Industrial Informatics*, vol. 5, no. 3, pp. 267-277, Aug. 2009.
- [39] J. J. Jung, "Service chain-based business alliance formation in services-oriented archietcture," *Expert Systems with appliaction* , vol. 38, no. 3, pp. 2206-2211, Mar. 2011.
- [40] G. li, V. Muthusamy, and H.-A. Jacobsen, "A Distributed Service-Oriented Architecture for Business Process Execution," *ACM Transaction on The WEB*, vol. 4, no. 1, pp. 1-33, Jan. 2010.
- [41] M. Themistocleous, N. Basias, and V. Morabito, "A Framework of Serive-Oriented Architecture Adoption in e-banking: the case of banks from a Transition and a Developed Economy," *Information Technology for Development*, vol. 21, no. 3 Special Issue, pp. 460-479, 2015.
- [42] G. Feuerlicht and S. Govardan, "SOA:Trends and Directions," in *Proc.17th International Conf. on Systems Integration*, 2009, pp. 149- 154.
- [43] X. Yang, "QoS-oriented service computing: Bringing SOA into cloud environment," in *Grid and Cloud Computing: Concepts, Methodologies, Tools and Applications*, USA, IGI Global, 2011, pp. 1621-1643.
- [44] B. Rida and E. Ahmed, "Multiview SOA : Extending SOA using a Private Cloud Computing as SAAS and DAAS," *International Journal of Software Engineering & Applications*, vol. 6, no. 6, pp. 1-11, Nov. 2015.

## Review of Retiming based Digital Filter for Low Power Consumption

\*Mandeep kaur, \*\*Ranjit Kaur

\*(Assistant Professor, Deptt. Of ECE, Punjabi University, Patiala) \*\* (Associate Professor, Deptt. Of ECE, Punjabi University, Patiala)

**ABSTRACT:** The power dissipation is the limiting factor in the digital signal processing. This paper provides the review of the low power techniques such as voltage scaling, positioning of the flip flop, reduce the switching activity along with the transformation techniques such as retiming, pipelining, and folding. Digital filters are the most common block in the signal processing applications. They are represented by data flow graph. Applying the retiming techniques, low power and high speed digital filter can be achieved. Retiming can be done through pipelining and cut set retiming. Low power filters can be achieved with the supply voltage scaling and the high speed can be controlled by minimization of register and clock period.

### I.INTRODUCTION

Digital filters are used in number of applications including predictive speech, video compression, echo cancellation, equalization, and multimedia system. Digital filters are an important class of a linear time invariant system (LTI) designed to modify the frequency properties of input signal  $x(n)$ . To meet certain specific design requirements, the causal digital filter can be characterized by the unit impulse response  $h(n)$ . The unit sample response or frequency response capture the time and frequency domain properties. The difference equation representation shows the computation required to implement the filter. A linear time invariant and causal infinite impulse filter (IIR) is described by the difference equation

$$y(n) = - \sum_{k=1}^N a_k y(n-k) + \sum_{k=0}^{M-1} b_k x(n-k) \quad (1)$$

If  $a_k = 0$  for  $1 \leq k \leq N$ , equation (1) reduced to

$$y(n) = \sum_{k=0}^{M-1} b_k x(n-k) \quad (2)$$

Which is an  $M$  – tap infinite impulse response (IIR) filter with unit sample response.

The digital filter can be studied using Very Large Scale Integrated (VLSI) to get reduced area and power and on the other hand, increase in speed which make the system suitable for high throughput applications. The reduction in power and area is used in time multiplexed application such as speech and mobile radio. On the other hand, applications such as video and radar system demands higher speed. In VLSI, the main source of power dissipation in a Complementary metal oxide semiconductor (CMOS) digital circuit is

$$P_{avg} = P_s + P_{sc} + P_l \quad (3)$$

Where  $P_{avg}$  is the average power dissipated by the circuit,  $P_s$  is the switching component of the power caused by charging and the discharging of the load capacitance  $C$ ,  $P_{sc}$  and  $P_l$  is the power dissipated due to short circuit and leakage current respectively. Both  $P_{sc}$  and  $P_l$  could be reduced to negligible level and  $P_s$  is the dominant factor for the power consumption. There are two components that contribute to the power dissipation in Complementary metal oxide semiconductor circuit (CMOS) [1]. The static power dissipation is due to leakage current but the dynamic power dissipation is due to the switching transient current and the charging and discharging of load capacitance. The major source of power dissipation is the dynamic power dissipation. The dynamic power dissipation appears in CMOS when the gates switch from one stable state to another. Thus the average power dissipation is

$$P_{avg} = \frac{1}{2} \times C_l \times V_{dd}^2 \times f_p \times N \quad (4)$$

Where  $C_l$  is the load capacitance,  $V_{dd}$  is the power supply voltage;  $N$  is the average number of gate output transition. The power consumption can be reduced if one can reduce the switching activity of a given logic circuit without changing its function [1]. The focus of low power design is to eliminate unnecessary switching activity in a logic circuit. Glitch which is due to the delay in the component can be reduced by retiming. As in the above equation, the power consumed is a quadratic function of operating voltage. Reducing the operating voltage will reduce the consumed switching power. This reducing of the voltage will increase the delay and result in reduced throughput. This reduction in throughput can be compensated by the no. of high level transformation techniques [3]. These transformation could be applied to Digital Signal Processing system by

hardware level, high algorithm level represented by using data flow graph. There are following transformation techniques:

## II.TRANSFORMATION TECHNIQUES

The following transformation provide the systematic techniques for designing the low power consumption circuit , high speed circuit and reduce in the silicon area as well as in the reduction in the hardware functional unit.

### (A)RETIMING

Pipelining and retiming are two different aspects of digital design. In pipelining, more registers are added that can change the transfer function of the system and in retiming, present registers are relocated in a structure to optimize the critical path. Digital filters are classified as the most common blocks in signal processing applications that can be represented by synchronous data-flow graphs (DFGs) [3]. Applying retiming techniques on the data flow graphs results high-speed digital circuits with the reduction of the clock period.

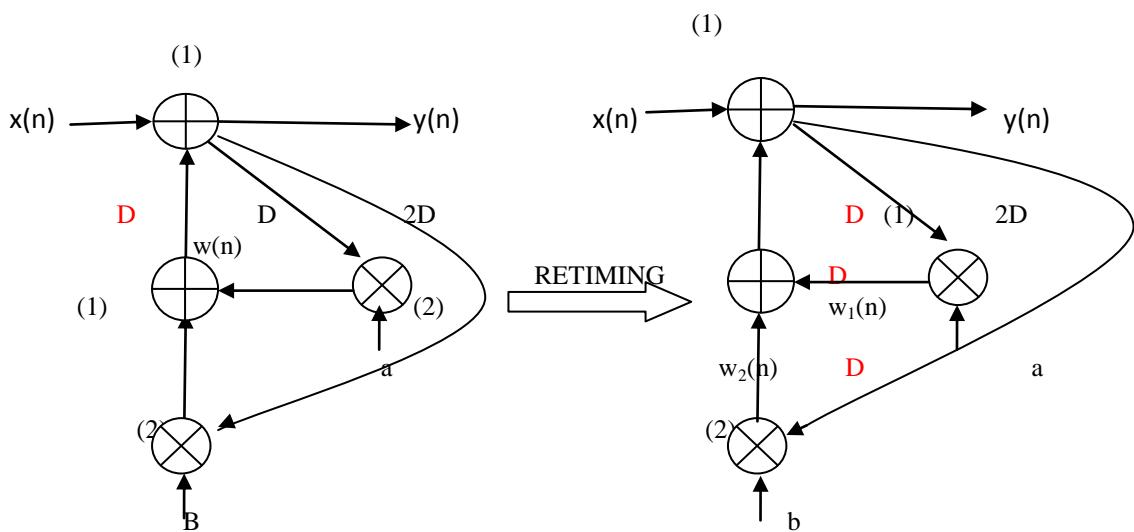


Fig 1: Retiming circuit (IIR filter)

Retiming assists in maximizing the testability of the design [4] and to enhance the performance of design. A given dataflow graph (DFG) can be retimed using forward or feedback cut-set or delay transfer approach . The weight of the retimed path p is given by

$$p = V_0 \xrightarrow{e_0} V_1 \xrightarrow{e_1} \dots \xrightarrow{e_{k-1}} V_k$$

$$w_r(p) = \sum_{i=0}^{k-1} w_r(e_i) = \sum_{i=0}^{k-1} (w(e_i) + r(V_{i+1}) - r(V_i)) \quad (4)$$

$$= \sum_{i=0}^{k-1} w(e_i) + \left( \sum_{i=0}^{k-1} r(V_{i+1}) - \sum_{i=0}^{k-1} r(V_i) \right)$$

$$= w(p) + r(V_k) - r(V_0) \quad (5)$$

Cut-set retiming is a special technique to retime a data flow graph. A cut set is usually a set of forward and backward edges within a DFG .These edges are removed from the graphthenhe graph becomes disjoint [4]. Cut set retiming only affect the weights of cut set in the cut set. If the two disconnected sub graphs are labelled as G1 & G2 then cut set retiming consist of adding d delay to each edge from G1to G2 and removing d delay from each edge from G2 to G1. Feasible solution:

$$\begin{aligned} W(G_1 \rightarrow G_2) &= W(G_1 \rightarrow G_2) + R(G_1) - R(G_2) \\ &= W(G_1 \rightarrow G_2) + K \end{aligned} \quad (6)$$

$$\begin{aligned} W(G_2 \rightarrow G_1) &= W(G_2 \rightarrow G_1) + R(G_1) - R(G_2) \\ &= W(G_2 \rightarrow G_1) - K \end{aligned} \quad (7)$$

Where  $K$  lies  $- \{W(G_1 \rightarrow G_2)\} \leq K \leq \{W(G_2 \rightarrow G_1)\}$

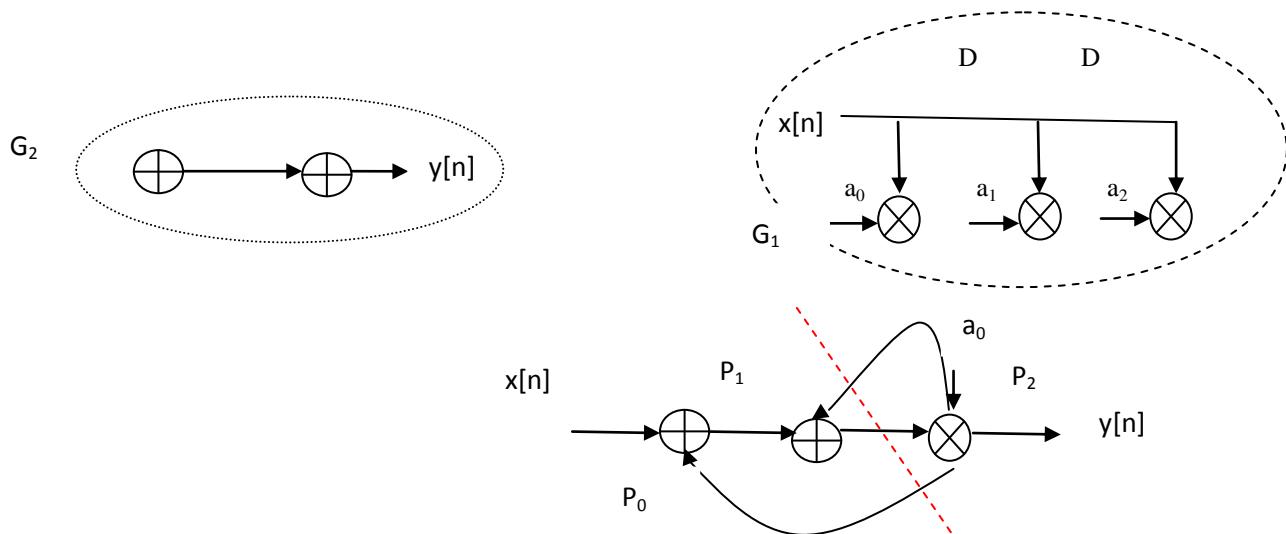


Fig 2: Cutset retiming

### (B)PIPELINING

Pipelining is a special case of feed forward cut set retiming .There is no edge in the cut set through the sub graph  $G_2$  to sub graph  $G_1$  i.e. pipelining refers to graph without loops.

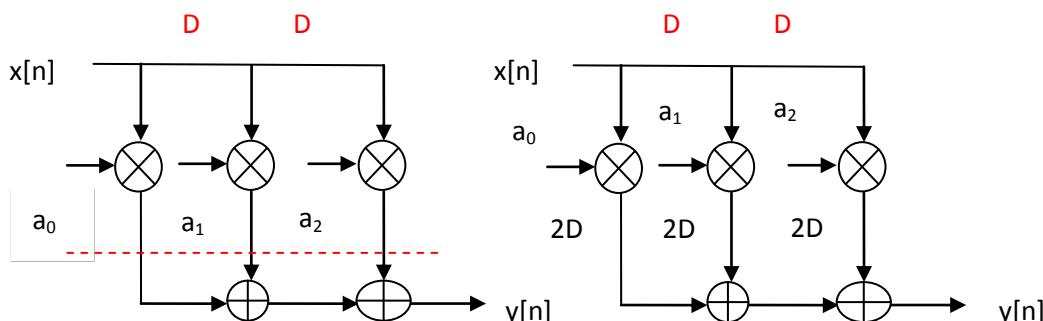


Fig 3: Cut set retiming structure with  $K= 2$  [4]

Figure 3 shows Cut set retiming structure of a FIR filter and its DFG with a valid cut-set line. The line breaks the DFG into two disjoint graphs, one consisting of nodes  $P_0$  and  $P_1$ and the other having node  $P_2$ . The edge  $P_1 \rightarrow P_2$  is a forward cut-set edge, while  $P_2 \rightarrow P_1$  and  $P_2 \rightarrow P_0$  are backward cut-set edges. There are two delays on  $P_2 \rightarrow P_1$  and one delay on  $P_2 \rightarrow P_0$ ; one delay each is moved from these backward edges to the forward edge. The retimed DFG minimizing the number of registers.

**(C) Retiming using Clock period minimization technique:** The retiming algorithm for clock period minimization is much more efficient than cut set retiming algorithm in terms of clock frequency improvement. But it has higher computational complexity of  $O(n^3 \log n)$ . The algorithm starts by building a new graph from

the original DFG. The new graph can give us a set of inequalities called the critical path constraints. The original DFG also presents a set of equalities called the feasibility constraints. A constraint graph can be built from the critical path constraints and the feasibility constraints. The retiming values for each node can be derived by applying a Floyd-Warshall shortest path algorithm to the constraint graph. The weight for each edge in the retimed DFG can be calculated using the original weight and the retiming values of the two nodes connected by this edge. The description of the algorithm is presented below.

1) Calculate  $M = t_{max}n$ , where  $n$  represents the number of nodes in the original DFG  $G$  and  $t_{max}$  is the maximum computation time of all the nodes in the DFG.

2) A new DFG  $G^*$  can be created from  $G$ .  $G^*$  has the same nodes and edges as  $G$ . For each edge in  $G^*$ , the edge weight is  $w^*(e) = M w(e) - t(u)$ , where  $w(e)$  is the edge weight of the same edge in  $G$ ,  $t(u)$  is the computation time of the node initiating this edge.

3) Apply the Floyd-Warshall algorithm to compute  $S^*_{UV}$ , which represents the shortest path from node U to node V.

4) From  $S^*_{UV}$ , calculate  $W_{UV}$  and  $D_{UV}$ . If  $U \neq V$ , then

$$W_{UV} = \left\lceil \frac{S_{UV}}{M} \right\rceil \quad (8)$$

$$D_{UV} = MW_{UV} - S_{UV} + t(V)t(U) \quad (9)$$

If  $U = V$ , then  $W_{UV} = 0$  and  $D_{UV} = t(u)$ . Here,  $t(u)$  and  $t(V)$  represent the computation times of node U and node V respectively.

5) Find the maximum value of  $D_{UV}$  and the minimum value of  $D_{UV}$ . Check all the possible clock periods starting from maximum value of  $D_{UV}$  to minimum value of  $D_{UV}$  one by one. Then find a clock period, that give a feasible solution, and find the minimal clock period. The solution contains the retiming values for all nodes. The retimed DFG from the original DFG can be obtained. This algorithm can efficiently compute the minimal clock period for a DFG.[4]

#### (D) FOLDING

In synthesizing DSP architectures, it is important to minimize the silicon area of the integrated circuits, which is achieved by reducing the numbers of functional units such as adders, registers, multipliers and interconnected wires. The folding transformation is used to systematically determine the control circuits in DSP architectures where multiple algorithm operations such as addition operations are time-multiplexed to a single functional unit such as pipelined adder. [3, 4]. By executing multiple algorithm operations on a single functional unit, the number of functional units in the implementation is reduced, resulting in an integrated circuit with low silicon area. Folding technique can be used for synthesis of DSP architectures that can be operated using single or multiple clocks.

### III. LOW POWER TECHNIQUES

Retiming is only one possible application for the presented switching activity estimation method. Glitches at the output of a logic gate can occur when the input signals arrive at different times. Switching activity can be drastically reduced by introducing latches or registers in the circuit. Different path lengths to the inputs of a gate can be compensated, so that the input data to the gate is synchronised and glitches cannot appear at the output. Another effect is that latches and registers are barriers for glitches as data can only pass through when the clock signal appears. Retiming, introduced in [7] to improve the speed of a design by rearranging register positions is now applied to reduce the switching activity. The retiming methodology is described for cyclic graphs. All primary inputs of the network are virtually connected with all outputs by a "dummy-vertex" with zero propagation delay. To every vertex  $v$  in the network a vertex-weight  $r(v)$  is assigned. In a first initialization step all  $r(v)$  are set to 0. Beginning at the primary inputs of the network, every vertex is observed stepwise in the order of the logical depth. In every step the glitch-weight is determined. If the glitch-weight for a vertex  $v$  is greater than a given maximum acceptable glitch-weight  $c_p$ , the vertex-weight is set to  $r(v) = 1$ . For every following vertex  $v_j$  of  $v$ ,  $r(v_j)$  is also set to 1 until again vertex  $v$  or a register is reached. Now a new register distribution is calculated. For every edge  $e$ ,  $u$ ,  $v$  in the network the new register count  $w_r(e)$  after retiming can be calculated from the actual register count  $w(e)$ :

$$w_r(e) = w(e) + r(v) - r(u) \quad (10)$$

If for any half the input edges of vertex  $v$ , where  $g_v > c_p$ , the edge weight  $w_r(e) > 0$  and  $w(e) = 0$ , the input edge is "marked". This means that a register was placed in this edge because of high glitching activity at the output of the considered vertex. After the new register distribution is calculated, all  $r(v)$  are set to 0 and the algorithm starts again at the primary inputs with the observation of the vertices.[14] If in one of the retiming steps a register is removed from a marked edge so that  $w_r(e) = 0$  leading to  $g_v > c_p$  at the following vertex, the algorithm stops. In this case the maximum acceptable glitch-weight  $c_p$  is not feasible. By stopping the algorithm for this case, oscillations due to altering register positions are not possible and the algorithm always terminates. If  $c_p$  is feasible, the proposed algorithm causes that registers are placed at all inputs of vertices where  $g_v > c_p$  and glitching is reduced [10]. A binary search algorithm can be used to find the minimum feasible glitch weight  $c_{p\min}$  starting with a given initial glitch-weight  $c_{init}$ . Besides low power considerations, timing constraints can also be involved so that a timing and power optimized design can be achieved. In this case the vertex-weight  $r(v)$  will be incremented if  $g_v > c_p$  or if  $\Delta(v) > c_t$ , where  $\Delta(v)$  is the critical path delay to vertex  $v$  and  $c_t$  is the desired minimum  $c$  another method is the positioning of flip flop .

Consider the following figure 4. If the average switching activity (during a clock cycle) at the output of the gate  $g$  is  $E_g$  and the load capacitance is  $C_L$ , then the power dissipated by the circuit is proportional to  $E_g \cdot C_L$ . Consider when the flip flop  $R$  is added to the output of  $g$ , then the power dissipated by the circuit is now proportional to  $E_g \cdot C_R + E_R \cdot C_L$  where  $C_R$  is the capacitance at the input of the flip flop and  $E_R$  is the average switching activity at the flip flop output and  $E_R < E_g$ . Suppose if the gate  $g$  may glitch and make three transitions but the flip flop will make at most one transition when the clock is asserted. This implies that it is possible that  $E_g \cdot C_R + E_R \cdot C_L$  is less than  $E_g \cdot C_L$  if both  $E_g$  and  $C_L$  are high. Thus the addition of flip flop to the circuit decreases the power dissipation.[12]

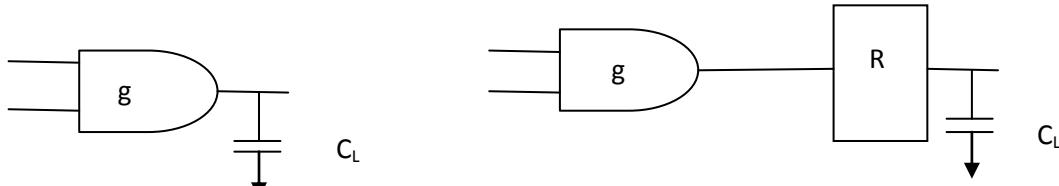


Fig 4: Flip flop in a circuit [12]

Consider the following circuit shown in figure 5. The power dissipated by the circuit is proportional to  $E_0 \cdot C_R + E_1 \cdot C_{L1} + E_2 \cdot C_{L2}$ . The power dissipated in fig (b) is proportional to  $E_0 \cdot C_{L1} + E_1' \cdot C_R + E_2' \cdot C_{L2}$ . One circuit may have lesser power dissipation than the other. Due to glitching,  $E_1'$  maybe greater than  $E_1$  but by the same  $E_2'$  may be less than  $E_2$ . The capacitance of the logic blocks and the flip flop along with the switching activity will determine which of the circuit is more desirable from the power standpoint.

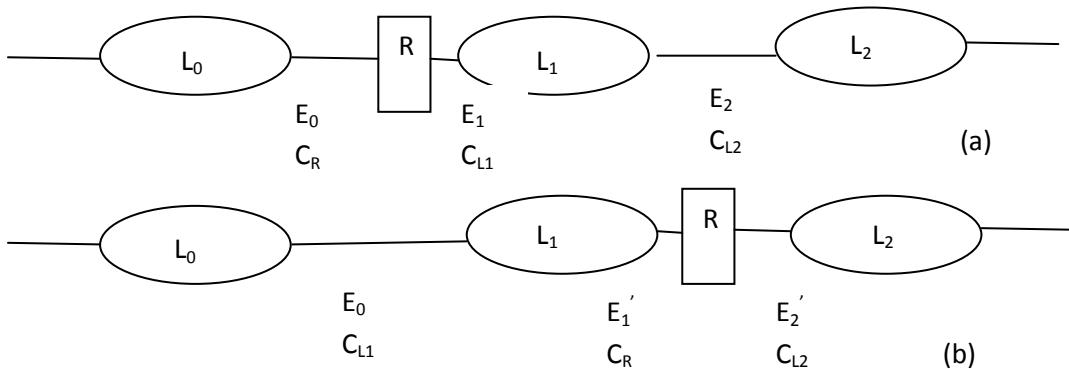


Fig 5: Moving a flip flop in a circuit.[12]

Retiming algorithm minimize the clock period and the delay varies linearly under retiming. The algorithm used for reducing the power dissipation in a pipelined circuit select the set of nodes which by having a flip flop placed at their outputs leads to the minimization of switching activity in the network. Nodes are selected based on the amount of glitching that is present at their output and on the probability that this glitching propagates through their transitive fan-out.[12,4]

For estimating the average switching activity, consider zero delay ( $E_{zeroD}$ ) and actual delay ( $E_{genD}$ ) for each gate , obtaining the amount of glitching ( $E_{glitch}$ ) at each gate by taking the difference of the expected number of transitions ( $E_{glitch} = E_{genD} - E_{zeroD}$ ). The probability of a transition at each gate propagates through its transitive

fan out. For each gate j in the transitive fan-out of node i , the probability of having a transition at node j caused by a transition at gate i is

$$s_{j,i} = \frac{P(i \uparrow \wedge j \uparrow)}{P(i \uparrow)} \quad (11)$$

Where  $P(i \uparrow)$  is the probability of a transition at node i , calculated using the method of [3]. The value of  $P(i \uparrow \wedge j \uparrow)$  can be calculating the primary input conditions under which a transition at i triggers a transition at j[3]. To reduce the power , place the flip flop at the output of a node i is:

$$\text{power\_red}(i) = E_{\text{glitch}}(i) \times \left( C_i + \sum_j^{\text{fanout}} (s_{j,i} \times C_j) \right) \quad (12)$$

Other factors that contribute to the power dissipation is the no. of flip flop in the network. The higher weight to the nodes had larger number of inputs ( $n_i(i)$ ) and output ( $n_o(i)$ ) . A flip flop placed at one of these nodes will be in a greater number of paths, reducing the total number of flip flop needed. The final cost function is

$$\text{weight}(i) = \text{power\_red}(i) \times C_i(i) + n_o(i) \quad (13)$$

#### IV. CONCLUSION

This paper has been discussed the low power estimation techniques which include the placement of registers, clock minimization with the help of retiming, voltage scaling, reducing the glitch in a circuit as well as the use of pipelining an folding with retiming to lower the power. The retiming technique is used to restructure the DFG representation in order to reduce the number of clock cycle, reduce the number of register. In future, these techniques can be generated with the evolutionary algorithm to reduce the power consumption.

#### V. REFERENCES

- [1] Yu-Lung Hsu , Sying-Jyan Wang , Retiming-Based Logic Synthesis for Low-Power *ISLPED'02* ACM August 12-14, 2002.
- [2] P.Duncan , S.Swamy, R.Jain . Low Power DSP circuit Design using retimed maximally Parallel architectures. In Proceedings of the 1<sup>st</sup> Symposium on Integrated Systems, March 1993, pp. 266- 275.
- [3] C.E. Leiserson , F. M . Rose, J.B. Saxe. Optimizing synchronous Circuitary By Retiming, In Proceeding of 3<sup>rd</sup> CalTech Conference on VLSI, March 1983, pp. 23-36 .
- [4] K.K. Parhi, VLSI Digital Signal Processing Systems (Design and Implementation), John Wiley & Sons, Inc., New York, 2000.
- [5] T. C. Denk, K. K. Parhi, Synthesis of Folded Pipelined Architectures for Multirate DSP Algorithms, IEEE Transaction on Very Large Scale Integration (VLSI) Systems, Vol.6, No. 4, Dec. 1998, pp. 595-607
- [6] N. Chabini, I. Chabini, E.-M. Aboulhamid, and Y. Savaria, “Unification of basic retiming and supply voltage scaling to minimize dynamic power consumption for synchronous digital designs,” in *Proc. Great Lakes Symp. VLSI*, Washington, DC, Apr. 2003.
- [7] J.-M. Chang and M. Pedram, “Energy minimization using multiple supply voltages,” *IEEE Trans. VLSI Syst.*, vol. 5, pp. 1–8, 1997.
- [8] C. E. Leiserson and J. B. Saxe, “Retiming synchronous circuitry,” *Algorithmica*, Jan. 1991, pp. 5–35.
- [9] J. Monteiro, S. Devadas, and A. Ghosh, *Retiming sequential circuits for low power*, in *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, 1993, pp. 398–402.
- [10] A. Raghunathan, N. K. Jha, and S. Dey, *High-Level Power Analysis and Optimization*. Norwell, MA: Kluwer, 1997.

[11] Noureddine Chabini and Wayne Wolf ,*Reducing Dynamic Power Consumption in Synchronous Sequential Digital Designs Using Retiming and Supply Voltage Scaling*, IEEE Transactions on VLSI System, Vol. 12, No. 6, June 2004

[12]A. K. Allam and J. Ramanujam , *Simultaneous Peak and Average Power Optimization in Synchronous Sequential Designs Using Retiming and Multiple Supply Voltages* , ICICDTO6.

[13] A.Chandrakasan , T. Sheng ,R.W . Brodersen. *Low Power CMOS Digital design in Journal of solid State Circuits*, April 1992 , pp. 473-484.

[14] Christian V. Schimpfle, Sven Simon and Josef A. Nossek , *Optimal Placement of Registers in Data Paths for Low Power Design IEEE International Symposium on Circuits and Systems, June 9-12, 1997, Hong Kong*

# De-speckling of B-Mode Breast Ultrasound Images Using Wavelet shrinkage filters: A Comparative Analysis

Madan Lal, Dr.Lakhwinder Kaur  
Department of Computer Engineering  
Punjabi University, Patiala  
India

**Abstract-** Speckle noise is an inherent yet undesirable residual part of the breast ultrasound images, which significantly demean the visual quality and limits the accuracy of automatic diagnostic techniques. Therefore, speckle elimination is a necessary task before further processing of ultrasonic images. Speckle reduction from breast ultrasound images results in blurring of lesion margins and other sharp details which may carry the significant diagnostic information. Among various denoising methods used in the literature for enhancement of breast ultrasound images, wavelet based techniques are gaining importance, because of their time-frequency and multi-scale analysis. Basic de-noising methods that use the wavelet transform, make use of three steps – In the first step, it computes the wavelet transform of the noisy input image, the second step is used to apply thresholding in order to remove noise on the detailed coefficients and finally inverse wavelet transform is applied of the modified coefficients to get the denoised image. In this paper, the performance of various wavelet shrinkage techniques is reviewed. An Experimental analysis of wavelet based methods including Visu Shrink, Sure Shrink, Bayes Shrink, and a hybrid method (wavelet shrinkage with guided filter) is carried out. For performance comparison, signal to noise ratio, structural similarity index and edge preservation index, parameters are used. From the experimental results, it is concluded that hybrid filter outperform the traditional wavelet shrinkage methods.

**Keywords:** Speckle Noise, Breast Ultrasound (BUS) image, Wavelet shrinkage, Signal to Noise Ratio (SNR), Structural Similarity Index (SSIM), Edge Preservation Index (EPI)

## I. INTRODUCTION

Breast cancer is the most recurrent cancer and a leading cause of deaths among women worldwide [1]. In the United States, the likelihood of developing invasive breast cancer in the woman's life is nearly 1 in 8 [2]. For breast cancer detection, mammography and sonography are most sensitive modalities [3]. Earlier, mammography was considered as a most effective way for cancer disclosure, but due to its low specificity, 65%-68% unnecessary biopsies are performed [4], which is an expensive and painful process. It also adds to the emotional burden of the patient.

Now, among all the imaging techniques used as a part of the medical science, for diagnostic purpose of breast cancer, ultrasonic frameworks are opted for revealing a minimal danger to the patients. This is on the basis that non-perceptible sound waves with frequencies above 20 kHz are not known to cause any unfavorable effects in patients. Accordingly, the clinical application of ultrasonic imaging technology has turned out to be more essential. On the other hand, like all coherent imaging techniques, the major flaw of an ultrasonic imaging is that it is contaminated by speckle noise, which is produced by an interaction of the reflected waves from different autonomous scatters inside a cell determination [5].

In BUS images, Speckle is considered to be the foremost source of noise in BUS images and it should be filtered out without affecting fine details i.e. object boundaries, sharp discontinuities, and other important features of the input image. In the literature, many de-speckling methods have been proposed. One variety of speckle reduction technique is called linear filtering process [6-8], which works based on the linear combination of the intensity values of pixels in the input image, but these methods results in blurring of edges, and sometimes destroys some important diagnostic features. Another variety of despeckling methods are called nonlinear filters,. These filters produce better results, but may produce over-smoothening in non-uniform regions [9].

To deal with the harms of linear filtering techniques, Perona and Malik [10] introduced a nonlinear filter, called anisotropic diffusion (AD) filter. It is a partial differential equation based filter, which simultaneously perform contrast enhancement and speckle reduction of input image. Yu and Acton [11], proposed another diffusion based filter, called speckle reducing anisotropic diffusion (SRAD) filter, which iteratively process a noisy image using adaptive weighted filters, reduce noise and preserve edges. However, it emphasizes edge enhancement rather than visualization improvement.

Fast Fourier Transformation (FFT) based de-noising technique is also introduced, but it is unable to preserve sharpness of edges because it is a low pass filtering technique whose basic function is not being localized in terms of time and space domain [12]. This problem can be resolved by using wavelet transform, due to its localized nature in terms of time and space domain.

In this paper, various wavelet based speckle reduction techniques are reviewed and their performance is compared using quality parameters.

## II. THEORETICAL BACKGROUND

### A. Speckle Noise

Speckle is a multiplicative noise; it occurs when a sound wave pulse randomly interferes with small particles that are classically much smaller than the wavelength of an ultrasound wave [13]. The statics of the speckle depend on the scatter density within the resolution cell. It has a granular pattern and it is the inherent property of BUS images.

### B. Speckle Noise model

It is imperative to understand the speckle model before the noise removal in BUS images. When high frequency sound waves are emitted on the parts of the human body, a number of scatters reflect the incident wave towards sensor for each resolution cell. The received signal consists of two parts, the actual signal reflected by human body cells and added noise component. The noise component is thus a random granular pattern called speckle. Many statistical models have been investigated for describing the characteristics of the US speckle under different scattering conditions. These models include the well-known Rayleigh model (it works based on the assumption that a resolution cell contains a large number of scatters) and a class of non-Rayleigh models comprising of  $K$ -distribution, Weibull distribution, log-normal distribution and Nakagami distribution [14-16].

A generalized model of ultrasonic envelope resolution can be written as:

$$f(x, y) = g(x, y) * \eta(x, y) \quad (1)$$

Where  $f(x,y)$  and  $g(x,y)$  are observed noisy image and noise free image respectively, and  $n(x,y)$  is the fading variable modeled as a stationary unit-mean random process independent of noise free image  $g(x,y)$  [17].

By applying logarithm operation, the multiplicative model can be converted into an additive one as represented in (2). Log compression is used to minimize the dynamic range of an image.

$$\log(f(x,y)) = \log(g(x,y)) + \log(\eta(x,y)) \quad (2)$$

### C. Discrete wavelet transforms

In discrete wavelet transform (DWT), the input signal is discretely sampled. It tries to represent a function in terms of small waves, thus called wavelet. DWT gives non-redundant and unique representation to a signal. Using DWT, both the time and frequency analysis of a signal can be done simultaneously; consequently wavelet transform has turned out to be a unified framework for de-noising.

#### I. Decomposition process

Using decomposition, initial two dimensional data is replaced with four blocks. These blocks represent the sub bands which correspond to either low pass or high pass filtering in each direction. Wavelet decomposition is performed on rows and columns of two dimensional data consecutively. In the first step, transform applies to all rows yielding a matrix of whose left side contains down sampled low pass coefficients of each row and right side contains the high pass coefficients. The second step of decomposition is applied to all columns resulting in four types of coefficients; LL, HL, LH and HH as shown in Fig.1.

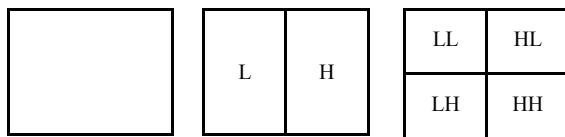


Fig.1 Wavelet decomposition

Thus, the original image is divided into four sub-images of size  $N/2 \times N/2$ . Here the LL sub band contains the low frequency components in both directions. It is also known as an approximation sub band. The LH, HL and HH sub bands contain the details components in vertical, horizontal and diagonal directions respectively. The HH sub band is obtained by high-pass filtering in both directions and includes the high-frequency components along the diagonals as well. The HL and LH sub bands are the outcome of low-pass filtering along rows and high-pass filtering along the column. All three sub bands HL, LH and HH are also known as the detail sub bands, because they append the high-frequency information or details to the approximation sub band.

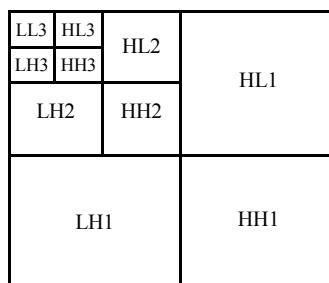


Fig.2 Pyramidal decomposition

For pyramidal decomposition, the LL sub band is further decomposed as shown in Fig.2

### III. WAVELET BASED DE-NOISING

The main thought of wavelet based de-noising methods is to preserve the wavelet coefficients of low frequency components (LL) and to shrunk the high frequency components (LH,HL,HH) using a wavelet threshold function.

#### A. Wavelet de-noising procedure

Wavelet based image de-noising using thresholding procedure involves following basic steps.

- Step1. Apply the log transformation on the noisy input image
- Step2. Apply wavelet Transform
- Step3. Change noisy coefficients using threshold value
- Step4. Apply Inverse Wavelet Transform
- Step5. Apply exponential transformation to produce de-noised image

In images, noise has a fine grained structure so most of the noise represented by wavelet coefficients is also at finer scales. But at finer scales, wavelet coefficients also carry edge information so wavelet coefficient values below a certain threshold are set to be zero, because edge related coefficients are usually higher than the threshold [17]. The inverse wavelet transform of modified wavelet coefficients represents the de-noised image.

#### B. Wavelet shrinkage

Wavelet shrinkage is the main process responsible for image de-noising and its functionality depends upon the threshold selection method. If the threshold value is small then some noise coefficients which have a larger value than the threshold will be retained as useful signal this in turn leads to the results that the output image still possesses noise. On the other side if the threshold value is large then some useful coefficients values will be zeroed, which results in a loss of important information. So, an appropriate threshold selection is of paramount importance because its values directly affect the final de-noising capability of the wavelet process.

##### 1) Thresholding function

There are mainly two thresholding functions proposed, hard thresholding and soft thresholding [18]. In hard thresholding, coefficients less than certain threshold value T are set as zero while other coefficients remain unchanged. It is represented by the following equation (3).

$$w_a(t) = \begin{cases} w_b(t), & \text{for } |w_b(t)| \geq T \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Where  $w_a(t)$  represent the de-noised wavelet coefficient and  $w_b(t)$  represents the noisy wavelet coefficient and T denotes the threshold value. While in Soft thresholding, coefficients less than threshold T are set to zero while important coefficients reduced by absolute threshold value [19] as represented by the equation (4).

$$w_a(t) = \begin{cases} \operatorname{sgn} w_b(t), (w_b(t) - T), & \text{for } |w_b(t)| \geq T \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Where the  $\text{sgn}(\cdot)$  is a Signum function, which return 1 if value is greater than zero, zero if value is zero and -1 if value is less than zero. Generally soft thresholding outperforms the hard thresholding, but in some cases hard thresholding yield superior results [20].

## 2) Threshold selection methods

Different methods available for threshold selections are universal threshold, sub-band adaptive threshold and spatially adaptive threshold. In universal method, a threshold value is uniquely selected for all wavelet coefficients [21]. In sub-band adaptive technique, threshold is selected differently for each detailed sub band [18] and in spatially adaptive threshold selection method; each detail coefficient has its own threshold value [22].

## IV. WAVELET DE-NOISING METHODS

Numerous wavelet based de-noising methods have been developed. Some popular wavelet shrinkage methods are Sure shrink, Visu shrink, Bayes shrink and hybrid methods. The methods are described in the following section.

### A. Visu Shrink

In this method, the threshold is calculated by using a universal threshold method [18]. It uses hard thresholding rule where threshold value is directly proportional to noise standard deviation in each sub-band of ultrasound image after decomposition. If  $m$  is the size of sub-band in the wavelet domain, then the threshold of Visu shrink is defined as:

$$\lambda = \sigma \sqrt{2 \log m} \quad (4)$$

Where  $m$  is the number of wavelet coefficients in the corresponding wavelet domain and  $\sigma$  is the standard deviation of noise that is obtained using following equation (5).

$$\sigma = \frac{\text{median}(w_y)}{0.6745} \quad (5)$$

Where,  $w_y$  is the wavelet coefficient of the HH1 band of Noisy Image. The main limitation of Visu shrink is that it does not deal with the mean squared error and generates an over smoothed output.

### B. Sure Shrink

Donoho and Johnstone [18] developed another threshold selection method which is based on Stein's Unbiased Risk Estimator (SURE) and known as SureShrink. Level dependent threshold value  $\lambda_j$  for each resolution level  $j$  in wavelet transform is calculated using equation (6).

$$\lambda_j = \min(\lambda, \sigma_n \sqrt{2 \log m}) \quad (6)$$

Where,  $\sigma_n$  is the estimated noise variance,  $\lambda$  indicates the values that decrease Stein's Unbiased Risk Estimator and  $m$  is the size of the image. It uses the soft thresholding technique. This shrinkage rule uses the thresholding that is adaptable, i.e., a threshold level is allotted to every resolution level by the standard of reducing the Stein's Unbiased Risk Estimator for threshold estimates. The main improvement of Sure Shrinkage rule is to reduce the mean squared error (MSE), certainly not like Visu Shrink.

### C. Bayes shrink

Bays shrink uses soft thresholding and it is also sub-band dependent as sure shrink, which implies that thresholding level is chosen at every sub-band of resolution in discrete wavelet decomposition [23]. As noisy image is an additive sum of original image and noise. It can be represented in terms of variance as:

$$\sigma_y^2 = \sigma_x^2 + \sigma_n^2 \quad (7)$$

Where  $\sigma_x^2$  is the variance of original image,  $\sigma_n^2$  is the variance of noise and  $\sigma_y^2$  is the variance of a noisy image. The bays threshold on a given sub-band of the image S is defined as:

$$\lambda = \frac{\sigma_n^2}{\sigma_s^2} \quad (8)$$

Where  $\sigma_n^2$  represents the variance of noise  $\sigma_s^2$  is the variance of noise-free original signal. The noise variance is evaluated from the detailed coefficient of finest sub- band HH1. Value of  $\sigma_s^2$  is calculated using equation (9).

$$\sigma_s^2 = \sqrt{\max(\sigma_y^2 - \sigma_n^2), 0} \quad (9)$$

In Bayes shrink, thresholding is done in the wavelet decomposition at each sub band which improves outcome and also completely denoise the flat regions of the image, but it is less sensitive to the noise around edges.

### D. Hybrid method [26]

The threshold selection method given in equation (4) by Donoho and Johnstone [18] depends on the number of wavelet coefficients. When  $m$  becomes very large, it results in larger threshold value which in turn smooths out some useful information. So, this threshold method is not very useful for noise removal in BUS images. To balance the speckle suppression and edge preservation in ultrasonic images, J. Zhang et al. [26] proposed a new threshold selection method given by the following equation (10).

$$\lambda_i = k_i \sigma_n \sqrt{2 \log m} \quad (10)$$

Where ( $i=1,2\dots I$ ) represents the number of decomposition layers of wavelet transformation and  $I$  represent the largest decomposition layer,  $a_j$ , represents the adaptive parameter of  $i^{th}$  layer and its value is calculated as:

$$a_i = \frac{1}{\ln(i+1)} \quad (11)$$

This new threshold function weakens the influence of  $m$ . The high frequency sub-band in each layer is processed by improved wavelet shrinkage threshold function given in equation (10), but the low frequency sub-band of last layer also contains a large amount of noise, which is processed by the guided filter [27]. Finally inverse wavelet transform (IWT) is applied to the denoised wavelet coefficients and a noise free BUS image is produced. Guided filter not only denoise the input image, but also have greater edge preservation ability.

## V. EXPERIMENT AND RESULTS

Experiments were conducted on the machine comprises Intel(R), Core i5 processor with 4 GB system memory and Matlab-2012 as simulation software. For performance evaluation of wavelet based filtering techniques, 45 BUS images were collected from Rajindra medical college and hospital, Patiala. For quantitative evaluation, two types of images were used, a synthetic test image and real time BUS images (Fig.3). Synthetic test image consist of regions with uniform intensity, sharp edges and strong scatters (as shown in Fig.3(a)) and real time BUS images consist of real time ultrasonic patterns. For performance evaluation, Speckle of different values ( $\sigma=0.1$  to  $0.5$ ) and ( $\sigma=0.01$  to  $0.05$ ) were simulated on Test image Synth.tif [29] and on Real time BUS image (Fig.3(c)), respectively. Speckle noise was added using speckle simulator given by Aleksandra Pizurica et al. [24]. For noise reduction, Haar wavelet is used up to two decomposition levels.

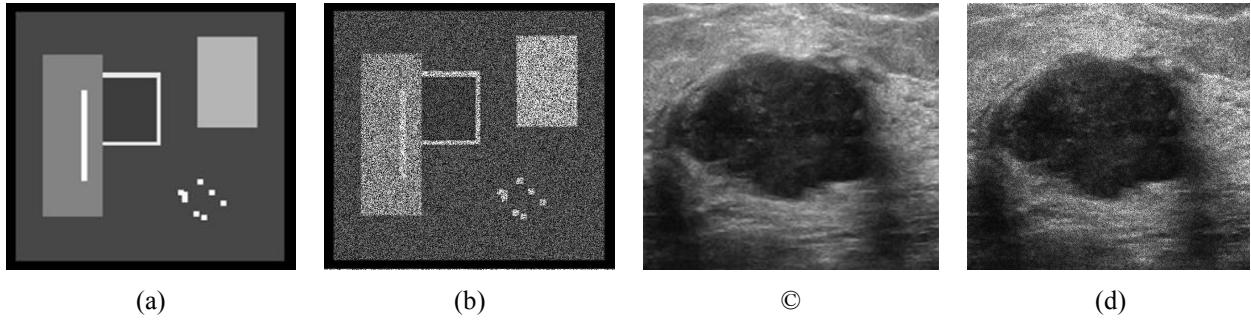


Fig.3: a) Test image (b) Speckle simulated test image ( $\sigma=0.5$ ) (c) Real time BUS image (d) Speckle simulated BUS image ( $\sigma=0.05$ )

### A. Quality metrics

The quality of de-noised image is measured by using traditional distortion measures, such as Signal to Noise Ratio, Edge Preservation Index and Structural Similarity Index between the original and reconstructed images.

#### 1) Signal to noise ratio (SNR)

It is expressed in db and defined as [24]:

$$SNR = 10 \log_{10} \left( \frac{\sigma_g^2}{\sigma_e^2} \right) \quad (10)$$

Where  $\sigma_g^2$  is the variance of original noise free image and  $\sigma_e^2$  is the variance of the error (between original and de-noised image). The larger SNR value corresponds to good image quality.

#### 2) Structural Similarity Index (SSIM) [25]

SSIM is a measure of similarity between original and filtered image. It is calculated using equation (11). The measure between two images X and Y of common size is calculated as:

$$SSIM = \frac{(2\mu_x\mu_y + 2.55)(2\sigma_{xy} + 7.65)}{(\mu_x^2 + \mu_y^2 + 2.55)(\sigma_x^2 + \sigma_y^2 + 7.65)} \quad (11)$$

Where  $\mu_x, \mu_y$  are mean of X and Y.  $\sigma_x^2$  is the variance of X and  $\sigma_y^2$  is the variance of Y.  $\sigma_{xy}$  is the covariance of X and Y. The resulting value of SSIM lies between -1 and 1.

3) *Edge Preservation Index [28]*

EPI is used to measure the edge preservation capability of a filter. It is calculated as:

$$EPI = \frac{\sum(\Delta I - \bar{\Delta I}) \sum(\Delta F - \bar{\Delta F})}{\sqrt{\sum(\Delta I - \bar{\Delta I})^2 \sum(\Delta F - \bar{\Delta F})^2}} \quad (12)$$

Where  $\Delta I$  and  $\Delta F$  are high pass filtered versions of the image  $I$  and  $F$ , obtained with a 3x3 pixel standard approximation of Laplacian operator. The larger value of EPI means more ability to preserve edges.

#### A. Results

Results of various quality parameters for different speckle noise values are stored in the following tables. Table 1 shows the output values for SNR, SSIM and EPI. These values are recorded as an output of different shrinkage filters, after adding different speckle noise ( $\sigma=0.1, 0.3, 0.5$ ) in the input test image [29] shown in Fig. 3(a).

Table.1 Quantitative results for Test image.

	Speckle Noise value ( $\sigma = 0.1$ )			Speckle Noise Value ( $\sigma = 0.3$ )			Speckle Noise Value ( $\sigma = 0.5$ )		
	SNR	SSIM	EP1	SNR	SSIM	EP1	SNR	SSIM	EP1
Visu Shrink [18]	17.31	0.6733	0.6433	16.12	0.6453	0.6217	15.01	0.6113	0.6044
Sure Shrink [19]	19.37	0.6905	0.6521	18.15	0.6607	0.6438	17.05	0.6340	0.6217
Bays Shrink [23]	20.45	0.7062	0.6720	19.36	0.6742	0.6689	18.33	0.6495	0.6395
Wavelet shrinkage with guided filter. [26]	21.87	<b>0.7311</b>	<b>0.7066</b>	<b>20.27</b>	<b>0.6912</b>	<b>0.6933</b>	<b>19.21</b>	<b>0.6717</b>	<b>0.6672</b>

In Table 2, values of various quality parameters are recorded after applying different filters on the real time BUS image and after adding the speckle noise of different values ( $\sigma=0.01, 0.03, 0.05$ ).

Table.2 Quantitative results for real time BUS image.

	Noise value ( $\sigma = 0.01$ )			Noise Value ( $\sigma = 0.03$ )			Noise Value ( $\sigma = 0.05$ )		
	SNR	SSIM	EP1	SNR	SSIM	EP1	SNR	SSIM	EP1
Visu Shrink [18]	26.13	0.8735	0.8833	25.33	0.8469	0.7827	24.05	0.8123	0.6841
Sure Shrink [19]	27.73	0.8905	0.8924	26.17	0.8627	0.7941	25.34	0.8349	0.7012
Bays Shrink [23]	28.26	0.8967	0.9127	27.27	0.8742	0.8052	26.45	0.8428	0.7165
Wavelet shrinkage with guided filter. [26]	29.67	<b>0.9013</b>	<b>0.9367</b>	<b>28.71</b>	<b>0.8802</b>	<b>0.8313</b>	<b>27.96</b>	<b>0.8537</b>	<b>0.7405</b>

#### B. Analysis

According to the quantitative results stored in Table.1 and Table.2, it can be observed that SNR values for hybrid filter (wavelet shrinkage with guided filter) are higher than the SNR values of other wavelet shrinkage methods at

different noise levels. SSIM values are also better for Hybrid method, which implies that the output image given by the wavelet shrinkage with guided filter is more similar to the input image than the resulting images of simple wavelet shrinkage filters. Similarly EPI values are also desirable, i.e. EPI values given by Hybrid method is 0.9367 for noise level 0.01, where the input image is a real time BUS image. It indicates that the combination of wavelet shrinkage (for high frequency components) with guided filter (for low frequency components) reduces more noise from BUS images along with better edge preservation.

## VI. Conclusion

In this paper, the performance of various speckle noise removal techniques using different wavelet shrinkage methods is compared. For comparison purpose, various image quality parameters (SNR, SSIM, and EPI) are used. The values of these parameters are recorded for a synthetic test image with noise levels  $\sigma=0.1, 0.3, 0.5$  and for real time BUS images with noise levels  $\sigma=0.01, 0.03, 0.05$ , respectively. From the results, it can be concluded that wavelet shrinkage with guided filter outperform the other traditional shrinkage filters. From the quantitative outcome, it is also clear that the hybrid filter reduces more noise and simultaneously it preserves more edges. However the different numerical implementation of these methods may give different results under varied conditions.

### ACKNOWLEDGMENT

Authors are thankful to Dr. Navkiran Kaur, Head, Department of Radiology, Rajindra Medical College and Hospital, Patiala, India for providing real time Bus image.

### REFERENCES

- [1] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.
- [2] A. C. Society. Learn about breast cancer, <http://www.cancer.org>
- [3] H. Zhi, B. Ou, B. Luo, X. Feng, Y. Wen, H. Yang, 'Comparison of ultrasound elastography, mammography, and sonography in the diagnosis of solid breast lesions', *Journal of Ultrasound in Medicine*, Vol.26, No.06, pp. 807–815, 2007.
- [4] J. Jesneck, J. Lo, J. Baker, 'Breast mass lesions: computer-aided diagnosis models with mammographic and sonographic descriptors', *Radiology*. Vol. 244, No.02, pp.390–398, 2007.
- [5] Indrajeet Kumar, H.S Bhadauria, Jitendra virmani, Jyoti Rawat, "Reduction of Speckle noise from Medical Images using Principal Component Analysis Image Fusion" 9th International Conference on Industrial and Information Systems (ICIIS) , pp.1 – 6, 15-17 Dec. 2014.
- [6] J. S. Lee, "Speckle suppression and analysis for synthetic aperture radar," *Opt. Eng.*, Vol. 25, no. 5, pp. 63643, 1986.
- [7] S. Frost, J. A. Stiles, K. S. Shanmugam, and J. C. Holtzman, "A model for radar images and its application to adaptive digital filtering of multiplicative noise," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. PAMI-4, no. 2, pp. 15765, 1982.
- [8] D. T. Kaun, A. A. Sawchuk, T. C. Strand, and P. Chavel, "Adaptive restoration of images with speckle," *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-35, pp. 37383, 1987.
- [9] S. K. Ueng, C. L. Yen, and G. Z. Chen, "Ultrasound image enhancement using structure-based filtering," *Comput. Math. Methods Med.*, Vol. 2014, pp. 114, 2014.
- [10] P. Perona, and J. Malik, "Scale space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 12, no. 7, pp. 62939, 1990.
- [11] Yongjain Yu and Scott.T. Acton "Speckle Reducing Anisotropic Diffusion", *IEEE Transactions on image processing*, Vol. 11, No, 11, pp-1260-1270,2002.
- [12] E. R. McVeigh, R. M. Henkelman and M. J. Bronskill, Noise and Filtration in Magnetic Resonance Imaging, *Med. Phys.*, vol. 12, No. 5, pp. 586-591, 1985.
- [13] Burckhardt CB, Speckle in ultrasound B-mode scans, *IEEE T Son Ultrason*, Volume 25, 1-6, 1978.
- [14] P.M. Shankar, Ultrasonic tissue characterization using a generalized Nakagami model, *IEEE Trans. Ultrason. Ferroelectr. Frequency Control* 48 (6), pp. 1716–1720.2001.

- [15] P. Shankar, J. Reid, H. Ortega, C. Piccoli, B. Goldberg, Use of non-Rayleigh statistics for the identification of tumors in ultrasonic B-scans of the breast, *IEEE Trans. Med. Imag.* 12 (4), pp. 685–692,1993.
- [16] P.M. Shankar, V.A. Dumane, J.M. Reid, V. Genis, F. Forberg, C.W. Piccoli, B.B. Goldberg, Classification of ultrasonic B-mode images of breast masses using Nakagami distribution, *IEEE Trans. Ultrason. Ferroelectr. Frequency Control* 48 (2) pp. 569–580.2001.
- [17] A. Dixit, P.Sharma, “A comparative study of wavelet thresholding for image de-noising”, *I.J. Image, Graphics and Signal Processing*, 12, pp. 39-46,2014 .
- [18] D. L. Donoho and I. M. Johnstone, “Adatpting to unknown smoothness via wavelet shrinkage” , *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200-1224, December 1995.
- [19] D. L. Donoho, “De-noising by soft thresholding,”, *IEEE Trans. on Information Theory*, Vol.41, no. 3, pp. 613-627, 1995.
- [20] F. Xiao, Yungang Zhang, “A Comparative Study on Thresholding Methods in Wavelet-based Image Denoising,” *Elsevier Advanced in Control Engineering and Information Science*, vol. 15, pp. 3998 – 4003, 2011.
- [21] D. L. Donoho, I. M. Johnstone, “Ideal spatial adaptation by wavelet shrinkage,” *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
- [22] M. K. Mihcak, I. Kozintsev, K. Ramchandran, P. Moulin, “Low-complexity image denoising based on statistical modeling of wavelet coefficients,” *IEEE Signal processing*, vol. 6, pp. 300-303, 1999.
- [23] S.G. Chang, Y.Bin and Martin Vetterli, Adaptive Wavelet Thresholding for Image Denoising and Compression, *IEEE Trans. Image Processing*, 9(9): pp.1532-1546, 2000.
- [24] A. Pizurika et al. “A versatile wavelet domain noise filtration technique for medical imaging”, *IEEE Transactions Medical Imaging*, Vol. 22 No.2, 2003.
- [25] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, *IEEE Trans. Image Process.* 13 (4), pp. 600–612, 2004.
- [26] Ju Zhang, Guangkuo Lin, Lili Wu, Yu Cheng, "Speckle filtering of medical ultrasonic images using wavelet and guided filter" Published in *Ultrasonic Volume* 65, pp. 177-193, 2016.
- [27] Kaiming He, Jian Sun, Xiaou Tang, Guided image filtering, in: *Proceedings of 11th European Conference on Computer Vision*, Berlin, pp. 1–14,2010.
- [28] S. Gupta, L. Kaur, R. C. Chauhan, and S. C. Saxena, “A versatile technique for visual enhancement of medical ultrasound image,” *Journal of Digital Signal Processing*, Elsevier, vol. 17, pp. 542-560, January 2007.
- [29] <http://telin.ugent.be/~sanja/>

# WEIGHTED CLUSTERING ALGORITHM IN ADHOC NETWORK-REVIEW

\*Jagdeep Singh

M-Tech Research Scholar

Department of E.C.E. in Punjabi university Patiala

Jagdeep.mann3@gmail.com

\* \*Amandeep Kaur Brar

Assistant Professor

amanbrar123@gmail.com

**Abstract:** WCA is a cluster election and cluster formation algorithm. In mobile ad hoc network its make cluster that is called dominant set of network. MANET has many application in field of military rescue operation, disaster management and utmost emergency. WCA increased the capacity of network, reduce computation and communication cost of mobile ad hoc network. This algorithm select the minimum weight node is elect as a cluster head. Cluster head election is based on the factors such as degree of nodes, stability, battery energy, mobility, balancing factor and compute cumulative time node as a cluster head. This review paper provide the knowledge about WCA and improvement in WCA. I also considered the problem formulation in various paper and future scope of this algorithm.

**Keywords-component;** weighted clustering algorithm, mobile ad hoc network.

## I. INTRODUCTION

Current wireless cellular networks rely on the wired backbone by which all base stations are connected, implying that networks are fixed to a geographical area with a pre-defined boundary. Deployment of such networks takes time and cannot be set up in times of utmost emergency. Therefore, mobile multi-hop radio networks play a critical role in places where a wired backbone is neither available nor economical to build, such as battle field communications and disaster recovery situations. Such situations demand a network where all the nodes including the base stations are potentially mobile nodes and communication must be supported untethered between any two nodes.

Ad hoc network offers many excellent technology over the traditional cellular network. Flexibility and fast deployment without any infrastructure are the main advantage of an ad hoc network. All nodes in network can be used as relay, for end to end radio transmission using the multi-hop concept. Ad hoc network are wireless, multi-hop, and dynamic network established by a collection of mobile nodes [23].

The concept of clustering is division of the network into different virtual groups, based on rules in order to discriminate the nodes into different subnetwork. Its goal achieve scalability in presence of large network and high mobility. Various properties of clustering are geographical allocated, balance resource use and service localization. . Classification of clustering are followings

- DS-based clustering
- Mobility aware clustering
- Energy efficient clustering
- Load balanced clustering

- Combined metric clustering
- Low maintenance clustering

**DS-based clustering** a basically dominant set in graph is a subset , in which nodes are assume at the beginning a cluster head role and connected cluster can be merged. DS includes dominating sets and non-dominating sets or nodes. The nodes are connected to cluster head is called dominating set and otherwise it's called non dominating sets or nodes.

**Mobility aware clustering:** Cluster nodes and cluster head are similar moving pattern together in term direction and speed. Every mobile nodes in a cluster has a path to every other node that will be available for some time period with a probability. These consideration are taken into account.

**Energy efficient clustering** technique balanced the energy consumption on nodes by moving the cluster head. It achieved limit the time a node act as cluster head using time counter and limit the size of dominating set.

**Load balanced clustering** limit the minimum and maximum number of cluster in graph. Cluster members and cluster head are periodically broadcast of clustering information. Each cluster head are optimal number of nodes.

**Combined metric clustering** uses multiple metrics to elect a cluster head. Weighted clustering algorithm is example of combined metric clustering. Its parameters are degree difference, distance to neighbor nodes, average moving speed and cluster head serving time. Cluster heads local area are minimum for combined weighting factor, where the sum of weighting factor is 1.

**Low maintenance clustering** reduce the re-affiliation and re-clustering effect. This type of clustering increase the tolerance to topology changes. Least cluster change, 3-hop between adjacent cluster and passive clustering are example of low maintenance clustering.

Weight-based clustering algorithm in ad hoc network is an on demand clustering algorithm for multi-hop packet radio network. These type of networks are ad hoc networks and dynamic in nature due to mobility of nodes. Clustering in mobile ad hoc network can be defined as various partition into various groups. It is an important concept of MANET, because clustering makes it possible to guarantee of system performance, such as throughput delay and also security issues. So in this algorithm they select the cluster head whose performance is well and formation of cluster in ad hoc network. The cluster head is elected based on weight factor, so it is called weighted clustering algorithm. Weighted clustering algorithm is basically appropriate cluster selection in wireless ad hoc network where it is necessary to provide well in face of topological changes caused by node motion and node failure etc. The WCA select the cluster head which has lowest weight among the nodes and some other factor also consider the election for cluster head. The cluster head, form a dominant set in the network, determine the topology and its stability. The weighted clustering algorithm takes into consideration the ideal degree, transmission power and mobility and battery power of nodes.

The main contribution of work [13] is new strategy for clustering a wireless ad hoc network and improvement in WCA. Author derived a simple stability model and a load balancing clustering scheme. They showed that the algorithm outperform in term of cluster formation and stability. Main idea of approach is to avoid cluster re-election and reduce the computation and communication cost. The improved weighted clustering algorithm, the goals of algorithm are maintaining stable clustering structure, minimizing the overhead for clustering set up, maximize lifetime of nodes in the system and achieving good performance.

## ***II. LITERATURE SURVEY***

**Mainak Chatterjee et.al.2002.** In this paper, an on demand distributed clustering algorithm for multi-hop radio network. Author proposed that weighted clustering algorithm overcome the highest degree, lowest id and node weight mathematical strategy. In which network lowest node weight are selected as a cluster head. The cluster head form a dominant set in the network, determine its topology and its stability. The weighted clustering algorithm takes the consideration of ideal degree, transmission power, mobility of nodes, summation of neighbor node distance and battery power of mobile nodes. Dual power mode supply is used by cluster head in this network. The four weighting factor consideration which is adjustable according to the system needs. Simulation experiment is shown average number of cluster, re-affiliation per unit time and number of dominant set updates versus transmission range and maximum displacement. Result shown that algorithm perform better than existing one such as highest degree, lowest id and node weight clustering algorithm.

To verify the performance of the system, the nodes were assigned weights which varied linearly with their speeds but with negative slope. Results proved that the number of updates required is smaller than the Highest-Degree and Lowest-ID heuristics. But this algorithm is compute cumulative time (node work as a cluster head) for each cluster head, this work is very complex.

**Gaurav Gupta et.al. 2003.** In this paper author proposed that sensor energy node, load balanced clustering and given pseudo code for clustering algorithm. They proposed an algorithm to network these sensor into well define cluster with less energy constrained gateway node acting as a cluster head. Gateway node can hear the message request from two or three cluster head. This paper is also described about the different instant of time for select the various cluster head and its neighbor node. This description is consider because cannot any node is attached two cluster head. It is also called multi-gateway cluster sensor network and each cluster head is command given the command nodes. This paper also describe the formula about communication energy dissipation. Simulation result show how our approach can balance the load and improve the lifetime of system. This paper is not consider the information about cluster election and if any gateway node is fail then does not provide any backup.

**Christian bettstetter in 2004.** The network employ's Besagni's distributed mobility adaptive clustering algorithm is self-organizing network structure. Author show that the cluster density and the expected number of cluster heads per unit area formula described. In this paper number of cluster head is reduced by using increase the transmission range

of each node. This paper is showed that if number of cluster head is reduced then performance of communication system will be better. So this paper is give great knowledge about density of cluster head.

**Jane Y. et.al. 2005** .Clustering is an important research topic for mobile ad hoc networks (MANETs). Because clustering makes it possible to guarantee basic levels of system performance, such as throughput and delay, in the presence of both mobility and a large number of mobile terminals. A large variety of approaches for ad hoc clustering have been presented, whereby different approaches typically focus on different performance metrics. This article presents a comprehensive survey of recently proposed clustering algorithms, which we classify based on their objectives. This survey provides descriptions of the guarantee basic levels of system performance such as throughput, delay and also security issues such as availability, in the presence of both mobility and large number mobile terminals. Many clustering protocols for MANETs have been proposed in the literature. Basically this paper is overcome the limitation of high frequency re-affiliation. As a newly proposed weighing-based clustering algorithm, the Weighted Clustering Algorithm (WCA) has excellent performance compared with other previous clustering algorithms. Mechanisms, evaluations of their performance and cost, and discussions of advantages and disadvantages of each clustering scheme. With this article, I can have a more thorough and delicate understanding of ad hoc clustering and the research trends in this area.

**Vincent Bricard et.al. 2006.** In this paper, author proposed a new distributed Weighted Clustering Algorithm with Local cluster-heads election (WCA-L) based on demand distributed clustering algorithm for multi-hop packet radio networks. The multi-hop packet radio networks, also named mobile ad hoc networks (MANETs) have a dynamic topology due to the mobility of their nodes. This mobility makes the challenge harder for routing protocol. Moreover, the well-known routing protocols are not able to offer QoS (Quality of Service) that is why we need to manage MANETs. Such task can be done using clustering techniques but the association and dissociation of nodes to and from clusters perturb the stability of the network topology, and hence reconfiguration of the system is often unavoidable. However, it is vital to keep the topology stable as long as possible. The nodes called cluster-heads form a dominant set and determine the topology and its stability. Simulation experiments are conducted to evaluate the stability of the dominant set in terms of updates of the dominant set, handovers of a node between two clusters and the QoS in terms of packet delivery rate and overhead provided by both our algorithm (WCA-L) and the Weighted Clustering Algorithm (WCA), which does not consider prediction and local election. Results show that algorithm show that better than WCA.

**Wei-dong et.al. 2007.** Ad Hoc network is a kind of multi-hop self- organizing network, which is dynamic in nature due to the mobility of nodes. Many mobile ad hoc applications depend on the hierarchical structure, and clustering is the most popular method to impose a hierarchical structure in the mobile ad hoc networks. Only is some single factor is considered in most existing clustering algorithms, they have very limited application scenarios. In this paper, author propose a novel weight-based clustering algorithm (WBCA). The proposed clustering algorithm takes the mean connectivity degree and battery power of mobile nodes into consideration. In which paper only two weighting factor  $c_1$  and  $c_2$  are taken and values of  $c_1$  and  $c_2$  can be adjusted according to the system requirement. It can improve its

applicability and universality Analysis and simulation of the algorithm have been implemented and validity of the algorithm has been proved. The simulation result shown that the performance of our clustering algorithm are better than of highest id and lowest id heuristic.

**Burst et.al. 2007.** I have studied associate cluster techniques produce stratified network structures, referred to as clusters, on associate otherwise flat network. In a very dynamic environment in terms of node quality additionally as in terms of steady ever-changing device parameters the cluster heads election method should be re-invoked consistent with an appropriate update policy. Cluster re-organization causes extra message exchanges and procedure quality and its execution should be optimized. Our investigations specialize in the matter of minimizing cluster heads re-elections by considering stability criteria. These criteria are supported topological characteristics additionally as on device parameters.

**Hui cheng et.al. 2008.** Clustering can help aggregate the topology information and reduce the size of routing tables. The maintenance of the cluster structure should be as stable as possible to reduce overhead and make the network topology less dynamic. Hence, stability measures the goodness of clustering. However, for a complex system like MANET, one clustering metric is far from reflecting the network dynamics. Some prior works have considered multiple metrics by combining them into one weighted sum, which suffers from intrinsic drawbacks as a scalar objective function to provide solution for multi-objective optimization. Authors proposed a stability-aware multi-metric clustering algorithm, which can achieve stable cluster structure by exploiting group mobility and optimize multiple metrics with the help of a multi-objective evolutionary algorithm. Performance evaluation shows that our algorithm can generate a stable clustered topology and also achieve optimal solutions in small-scale networks.

**Jing An et.al. 2009.** Clustering is an important concept for mobile ad hoc networks (MANETs), because clustering makes it possible to However, the high mobility of nodes will lead to high frequency of re-affiliations which will increase the network overhead and minimize the network lifetime. To solve this problem, author propose an improved weight based clustering algorithm (iWCA), the goals of the algorithm are maintaining stable clustering structure, minimizing the overhead for the clustering set up, maximizing lifetime of nodes in the system, and achieving good performance. They proposed a new method for find relative speed of node versus other nodes. The proposed algorithm taken into consideration degree difference, distance, mobility and consumed energy of nodes. The simulation results demonstrate the superior performance of proposed algorithm in terms of average number of re-affiliation and minimum lifetime of nodes.

**S. Muthuranaligam et.al. 2010.** This paper presents a novel algorithm for clustering of nodes by transmission range based clustering (TRBC). TRBC algorithm does topology management by the usage of coverage area of each node and power management based on mean transmission power within the context of wireless ad-hoc networks. By reducing the transmission range of the nodes, energy consumed by each node is decreased and topology is formed. A new algorithm is formulated that helps in reducing the system power consumption and prolonging the battery life of mobile nodes. Formation of cluster and selection of optimal cluster head taking weighted metrics like battery life,

distance, position and mobility. These factors are adjusted based on the factors such as node density, coverage area, contention index, and current node degree of the nodes in the clusters.

**Yang Wei-dong 2011.** Clustering has often been used to impose structure in wireless ad hoc networks. In this paper, author proposed a novel weight-based clustering algorithm (NWBCA) that tries to increase the stability of the created clusters. It takes into consideration the mean connectivity degree (MCD) and energy status of the mobile nodes. The concept of MCD of nodes is firstly introduced and the calculation of MCD is also proposed. To balance the energy consumption among the nodes, energy status of the mobile nodes is firstly quantified and calculation method of energy status of the nodes is proposed in this paper. The non-periodic procedure for cluster head election is invoked on demand, and is aimed to reduce the computation and communication costs. Simulation experiments are carried out to validate our algorithm. All results show that our algorithm performs better than existing ones and is also tunable to different kinds of network conditions.

**Mohamed Aissa et.al. 2013.** Author consider the problem of appropriate cluster head selection in wireless ad-hoc networks where it is necessary to provide robustness in the face of topological changes due to node motion, node failure and node insertion/removal. They main contribution of work is a new strategy for clustering a wireless AD HOC network and improvements in WCA. Author first derived a simple stability model and thereafter a load balancing clustering scheme. Author showed that our algorithm out performs the Weighted Clustering Algorithm (WCA) in terms of cluster formation and stability .They proposed two new heuristic strategy are stability factor and relative dissemination degree of nodes. In which algorithm four consideration are taken such as distance of node from neighbor nodes, remaining battery energy, stability factor and relative dissemination degree of node. One of the main ideas of our approach is to avoid cluster head re-election and to reduce the computation and communication costs by implementing a non-periodic procedure for cluster head election which is invoked on-demand. Author strived to provide a trade-off between the uniformity of the load handled by the cluster heads and the connectivity of the network. The mobility factor is not considered for the election of cluster head.

**Mohamed Aissa et.al. 2014.** They are also consider the problem of appropriate cluster head selection in wireless ad-hoc networks where it is necessary to provide robustness in the face of topological changes caused by node motion, node failure and node insertion or removal. The main contribution of work is a new strategy for clustering a wireless AD HOC network and improvements in WCA and other similar algorithms. We first derived some analytical models and thereafter some clustering schemes. Our contribution also extends previous works in providing some properties and analyses of Quality of Clustering in AD HOC. Author showed that our algorithm outperforms the Weighted Clustering Algorithm (WCA) in terms of cluster formation and stability. One of the main ideas of our approach is to prioritize favorable nodes in cluster head election and re-election processes. In which paper author proposed a fastest weighted clustering algorithm, but performance of this paper is same a scalable clustering algorithm. But weighting factor consideration is not appropriate and remaining battery energy factor consideration are not right.

**Dahane amino alkane et.al. [2015]** I have studied cluster approaches for mobile Wireless sensing element Networks (WSNs) is to prolong the battery lifetime of the individual sensors and therefore the network period of time. During

this paper, we tend to propose a distributed associated safe weighted cluster algorithmic rule that is an extended version of our previous algorithmic rule (ES-WCA) for mobile WSNs employing a combination of 5 metrics. Among these metrics lie the activity level metric that promotes a secure alternative of a cluster head within the sense wherever this last one can ne'er be a malicious node. The goals of the planned algorithmic rule are: detective work common routing issues and attacks in clustered WSNs, supported behavior level. But five metric is increase the time of select the cluster head so delay time is increased.

### **III. PROBLEM FORMULATION**

Clustering is thought as a graph partitioning drawback with some side constraints. Because the underlying graph doesn't show any regular structure, partitioning the graph optimally (i.e., with minimum variety of partitions) with relevancy sure parameters becomes associate NP-hard drawback. Ensuing drawback consists of proposing our new cluster degree constraint and our changed degree-difference schemes. For the heuristic, the system performance is best compared with the Highest-Degree heuristic in terms of outturn. If in WCA densely inhabited as a result of migration of nodes from alternative zones, then the cluster head may not be ready to handle all the traffic generated by the nodes as a result of theirs associate inherent limitation on the quantity of nodes a cluster head will handle.

Basis for our algorithmic program to determine however well matched a node is for being a cluster heads, we have a tendency to take under consideration its degree, transmission power, quality and battery power. The subsequent options area unit thought of in our agglomeration algorithm:

- Every cluster heads will ideally support solely  $\delta$  (a pre-defined threshold) nodes to make sure economical medium access management (MAC) functioning. If the cluster heads tries to serve additional nodes than it's capable of, the system potency suffers within the sense that the nodes can incur additional delay as a result of they need to attend longer for his or her flip (as in TDMA) to urge their share of the resource. A high system outturn is achieved by limiting or optimizing the degree of every cluster heads.
- The battery power is with efficiency used among nodes. The alternative nodes are take less power than cluster head. But cluster head consume additional power than member nodes. Because cluster head has further responsibility of its member and also communicate with another cluster heads. Whenever cluster head battery is weak then other strong battery life time node is elect as a cluster head. So battery power factor is efficiently consideration is necessary.
- Quality is a crucial consider deciding the cluster heads. So as to avoid frequent cluster heads changes, it's fascinating to elect a cluster heads that doesn't move terribly quickly. Once the cluster heads moves quick, the nodes is also detached from the cluster heads and as a result, a re-clustering happens. Affiliation takes place once one among the normal nodes moves out of a cluster and joins another existing cluster. So mobility factor consideration is necessary. In efficient scalable algorithm does not take the mobility factor.
- Due to the more energy consumption of cluster head. Author used the cumulative time (node act as a cluster head) consideration in weighting factor. The cumulative time calculation is very complex.

- Ripple effect occurs due to the re-election of single cluster head may affect the cluster structure and many other clusters and change the whole topology of network.

#### **IV. FUTURE SCOPE**

In this paper I actually have seen the various issues of various authors. The main downside of constructing a framework for dynamic organizing mobile nodes in wireless ad-hoc networks into clusters wherever it's necessary to produce hardness within the face of topological changes due to the mobility of nodes. In future I am going to derive a straightforward clump load leveling theme. In future we shall be work on problem formulation and simulation result for improving the performance of weighted clustering algorithm.

#### **REFERENCES**

- [1]. S. Basagni. Distributed clustering for ad hoc networks. In: Proc. Intern. Symp. Parallel Architectures, Algorithms, and Networks (ISPAN); (Perth/Fremantle, Australia); June 1999.
- [2]. Mainak Chatterjee, Sajal K. Das and DAMLA TURGUT “WCA: A Weighted Clustering Algorithm for Mobile Ad Hoc Networks” Cluster Computing 5, 193–204, 2002 □ 2002 Kluwer Academic Publishers. Manufactured in The Netherlands.
- [3]. G. Gupta, Mohamed Younis, Load-balanced clustering of wireless sensor networks, IEEE International Conference on Communications, ICC '03, Anchorage, Alaska, USA, May 11-15, 2003
- [4]. Christian Bettstetter. The Cluster Density of a Distributed Clustering Algorithm in Ad Hoc Networks. IEEE International Conference Communications; Page(s): 4336 - 4340 Vol.7; 2004.
- [5]. Jane Y. Yu and Peter H.J.Chong “a survey of clustering scheme for mobile ad hoc network” in 2005, volume 7 and no.1 IEEE.
- [6]. Vincent bricard –vieu, Nidal Nasser and Nofissa, “A Weighted Clustering Algorithm using local heads election for QoS in MANETs” in 2006.
- [7]. L. Agba, F. Gangnon and A.kouri, “Scenario generation for ad hoc network” in 2006.
- [8]. Wei-dong Yang and Guang-zhao Zhang “a weight based clustering algorithm for mobile ad hoc network” in 2007 IEEE.
- [9]. Matthias R. Brust, Adrian Andronache and Steffen Rothkugel “WACA: A Hierarchical Weighted Clustering Algorithm optimized for Mobile Hybrid Networks” pp 193-204 in 2007.
- [10]. Hui Cheng, et al. Stability-aware multi-metric clustering in mobile ad hoc networks with group mobility. Wireless Communications and Mobile Computing; 9:759–771, Published in 21 April 2008
- [11]. Jing An and Chang Li “An Improved Weight Based Clustering algorithm” in 2009 IEEE.
- [12]. S. Muthuramlingam and R. Rajaram, a transmission range based clustering algorithm for topological MANET in 2010.
- [13]. Yang Wei –Dong “weight based clustering algorithm for mobile ad hoc network in 2011.
- [14]. Mohamed Aissa, Abdelfettah Belghith “An Efficient Scalable Weighted Clustering Algorithm for Mobile Ad Hoc Networks” IEEE 2013.
- [15]. Mohamad Aissa and Abdelfettah Belghith, Quality of clustering in mobile ad hoc networks in 2014 (Elsevier).
- [16]. Amine Dahane “A Distributed and Safe Weighted Clustering Algorithm for Mobile Wireless Sensor Networks” The 6<sup>th</sup> international conference in 2015.
- [17]. R. Pandi Selvam et al, Stable and Flexible Weight based Clustering Algorithm in Mobile Ad hoc Networks / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (2), 2011, 824-828.
- [18]. Suchismita Chinara, Santanu Kumar Rath. A Survey on One-Hop Clustering Algorithms in Mobile Ad Hoc Networks. Journal Network System Management; 17:183–207; 2009.
- [19]. Naveen Chauhan et al. A Distributed Weighted Cluster Based Routing Protocol for MANETs. Wireless Sensor Network; 2011; 3, 54-60, doi:10.4236/wsn.2011.32006 Published Online February 2011.
- [20]. S. K.B. Rathika and J. Bhavithra. An Efficient Fault Tolerance Quality of Service in Wireless Networks Using Weighted Clustering Algorithm; Bonfring International Journal of Research in Communication Engineering; Vol. 2, Special Issue 1, Part 4; February 2012.

- [21]. Tolba, Magoni, D ; Lorenz, A stable clustering algorithm for highly mobile Ad Hoc networks, P. Second International Conference on Systems and Networks Communications, 2007. ICSNC 2007.
- [22]. Xi'an Jiaotong, et al. WACHM: Weight based adaptive clustering for large scale heterogeneous MANET; Communications and Information Technologies; ISCIT '07; 2007.
- [23]. M. Amine Abid, Abdelfettah Belghith. Stability routing with constrained path length for improved routability in dynamic MANETs. The International Journal of Personnal and Ubiquitous Computing, Springer, Volume 15, Issue 8, pp. 799-810, 2011.

# Big Data Study –Basics, Techniques and Tools

Tara Singh

Department of Computer Engineering  
Punjabi University Patiala, India.  
tarasingh26.ts@gmail.com

Assistant Professor Karandeep Singh

Department of Computer Engineering  
Punjabi University Patiala, India.  
[karan\\_rob7@yahoo.co.in](mailto:karan_rob7@yahoo.co.in)

**Abstract-**A lot of technologies revolution are driving rapid increase in data generation and data gathering. This is why big data is today's buzzword in every science and engineering domain. Big Data is similar to small data but bigger in size. Big data is not being greater than certain number of petabytes or terabytes. This can be assumed technology expands over time and data set size which is assumed as Big Data will also increases. Data is increasing day by day exponentially so today data is increased beyond the storage capacity of traditional databases. To process and analyze such datasets which cannot be stored in traditional databases Big Data tool like Hadoop is used. This paper will discuss the introduction, characteristics and various tools used to process and analyze big data.

**Keywords**—*Big Data; Data Types; Volume; Velocity; Traditional Data; Sources; Hadoop; MapReduce; HDFS; Tools;*

## I. INTRODUCTION

Big data is a dataset whose size is big enough so that it becomes difficult to store, analyze and manage in traditional database software system. We cannot say Big Data is being greater than certain number of petabytes or terabytes. We imagine technology expands and changes over some time the data set size that is considered as big data will also increases. Big data also defined as continuing change or expansion in data volume as well as velocity of data processing. Big Data is everywhere like Facebook, Twitter, Linked in, Google, shopping sites like Amazon etc. all are using Hadoop to handle Big Data. Big data is also present in manufacturing, healthcare, public or private clouds. By discovering relations and associations, understanding patterns and trends in hospital big data analyst in hospital can save lives, improve care and reduce costs [1]. Today data is increasing too fast, changes too fast so to process this expanding data we need the term Big Data. To extract useful value from this increasing and changing data one must choose alternative methods to process it.

- Oxford English Dictionary (OED) defined Big Data as Data of very big size, typically to level that its manipulation and handling present significant logistical challenges [6].
- Rodriguez (2012): “For years, statisticians have been working with large volumes of data in fields as diverse as astronomy, bioinformatics, and data mining [7]. Big Data is uncommon because it is generated on a largescale by countless online interactions among people, transactions between people and systems, and machinery embedded with sensors”[7].
- Horrigan (2013):-“I view Big Data as non sampled data, distinguished by the creation of databases from electronic sources whose primary goal is something other than statistical inference”[7].

Parameter	Traditional Data	Big Data
Size	GB	Continually increasing
Data production rate	Per hour, day	Rapid increase
Data type	Structured	Semi or unstructured
Approach	Centralized	Distributed
Data model	Fixed Schema	Schema less
Data store	RDBMS	Hadoop, NoSQL

Table 1: Difference between Big Data and Traditional Data

## II. 3V'S OR 3V MODEL OF BIG DATA

In 2001, industry analyst Doug Laney introduced “3Vs” volume, variety and velocity [6]. These 3Vs are used to define the Big Data. These V's are also called 3V model of Big Data [2].

There is number of data generation sources which provide data input to the Big Data system e.g. social networks, bank transactions, content of web pages, GPS trails, financial data etc. Now question is are all these are the same things?

To clarify this volume, variety and velocity are usually used to characterize various forms of Big Data.

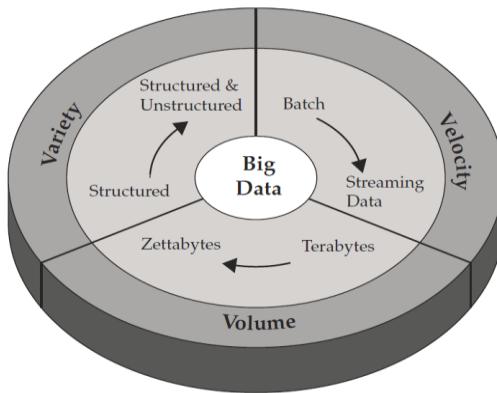


Figure 1: 3V's of Big Data [2]

### A. Volume:-

Volume refers to size of data which is increasing day by day typically starting at tens of terabytes. As an estimation 2.5 Quintillion data is generated every day. Facebook generates 500+ terabytes data daily. Twitter generates 8TB data per day. To handle such amount of increasing data Big Data term is used.

### B. Velocity:-

Velocity refers to speed at which data is being generated and changes. Today data is created and processed rapidly.

### C. Variety:-

This defines the fact that data came from different sources in different formats and structures with different data types. These data types can be structured, semi structured and unstructured. Big Data consists of semi structured or unstructured data.

### D. Different Types of Data:-

*Structured Data:-*

Structured data is the kind of data used by traditional database software where data is stored in defined relations or tables which can be easy to search, store, and retrieve based upon some conditions and rules [3]. Structured data is handled by SQL commands. Structured data of class students is shown in figure2.

Name	Roll_no	Class	Phone
Tara	100	MTech	9872400232
Sohal	101	MTech	9872400265
Amrinder	102	MTech	9872180265
Gurpinder	103	MTech	9826180265
Priya	104	MTech	9826170265
Navneet	105	MTech	9865170265
Boparai	106	MTech	9865170785
Abhi	107	MTech	9865196785
Pankaj	108	MTech	9865198528

9 rows in set (0.00 sec)

mysql>

Figure 2: Students Record in Structured Form

*Unstructured Data:-*

This type of data has no predefined format. This data cannot be stored in databases or other architectures. This type of data can be textual or non textual. Data generated by media like e-mail messages, word documents falls under textual category. JPEG images, audio files are examples of non textual unstructured data.

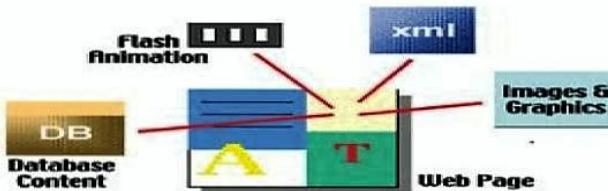


Figure 3: Unstructured Data [3]

*Semi-structured Data:-*

This type of data is the combination of the two discussed above. This data cannot be organized in databases. It is structured data but cannot be stored in relational model. XML, EDI (Electronic Data Interchange), web server log and data collected by remote sensors are examples of semi structured data.

### III. WHY IS BIG DATA NEEDED

1. To increase storage capacities
2. To increase processing power
3. To store different types of data (structured, unstructured)
4. By a survey performed in 2012, more than 950 million users, Facebook ingested 500+ terabytes of new data into their database every day [4].
5. In every second, 6000 tweets are tweeted on twitter which corresponds to 350,000 tweets in one minute, 500 million tweets in one day and 200 billion tweets in one year. Based upon this estimation twitter generates 8TB data daily.
6. A survey performed by IBM shows that today's 90% of stored data was generated in previous two years [3].

7.A report says that Data from the U.S. healthcare system alone reached in 2011,150 exabytes [1]. At this rate of growth, Big Data for U.S.healthcare will soon reach zettabytes (1024 Exabytes) not longer after yottabytes(1024 Zettabytes) [1].

8. In every second, 6000 tweets are tweeted on twitter which corresponds to 350,000 tweets in one minute, 500 million tweets in one day and 200 billion tweets in one year. Based upon this estimation twitter generastes 8TB data daily [5].

So, 2.5 Quintillion data is created by all the data sources per day [3].To manage and process such amount of changing and increasing data we need the term Big Data.

#### IV. VARIOUS DATA GENERATION SOURCES

##### *1.Data Collected From Sensors*

Sensor data analytics is the next step of information technology. In order to evaluate output and quality of work in organizations like manufacturing, are deploying sensors in their products and equipment. It is assumed that after few years we have to talk in Bronto Bytes when we consider data from sensors. Data like log data, geo location data, cpu utilization and temperature etc. is collected by sensors. In case of variety, we can see that remote sensing data consists of multiple sources (laser, radar etc.), multi temporal (data from different sites), multi spatial resolution data. In case of velocity, in remote sensing velocity does not involve rapid data generation speed only but also efficiency of data processing and analysis or we can say data should be analyzed in reasonable time to complete the given task e.g. seconds can save hundreds of thousands of lives in natural disasters like earthquake and tsunami.

##### *2.Point Of Sale:-*

Point of Sale (POS) and inventory control systems:-Online retail in US is about \$400 billion which is less than 8% of whole retail industry in the world. E-commerce company like Amazon each and every search , purchase and visit is mined and these are used to make the customer's shopping experience better.

##### *3.Internet Websites:-*

Websites are the best way to provide information to users. There are about one billion websites in the world and in every minute this number is increasing. Social sites like Facebook, Twitter and Google + creating huge amount of data per day which is continually increasing. 200 e-commerce items ordered in one second. Total of 2.5 Quintillion data is created by online website. That's huge data to analyze.

##### *4.Public, Private,Community Clouds:-*

Cloud computing is on- demand network approach to resources. Cloud computing can be classified into Software as a service (SaaS), Platform as a service (PaaS),Infrastructure as a service (IaaS.)

##### *5.Data Generated By Government Agencies:-*

Public sector is large area of the global economy facing challenges to improve its productivity. Govt. has rights to approach this digital data but it is difficult for them to take advantage of this data in efficient way.

#### 6.Bank / Credit Card Transactions:-

Data sources in banks are customer walk-in, emails, internet banking, voice call, social media, websites etc. HDFC is the first bank which moves to big data. Banks in average generate petabytes data, so big data in banking is no exception. Figure 4 shows the percentage of respondents corresponding to data sources.

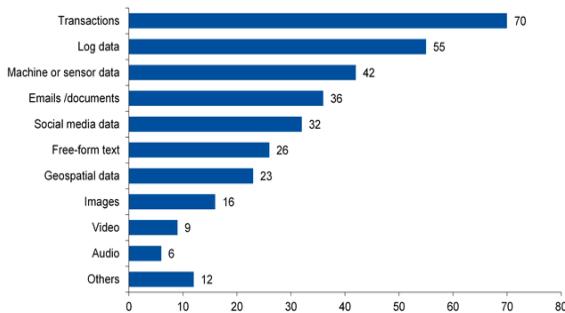


Figure 4: Percentage Of Respondents [2]

### V. BIG DATA IN DIFFIRENT COMPANIES

#### A. Big Data at Google:-

Google has not only significantly effected the way we can now analyze big data (Hadoop, MapReduce, BigQuery, etc.) but they are more responsible than anyone else for making it part daily lives. In 2007 Google launched its universal search which gather the data from hundred of sources including weather forcasting, language databases,financial and historical and many more[10]. In 2012 it involved knowledge graph which show the information regarding topic from a number of sources into search results. In 2010 Google launched its Google Query to process, store, analyze Big Data on cloud platforms [10]. Google is working on its Big Data project self driving cars which will use and create data from sensors, cameras, tracking devices etc.

#### B. Big Data at Amazon:-

Amazon is the one of the Big Data generation company in these days. Amazon pioneered e-commerce in many ways, but possibly one of its greatest innovations was the personalized recommendation system which, of course, is built on the big data it gathers from its millions of customer transactions [10].

### VI. USEFUL TOOLS AND TECHNIQUES IN BIG DATA

**A. Hadoop:-** Hadoop is a software platform used to run and write applications that process large data sets. It is scalable, efficient, reliable and economical way to process large data volume [8]. Organizations like Yahoo, AOL, Facebook using Hadoop technology. Yahoo has more than 100,000 CPUs in over 40,000 servers running Hadoop, with its biggest Hadoop cluster running 4,500 nodes That's big, and approximately four times larger than Facebook's beefiest Hadoop cluster [9]. Hadoop divides Big Data into number of units which can be processed and analyzed simultaneously. It is implemented by Google's MapReduce user friendly function developed in early 2000s through Hadoop Distributed File System (HDFS).MapReduce divides the work into small units and process them in parallel. MapReduce has two stages:-

**Map stage:-**Map's responsibility is to process input data, divide into small units and assign to worker nodes. Worker node then do it again that corresponds to multi-level tree structure [2].

*Reduce stage*:-Master node receives the answers from all sub units and add them to form output. It is the combination of shuffle and reduce stages basically [2].

*B. HDFS (Hadoop Distributed File System)*:- HDFS is a block structured distributed file system which follows client-server paradigm has Name Node and many Data Nodes, Name Node stores the meta data[2]. When there is a failure in Name Node then Hadoop cannot do automatic recovery

*C. NoSQL*:- A related new style of databases called NoSQL (Not Only SQL) has emerged to like hadoop,process large volumes of multi-structured data [3]. NoSQL provides a mechanism for storage and retrieval of data other than data storage and retrieval in database. NoSQL Database provides Java, C, Python and node.js drivers and a REST API to simplify application development [11]. Oracle NoSQL Database 12.1.3.5.2 was released late in 2015.Major features include Kerberos integration for external authentication [11]. Bulk Put API - this new API allows for the customer to perform Bulk Put operations for both rows and Key/Value inserts in a single API call [11].

## VII. CONCLUSIONS

We have entered in the word of Big Data. With increase in data size every organization must have to replace their traditional databases with Big Data technology to process and store large datasets in near future. There are so many challenges in order to analyze data. Organizations which are using Big Data are Google, Facebook, Twitter, eBay etc. There is number of advantages using Big Data like cost reduction, faster and better decision making. With Big Data technologies, we will possibly able to extract most relevant and most accurate information from various autonomous sources for better understanding and decision making. This study will help you to reveal the basics of Big Data, why Big Data is needed and What type of tools and techniques are used to process and analyze Big Data.

## REFERENCES

- [1] Wullianallur Raghupati, and Viju Raghupathi, “On Big data analytics: promise and potential”
- [2] Shilpa, Big Data and Methodology –A review, LPU, Phagwara, India
- [3] International journal of advanced Research in Computer and Communication Engineering Vol2., Issue 6, June 2013 “The Future Revolution on Big Data”.
- [4] <https://gigaom.com>
- [5] <http://www.internetlivestats.com/twitter-statistics/>
- [6] <http://www.forbes.com/sites/gartnergroup>
- [7] www.bls.gov/osmr/symp2013\_horrigan.pdf
- [8] <http://www.tutorialspoint.com/hadoop/>
- [9] <http://www.techrepublic.com/>
- [10] <http://www.ap-institute.com/big-data-books>
- [11] <http://www.oracle.com/technetwork/database/database-technologies/nosqldb/overview/>

# OUR SPONSORS

---



VARDHMAN COMPUTERS

---



**Organized By:**

Department of Computer Engineering, Punjabi University, Patiala – 147002

Web: <http://www.csepup.ac.in/icrtc/2/> Email: icrtc16@gmail.com

Ph: 0175-3046337

**Proceedings to be published by:**

International Journal of Engineering Sciences

(An Open Access, biannual, peer reviewed, indexed refereed journal)

ISSN: 2229-6913 (Print), ISSN: 2320-0332 (Online)